Rubin 系列全面量产，Vera 专为 Agent 设计性能超 x86

英伟达（NVDA.O）COMPUTEX 2026 跟踪报告

推荐（维持）

TMT 及中小盘/电子

事件：

英伟达 CEO 黄仁勋 6 月 1 日在 COMPUTEX 2026 大会发表主题演讲，介绍公司在 Agentic AI 领域和 Vera Rubin 机架最新进展，以及 Vera CPU 等硬件平台与 AI PC 等技术前沿应用。综合演讲及材料信息，总结要点如下：

评论：

1、Agentic AI 成为利润创造者，Harness 为 Agent 操作系统。

英伟达表示 Agentic AI 时代已经到来，标志着全新的计算模式，不仅仅是生成式 AI 的延续，更实现了“有用性”的跃升，AI 成为利润创造者，可以通过 Prompt 让它生成代码，并产生输出。智能体作为终极的解构式与分布式计算模型，由模型+Harness（软件底座）+工具+Runtime 组成，模型作为大脑，Harness 负责调度，使其完成生产性工作，英伟达 CUDA X 库作为智能体的工具，输入数据后，智能体进行理解、观察、推理、行动并使用工具。工具使用作为最大的突破，不需要担心智能体到来会导致软件公司倒闭，世界上将会出现数以亿计的智能体，将比以往任何时候都更频繁地使用工具。工具运行在 CPU 和 GPU 以及大语言模型上，安全 Harness 运行在 CPU 和 BlueField DPU 上，所有这些的编排工作都由 CPU 完成。

2、Vera Rubin 已进入全面量产阶段，机架组装时间缩短至 5 分钟。

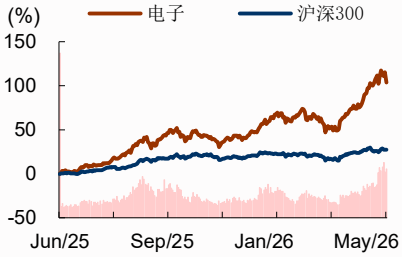
Vera Rubin 是端到端系统，包含 GPU、Vera Rubin NVLink 72，并由 Vera CPU 进行协调，目前已全面量产，建立的供应链规模是 Grace Blackwell 的两倍。架构核心是极端协同设计，其逻辑基于：每一个 Token 都是收入都是盈利的，计算即营收，计算即利润，而收入和利润的缺失即是损失。Vera Rubin AI 工厂构建中，**1）DSX 代表基础设施：** Nvidia DSX 作为参考设计，通过 DSX SIM Omniverse 蓝图，合作伙伴可以在动工建设前设计并验证 Vera Rubin AI 工厂；利用 DSX Max Lps 实现机架间的动态电力分配，将回收的搁浅电力转化为计算产出；通过 DSX Flex 实时协同电网信号，在电网压力大时动态调整工厂功耗，保持最高运行效率。**2）模块化集成：** Vera Rubin 系统由七款 3 纳米制程新芯片组成，结合 CoWoS-R 与 CoWoS-L 封装及各方 HBM 内存，由 Vera Rubin NVL72 执行模型思考、Prompt 处理及推理规划。**3）第三代 MGX 机架设计：** 系统采用全新的模块化计算链路与流线型 PCB 中板设计，将两侧直接连接，过去需要两个小时的装配时间，现在仅需五分钟。Super Chips 连接 X9 Super Nyx 和 BlueField 4 DPU 以提升韧性。**4）Groq 3 LPX 架构：** 16 个托盘中集成了 256 颗 Groq 3 LPU，Groq LPX 则负责以最低延迟完成 Token 生成，与 Vera Rubin NVL72 架构形成高吞吐与低延迟的互补算力组合。全新的 LPX LPU30 芯片是核心算力单元。**5）Vera CPU 机架：** 单机架内集成 256 颗液冷 CPU，在富士康和 Quanta 的工厂中运用，搭配 CX9 芯片及完善的软件栈，是有史以来建造的最先进的 CPU。

行业规模

		占比%
股票家数（只）	530	10.2
总市值（十亿元）	20660.8	17.8
流通市值（十亿元）	17372.4	16.6

行业指数

%	1m	6m	12m
绝对表现	16.1	48.7	105.3
相对表现	14.4	40.7	78.5



资料来源：公司数据、招商证券

相关报告

1、《英伟达 FY27Q1 跟踪报告—Rubin 按计划 Q3 出货，CPU 带来 2000 亿美元新增市场》2026-05-21

2、《英伟达 GTC 2026 跟踪报告—25-27 年 DC 收入超 1 万亿美元，Kyber 将使用铜光等多种互连形式》2026-03-18

3、《英伟达 CES 2026 跟踪报告—Vera Rubin 已正式量产，展示全新 Agentic 和 Physical AI 平台》2026-01-06

鄢凡 S1090511060002
yanfan@cmschina.com.cn
程鑫 S1090523070013
chengxin2@cmschina.com.cn
谌薇 S1090524070008
shenwei3@cmschina.com.cn
涂锟山 S1090525040004
tukunshan@cmschina.com.cn

3、Vera CPU 专为智能体设计，性能相比最高性能的 x86 处理器进一步提升。

传统的 CPU 追求单插槽核心最大化，智能体时代，CPU 已成为 GPU 利用率的瓶颈，直接影响 Token 的吞吐量、延迟和用户体验。Vera CPU 专为智能体设计，机架内集成了两颗 Vera CPU，负责管理 GPU、处理 KV 缓存以及处理机架内的所有软件调度，作为 Grace BlueField 的核心，用于安全与隔离，同时负责 Harness、AI 模型编排、工具调用及数据库访问。**针对智能体需求的四大特征：1）**Vera 拥有全球最高的指令执行效率（IPC）；**2）**拥有极高的带宽密度，是首个达到 reticle limits 的 CPU；**3）**I/O 能力方面，Vera 是首个支持 PCIe Gen 6 的 CPU，Vera 是首个支持 PCIe Gen 6 的 CPU；**4）**Vera 在提供超高性能的同时，为 GPU 留出足够的电力空间。Vera 是首个使用 LPDDR5X 内存的 CPU，与最高性能的 x86 处理器进行对比：Vera 将峰值内存延迟降低了 40%，agentic sandbox 性能达到了 x86 CPU 的 1.8 倍。Vera 将成为全球优化程度最高的智能体 CPU，将打开 CPU for Agents 这个更大的新市场。

4、NV 联手微软推出全新 AI PC 产品，构建本地化智能体计算终端。

NVIDIA 与 Microsoft 推出面向 Windows 的全新 PC 产品线，覆盖笔记本、台式机和工作站，兼容 Windows、支持 CUDA、搭载 NVIDIA AI Tensor Core。英伟达 AI PC 的核心在于将数据中心级 AI 能力下沉至个人终端，使 PC 从传统应用入口升级为智能体运行平台。发布的 RTX Spark 核心平台整合 Blackwell RTX GPU、定制 Grace CPU、NVLink、统一内存以及 Windows 智能体运行环境，让智能体可以在本地 PC 上运行，并调用本地应用和工具。

5、从 Cosmos 到 Isaac GROOT 机器人开发平台，构建物理 AI 生态底座

英伟达通过推出一系列开放模型与平台，构建了完整的企业级 AI 与物理 AI 生态系统：**1）Nemotron 3 Ultra:** 下一代混合架构（SSM+MoE）开放模型，开放训练数据与方法。对比领先模型，推理速度提升 5 倍，成本降低 30%，强化了长程推理与工具调用能力。**2）Cosmos 3:** 面向物理 AI 的全模态模型，通过自回归与扩散 Transformer 结合，实现对物理世界的理解与生成。它推动了“计算本身即可生成数据”的物理 AI 范式转变。**3）Alpamayo 2:** 自动驾驶专用开放模型，运行于 Nvidia Hyperion 平台与 HALOS 操作系统。它赋予车辆环境理解、路径规划及决策解释能力，赋能全球 80%的汽车产能。**4）Isaac GROOT:** 人形机器人全栈开发平台，集成 Isaac Lab、Omniverse、Jetson Thor 等核心组件。通过提供标准化软硬件参考设计，大幅降低了机器人研发与部署门槛。

风险提示：竞争加剧风险、贸易摩擦风险、行业景气度变化风险、宏观经济及政策风险。

附录：英伟达 COMPUTEX 2026 主题演讲纪要

时间：2026 年 6 月 1 日

主讲人：英伟达创始人黄仁勋

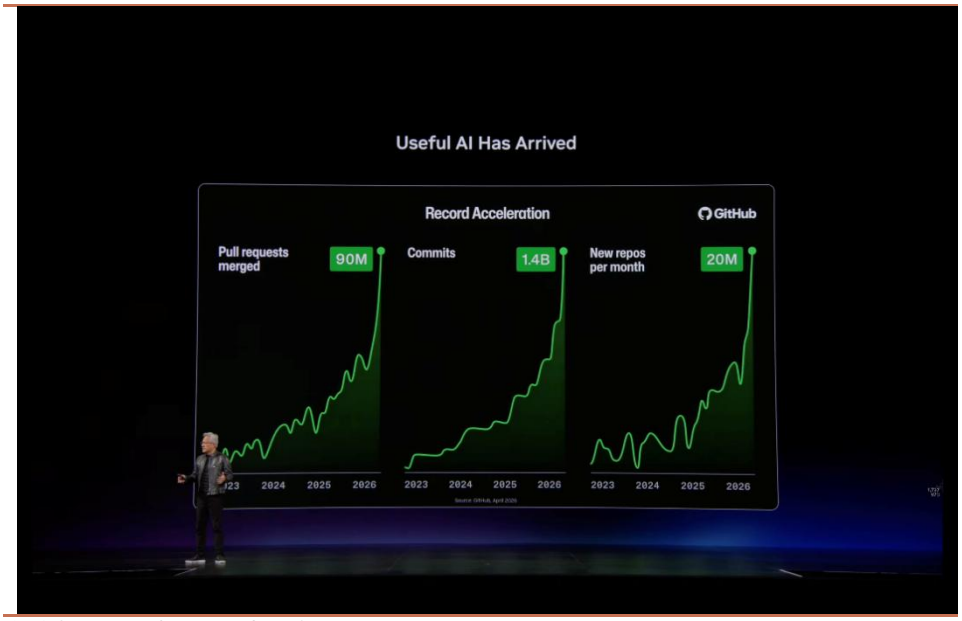
1、Agentic AI：新计算范式重塑 Token 经济，英伟达全栈构建 AI 基础设施生态

中国台湾生态系统的规模令人难以置信。通常人们提到“生态系统”时，会想到我们的软件栈以及英伟达构建的计算系统之上的开发者生态。但我们的生态系统贯穿整个上游，从中国台湾的整个供应链开始，一路向下游延伸至数据中心，并最终到达终端用户。今天，我们将探讨几乎涵盖整个生态系统的内容。中国台湾拥有全球最丰富的生态系统和全球最优秀的供应链，今年，我们的业务增长迅速。中国台湾的年度 GDP 预计将增长近 10%，这太不可思议了。

两年前我曾提到，AI 已从生成式 AI 演进，随后将涌现出其他 AI 浪潮。下一个浪潮就是 Agentic AI。今天我们可以断言：Agentic AI 已经到来，实用的 AI 已经到来。这究竟意味着什么？GitHub 就是智能体 AI 最早的应用之一。软件编程是全球最有价值的职业之一，拥有极大的生态系统——全球约有 3000 万到 4000 万名专业软件开发者，还不包括成千上万的学生和爱好者。GitHub 上的“Pull Request”代表着软件的下载与修改，而“Commit”代表着代码的提交。如果在 2023 年，提交量为 3 亿次，2024 年为 4 亿次，2025 年为 5 亿次，而在 2026 年的前几个月，这一数字几乎翻了三倍。

这说明了什么？这 3000 万软件开发者创造了约 3 万亿美元的 GDP。他们每年产生 3 万亿美元的薪资支出，进而带动了其他行业的经济增长。假设全球 100 万亿美元的产业价值受到影响，而这仅仅是由 3 万亿美元的薪资所驱动的。现在，这些薪资正在产生近三倍的产出，实际上形成了 9 万亿美元的生产力。这完全合乎逻辑。这正是 AI 的潜力和前景。软件工程师的数量实际上在增加。有些人说 AI 会减少工作岗位，这完全是错误的。恰恰相反，它促使企业雇佣更多的软件工程师。原因很简单：如果你能雇佣一名软件工程师，并产生 9 万亿美元的生产力，那么当产出如此惊人时，人们自然会想要雇佣更多的工程师。这将在我们的经济中很快显现出来。因此，第一件事是有用的。

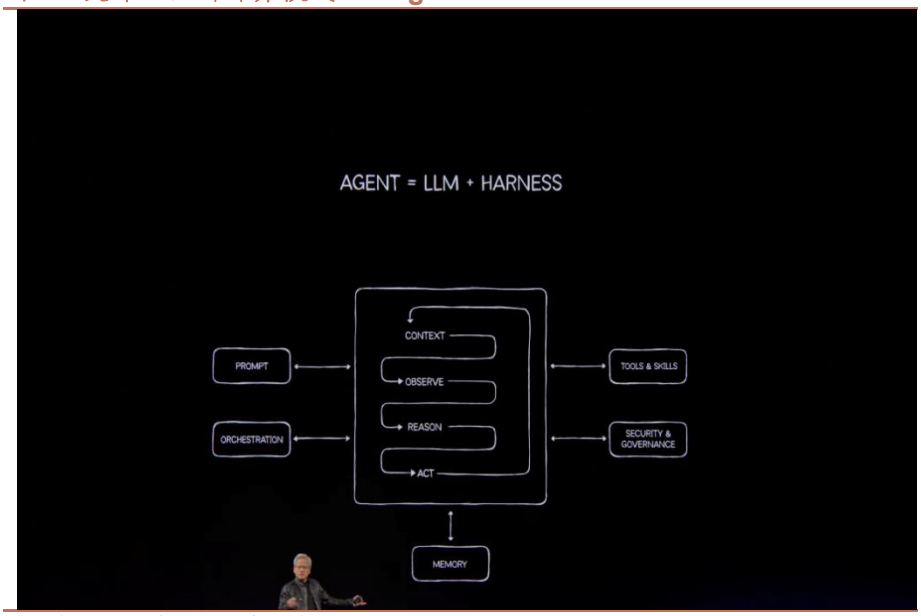
图 1：GitHub 使用次数与价值创造快速增长



资料来源：英伟达，招商证券

AI 已经到来。从行业角度看，这意味着 Token 需求激增。因为既然 AI 能带来收益，人们自然会想要生产更多。Token 已成为盈利的单位，因此 AI 公司渴望构建更多的 AI 工厂。这就是为何中国台湾的计算需求呈指数级增长。计算模式变了。AI 现在是利润创造者，是 GDP 创造者，其背后是一种全新的计算模式——不仅是大语言模型，而是“智能体”（agent）。过去的应用是程序代码在应用程序中运行，在操作系统中运行。而今天，智能体由大语言模型或多个模型组成，它们运行在一个“Harness”（软件底座/外壳）中。这个 Harness 负责调度，使其完成生产性工作。输入数据后，智能体必须理解、观察、推理、行动并使用工具——这些工具可以是电子表格、浏览器、数据库引擎等。Harness 协调每一次信息路由，处理上下文，理解正在发生的事情，进行推理，并制定出可执行的行动方案。

图 2：英伟达全新计算模式——agent



资料来源：英伟达，招商证券

因此，这本质上是一个代理。智能体处理短期记忆（即工作记忆）和长期记忆，

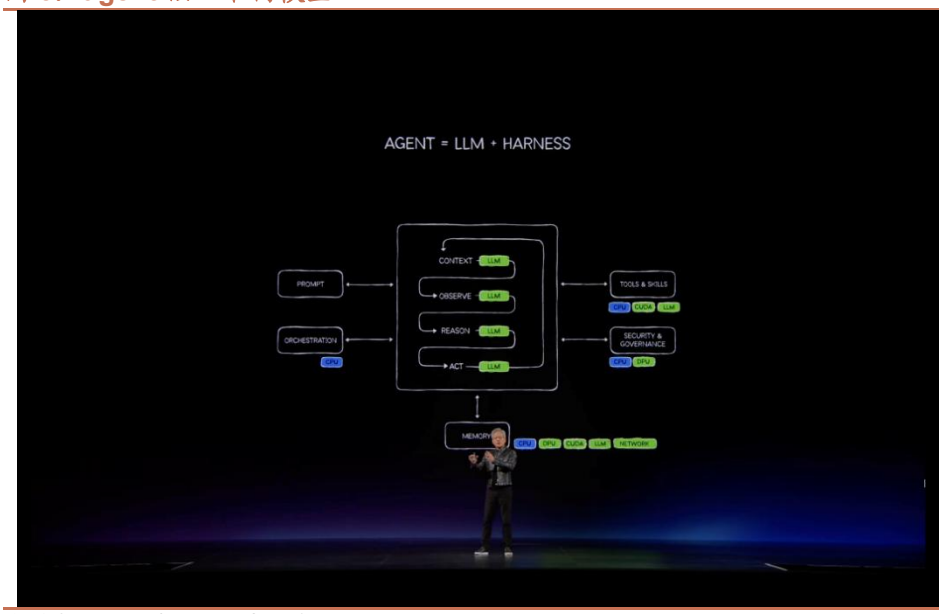
这与人类类似，因此内存管理系统至关重要。大型语言模型负责思考，而 Harness 将一切连接在一起，就像操作系统一样。这是新的计算模式。我们可以通过 Prompt 让它生成代码，并产生输出。这就是全新的计算模式。过去我们启动应用程序、点击并输入；现在我们只需向 AI 解释我们的意图，AI 就会生成代码、使用工具并产出结果。这就是未来计算机的工作方式，这就是 Agentic AI。

两年来我们一直致力于此，现在它已到来。最大的突破之一是“工具使用”。有些人担心智能体到来会导致软件公司倒闭，事实恰恰相反，因为世界上将会出现数以亿计的智能体，它们将比以往任何时候都更频繁地使用工具。这对软件公司而言是不可思议的机遇。但软件必须以智能体能够使用的方式呈现。英伟达的宝库正是我们的 CUDA 库，我称之为 CUDA X 库。今天，我们能够将这些 CUDA X 库呈现给智能体，它们使用这些库的效率甚至远超人类。二十年前，我们构建了 CUDA，一种用于加速计算的单一架构。现在，我们重塑了计算，1000 个 CUDA X 库帮助开发者在科学和工程的每个领域取得突破。CUDA X 库就是智能体的工具。例如：用于计算光刻的 cuLitho，用于决策优化的 cuOPT，用于直接稀疏求解器的 cuDSS，用于跨结构化与非结构化文档进行深度研究的 AIQ，用于 AI 的 Ariel，用于可微物理的 Warp，以及用于基因组学的 Parabricks。它们的基础都是算法。

这就是智能体，它是终极的“解构式”（disaggregated）和分布式计算模型。为了处理这个智能体，将有无数不同的计算机被激活。智能体由模型、Harness、工具、技能以及运行时环境组成，所有这些都在数据中心的不同位置并行运行。你可以把模型想象成大脑，Harness 想象成躯体，工具则是它在运行时环境下使用的设备。可以把它看作一个工作坊，一名工人正在工作坊里使用工具。当然，这是在极大规模下完成的，每一个步骤都在计算机的不同部分运行。你可以看到大语言模型在思考。上下文处理、观察、理解环境、推理、制定计划并付诸行动。每当这一过程发生，整个 Grace Blackwell NVLink 72 机架就会被激活。它利用大语言模型进行思考；每当它使用工具时，CPU 就会参与其中。这个工具可能是一个 C 编译器、Python、JavaScript 或者加速计算任务。

今天的智能体是相对简单的工具使用者。而未来，它们将成为非常复杂的工具使用者，这也是为何我刚才展示的 CUDA X 库将受到智能体热烈欢迎的原因。它们解决了世界上一些最重要的问题。我们所有的 CUDA X 库现在都将具备技能，AI 可以学习如何使用这些库。CUDA X 库不仅是工具，更包含了手册，AI 读取手册后就会明白：“啊，原来是这样使用它的。”智能体使用这些库的能力将是不可思议的。工具运行在 CPU 和 GPU 以及大语言模型上，安全 Harness 运行在 CPU 和名为 BlueField 的安全处理器（BlueField DPU）上，所有这些的编排工作都由 CPU 完成。这就是整个 Harness，由 CPU 统筹全局。

图 3：agent 核心架构模型



资料来源：英伟达，招商证券

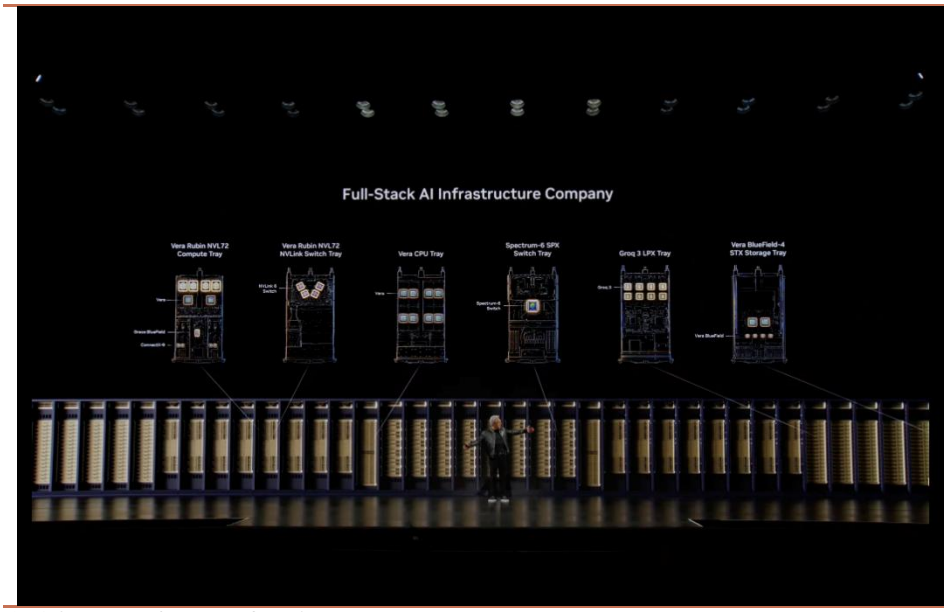
最困难的部分之一是内存。工作记忆被称为 KV 缓存，我们需要考虑如何压缩记忆，不仅仅是压缩，更重要的是如何检索。你是要检索结构化数据还是非结构化数据？ontology 是什么？所有这些不同数据之间的关系又是如何构成的？整个处理过程极其复杂。人工智能的内存系统将引发存储系统的彻底革命。正如你所见，这种计算模式、这种计算范式、这种被称为“智能体”的新应用，与过去在操作系统二进制文件中运行大量软件的方式有着本质的不同。这正是我们构建下一代 Vera Rubin 的原因。Vera Rubin 不仅仅是一块芯片，也不仅仅是一个 GPU。它始于 GPU，但 Vera Rubin 是从端到端不可思议的系统，它包含 GPU、Vera Rubin NVLink 72，并由 Vera CPU 进行协调。

存储系统、革命性的 Vera，以及配合 CX9 芯片的 DOCA 软件栈，还有内置的安全处理器，确保了一切在静态、传输及使用过程中均处于加密状态。由于 AI 模型如此珍贵，整个系统必须遵循“机密计算”（confidential computing）原则。每一个系统本身都是一场革命。Vera Rubin 是英伟达历史上最雄心勃勃的努力。公司全体 4 万名工程师都参与了 Vera Rubin 的研发，更不用说在座的各位也参与了这一整个系统的创建。Vera Rubin 确实是一个奇迹，它不仅仅是一艘“船”，它的意义远超于此。

很久以前，英伟达曾是一家 GPU 公司。但多年来，我们已经进化。如果你成为一家系统公司，你就是在设计最复杂的、从零开始构建的系统。但最终，我们的客户和合作伙伴并不只是想购买计算机，他们想要构建“AI 工厂”，这正是英伟达再次转型的动力所在。正如你所见，我们的技术现在已达到整个基础设施的规模。我们的合作伙伴涵盖了电力发电机、冷却系统、电网供应商等基础设施厂商。如此多的工业公司现在成为我们生态系统的一部分，因为我们试图构建一个完整的栈，就像我们建造 GPU 和 Grace Blackwell NVLink 72 一样。我们正在构建全栈系统，以便我们的客户能够建设令人惊叹的 AI 基础设施。

让我们来看一看。世界正在竞相建设 AI 工厂，这是人类历史上最大规模的基础设施建设。AI 工厂极其复杂。每一层级——芯片、机架、网络、电力、冷却网格——都必须进行端到端的协同设计，因为计算即营收。

图 4：Vera Rubin 架构

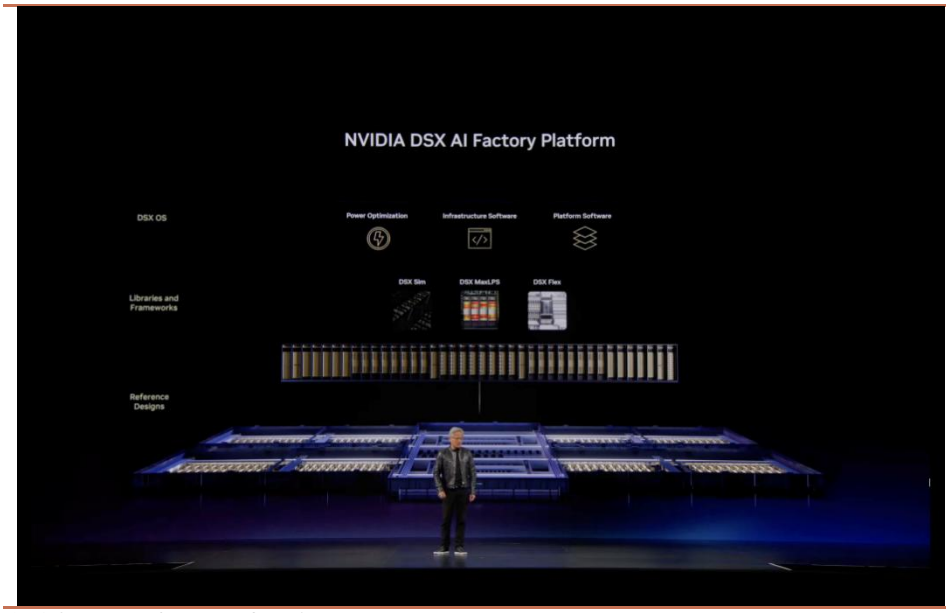


资料来源：英伟达，招商证券

Nvidia DSX 是构建和运营高效且盈利的 AI 工厂的蓝图和参考设计。它始于 DSXM（DSX 模块）。通过 DSX SIM Omniverse 蓝图，合作伙伴可以在动工建设前，设计并验证一个 Nvidia Vera Rubin AI 工厂。他们在数字孪生中规划布局，模拟电力与冷却系统，设计网络，验证每一项集成，并测试每一个变更。当工厂在 DSX 上启动时，操作系统将接管并负责基础设施的配置、运行、监控和修复，将所安装的系统转变为值得信赖、多租户且具有弹性的 AI 就绪容量。今天的 AI 工厂电力过度配置高达 40%，而 DSX Max Lps（Liquid Power System）让操作员能够在相同的电力预算下安全部署更多的 GPU，从而增加数十亿美元的年收入。我们突破性的 45 摄氏度热液冷技术不仅减少了水和能源消耗，还使更多的电力用于产生收益的计算。令人惊叹的动态功率分配技术可以将电力在机架间灵活调配，回收“搁浅电力”（stranded Watts），并将其输送到负载集中的机架；通过平滑电流峰值和电压浪涌，使工厂整体运行更加稳定。AI 智能体团队与 DSX Max Lps 协作，持续协调电力与冷却平衡以满足工作负载需求。DSX AI 工厂是灵活的能源资产，能够与电网协同运行。DSX Flex 可读取实时电网信号，在电网需要压力释放时动态调整工厂功耗。在本世纪末之前，将有 100GW 的 AI 工厂投入使用。Nvidia DSX AI 工厂以最高效率运行，生产出成本最低的 Token，并使电网更加稳固。

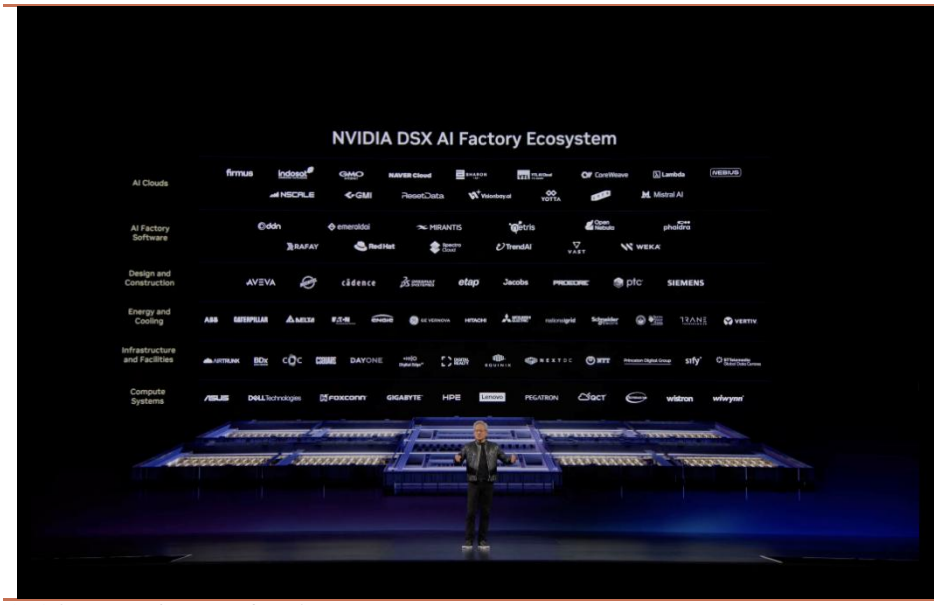
过去我展示的生态系统，重点在于英伟达的计算层和软件栈集成在他人平台或第三方库中以服务终端市场；那是一个“计算生态系统”。而现在，这是一个“AI 工厂生态系统”。这是位于你们所有人下游、而处于我上游的生态，我们共同构成了一个完整的产业链。归根结底，英伟达不仅是在制造 GPU 或系统，我们是在帮助客户构建这些极其复杂的 AI 工厂和 AI 基础设施。每一座 1GW 级别的工厂，投资额已从 200 亿、300 亿美元攀升至 500 亿、600 亿美元，不久的将来每 GW 投资将达到 800 亿甚至 1000 亿美元。投资 1000 亿美元建设 AI 工厂，必须确保一次性成功并立即投入运行，因为资本成本和复杂性都是不可思议的。正如你们所见，我们过去是在计算机内设计芯片，然后模拟系统；但今天，正如你们刚才看到的，一切都是在 Omniverse 中构建的。我与各位在 Omniverse 上合作已久，如今梦想成真，使我们能够在动工建设和投入资本之前，就在数字框架、数字模拟器和数字世界中构建这些庞大的系统。

图 5：英伟达 DSX AI 工厂平台



资料来源：英伟达，招商证券

图 6：英伟达 DSX AI 工厂生态



资料来源：英伟达，招商证券

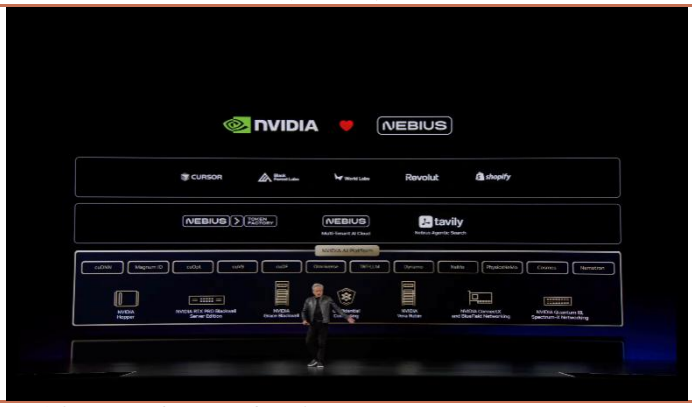
这是我们的生态系统，我们称之为 DSX。RTX 面向我们的 GPU，DGX 面向我们的系统，而 DSX 则代表基础设施。正是因为我们在整个技术栈——包括系统和软件——上的工作，我们才能够与小型公司合作，使它们成为世界级的 AI 云。每一个我即将展示的公司，不久前还都是小公司。如今 CoreWeave 的估值已达到 500 亿至 700 亿美元，且增长极其迅速。最近我们与 Nebius 合作，他们也在快速增长。这些云服务商拥有令人难以置信的客户：软件编码公司 Cursor、Black Mountain Labs、Image Generation、World Labs、World Foundation、Model、Revolut 以及 Shopify。此外还有 Nscale 及其客户。英国电信也在使用我们的 AI 云；谷歌正在利用 Thinking Machines 和 Frontier Labs 的技术。这令人非常振奋。韩国有 Naver Cloud、韩国银行、现代汽车；印度有 Yoda；新加坡有在澳大利亚建设数据中心的厂商。总之，AI 将无处不在。每一家公司都将被它驱动，每一个地区都将建设它。无论是印度的 Endoset，还是在中国台湾的 GMI，这些

都是不可思议的公司和机会。但它们都需要计算底座。这正是英伟达成名的地方：我们的硬件、软件、库，以及我们与全球第三方开发者生态的连接，使得任何人都可以建立一个 AI 云。然而，AI 云是如此复杂。

图 7：CoreWeave 与英伟达合作

资料来源：英伟达，招商证券

图 8：NEBBIUS 与英伟达合作

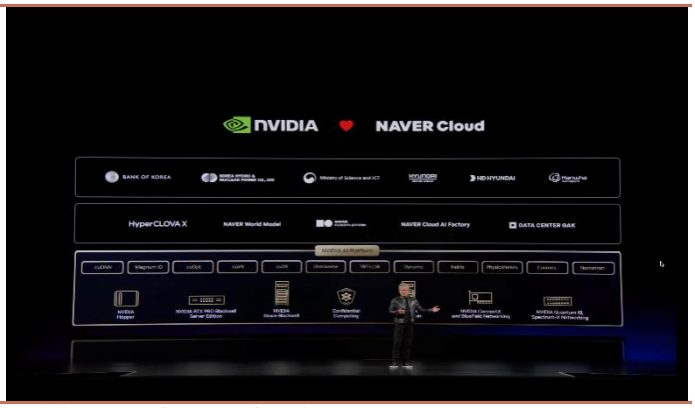


资料来源：英伟达，招商证券

图 9：NSCRLE 与英伟达合作

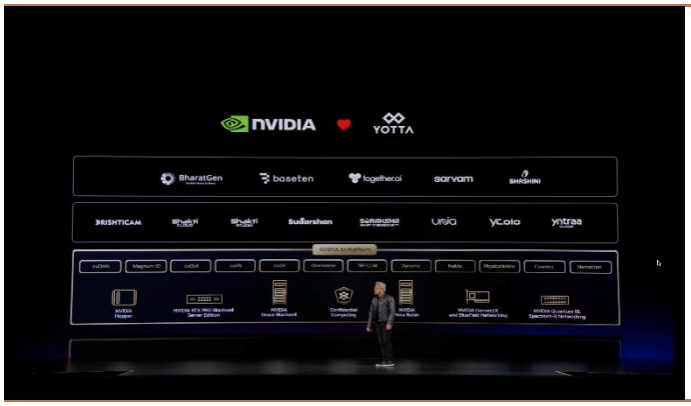
资料来源：英伟达，招商证券

图 10：NAVER Cloud 与英伟达合作



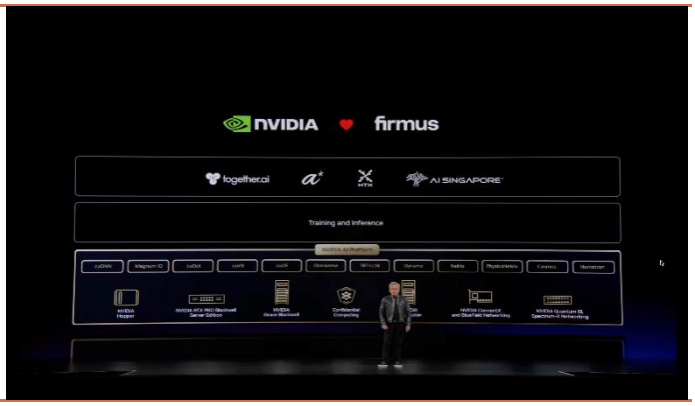
资料来源：英伟达，招商证券

图 11：YOTTA 与英伟达合作



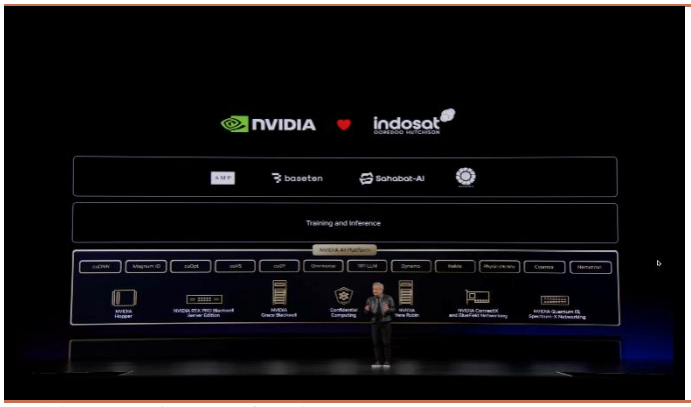
资料来源：英伟达，招商证券

图 12：firmus 与英伟达合作



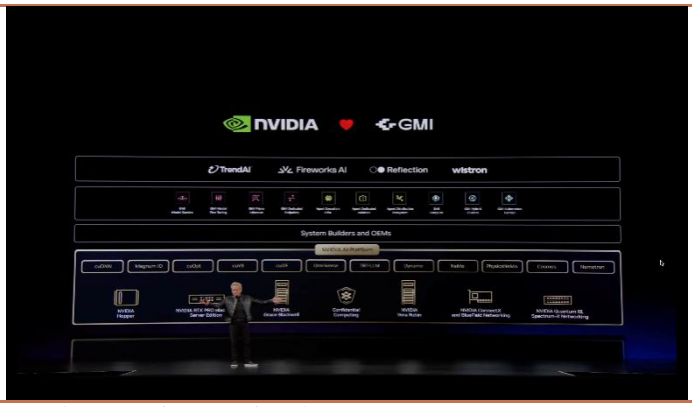
资料来源：英伟达，招商证券

图 13：indosat 与英伟达合作



资料来源：英伟达，招商证券

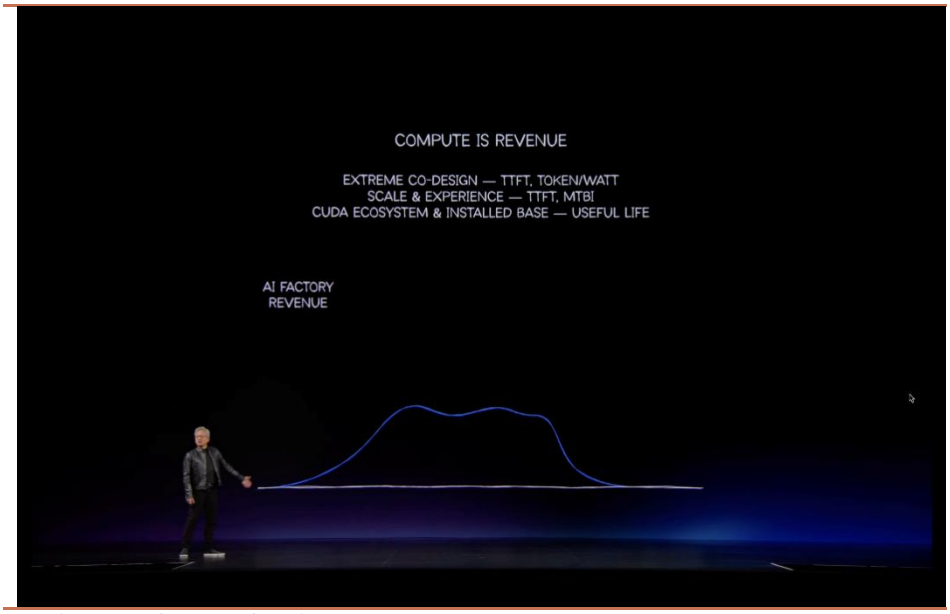
图 14：GMI 与英伟达合作



资料来源：英伟达，招商证券

现在，这不再仅仅是计算机科学层面的版本，它是资产层面、资本层面的版本。这就是我之前展示的内容——一个巨大的工厂。仅仅具备这种能力是不够的，这就是为什么英伟达已成为一家 AI 基础设施公司，在帮助客户建设和部署 AI 工厂方面变得极其专业。其根本原因在于：计算即营收，计算即利润，而收入和利润的缺失即是损失。因此，必须认识到，当一个 AI 基础设施上线时，其过程可能迅速，也可能耗时；其吞吐量可能很高，也可能很低；其弹性和可靠性可能优异，也可能糟糕；其使用寿命可能长，也可能短。因为这代表着 500 亿到 1000 亿美元的投资。这条性能曲线至关重要。这正是为什么英伟达能成为如此出色的合作伙伴，因为我们提供的是全集成能力。我们不仅仅是在 PowerPoint 演示文稿上谈论愿景，我们创建了整个基础设施，将所有组件连接在一起，并投入数十亿美元进行自我验证，确保一切运行顺畅。最终的结果是，我们极大地缩短了 Time to first token 和 Time to first inference。

图 15：计算即营收，计算即利润



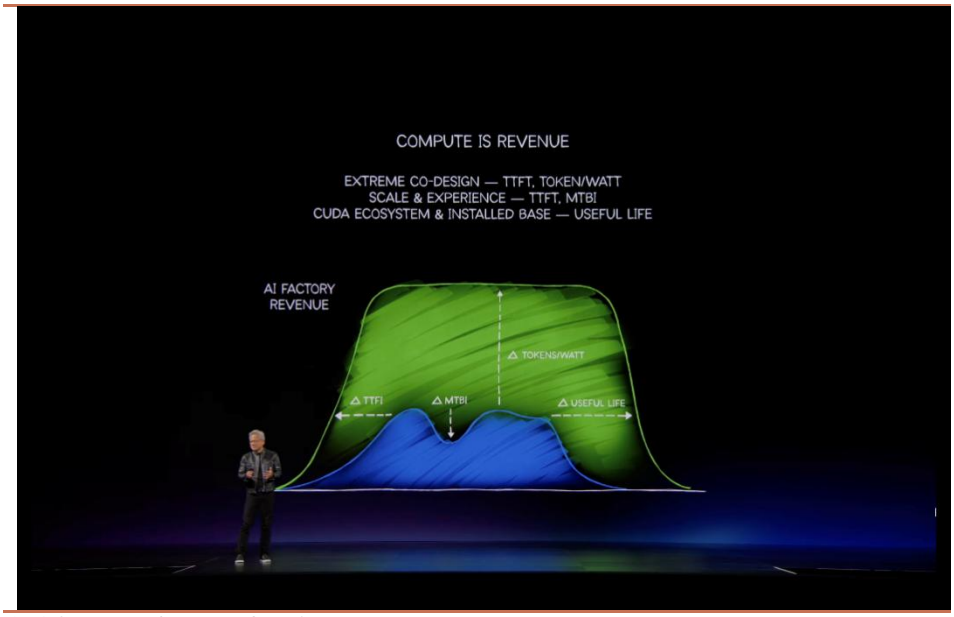
资料来源：英伟达，招商证券

第二，我们的训练启动速度更快。更重要的是，我们的“单位功耗吞吐量”，即我们每瓦特的 Token 产出是世界顶级的。原因在于我们集成了所有环节，从底层开始进行设计，模拟整个系统，并采用极致的“协同设计”，就像我刚才向大家展

示的 Vera Rubin 机架一样。一切设计皆为实现这种不可思议的吞吐量。如果你的数据中心或工厂拥有 1GW 的电力，这就是你的上限。1GW 意味着 1GW，这是你能做到的全部发电量。如果你拥有 1GW 的电力，那么“每瓦吞吐量”就是你的收入，因为每一个 Token 都是盈利的，每一个 Token 都是收入。这就是未来。计算即收入，性能供给即收入。仅仅因为芯片价格便宜就选择错误的架构是行不通的，这在商业逻辑上不合理。你必须确保你的基础设施能够持续贡献收入。买得越多，产出越多，因此 Token 供给至关重要。

最后，第三点是可靠性。如果你有机会参观这些数据中心，你会发现那里有数以百万计的线缆和复杂的运动部件。要使所有计算机和谐运转，实现高可靠性极难。我们已经在极大规模下运行了很长时间，这种经验至关重要，关于“平均中断间隔时间”（MTBF）的指标极其关键。此外，这非常艰难的一点是：这些系统的软件在不断演变。四年前，处于 Hopper AI 时代，技术完全不同；六年前，处于 Ampere AI 时代，技术也完全不同。我们从 CNNs 开始，接着是 Transformer，然后 Mixture of Experts，现在我们谈论的是智能体系统。每一代，甚至每隔几个月，软件行业就会出现新技术。如果你的架构不够灵活，如果你的生态系统不够丰富，那么这条增长曲线就不会长久。但我可以断言，英伟达的系统正遍布全球。软件开发者从 Nvidia CUDA 开始，因此其资产的生命周期、生态系统和使用寿命将会长得多。本质差异在于成本，你可以将其视为收入，但收入的另一面是成本。如果资产寿命长，TCO 就低。这就是当计算能力增强时，买得越多，产出越多的意义所在。各位正在与我共同见证这一点。在中国台湾，你们的工厂全负荷运转，因为每个人都意识到，有用的 AI 已经到来，盈利的 AI 已经到来。计算需求高得惊人，计算需求成为了唯一的瓶颈。所以，让我们继续全力以赴，帮助世界建立起遍布全球的 AI 工厂。

图 16：AI 工厂营收三大关键点



资料来源：英伟达，招商证券

2、Vera Rubin：已进入全面量产阶段，极端协同设计重塑 Token 成本曲线

Vera Rubin 目前已全面进入量产阶段。我们为 Vera Rubin 建立的供应链规模是

Grace Blackwell 的两倍。这令人难以置信。过去组装一台 Grace Blackwell 机架需要两个小时，而现在仅需 5 分钟。不仅产能更高，吞吐速度也快得多，我们需要这一切来支持市场需求。这个生态系统非同凡响，为了支持 Grace Blackwell，数百万平方英尺的工厂空间已投入使用，目前正在扩产中。

Vera Rubin 现已全面进入量产。大型语言模型负责生成答案，而 AI 智能体则负责执行任务，处理 Agentic AI 是一项完全不同的挑战。智能体需要进行观察、推理、规划、使用工具，并同时管理大规模上下文，即“平衡”工作记忆与长期记忆。它们会根据需求实时生成子智能体（sub-agents）专家。Vera Rubin 是一个多机架规模的计算 Pod 系统，专为处理 Agentic AI 而构建，目前已全面投产。整个供应链在制造、自动化与编排上的配合，令人叹为观止。

我们的旅程始于第一台 AI 超级计算机 Nvidia DGX1。在过去的十年里，我们将每一块芯片和系统推向极限——从 Pascal 架构到第一代 NVLink，再到 Grace Blackwell（第一台机架级 AI 超级计算机），直至现在的 Vera Rubin——这是首台为智能体时代构建的多机架 Pod 级超级计算机，而这一切始于台积电。构成 Vera Rubin 的七款新芯片，经历了数百道加工步骤，采用 3 纳米制程，集成了 CoWoS-R 和 CoWoS-L 封装技术，并搭载了来自美光、SK 海力士和三星的 HBM 内存。Vera Rubin 计算板在一块板卡上集成了超过 1.8 万个组件和 6 万亿个晶体管。Vera Rubin NVL72 负责执行思考、Prompt 处理、上下文理解、推理与规划。

紧接着，是一种全新的模块化计算链路，采用了流线型的 PCB 中板设计。Super Chips 连接 X9 Super Nyx 和 BlueField 4 DPU，且在工厂级规模上完全抛弃了线缆设计，以保证极高的韧性。该系统包含 18 个计算托盘、9 个热插拔 NVLink 交换机托盘，以及新型高效 manifolds 和承载超过 5000 安培电流的液冷母线（bus bars）——这相当于 20 辆电动汽车全速加速所需的电流。所有这些 130 万个组件，共同构成了第三代 MGX 机架设计。在此，我要向微软在 Vera Rubin MVL72 工程实现上的卓越工作表示祝贺；同样祝贺 Dell 和 CoreWeave 在建设 Vera Rubin 工程路径上的贡献。

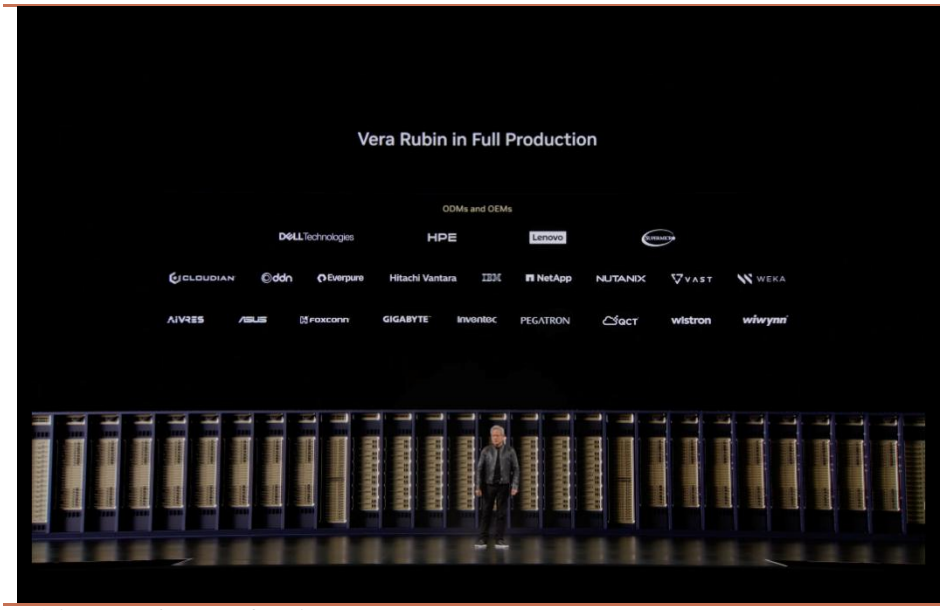
随后是 Vera CPU 机架，单机架内集成 256 颗液冷 CPU，用于编排模型、进行内存调度以及在富士康和 Quanta 的工厂中运行工具。Groq 3 LPX 架构也随之成型，它在 16 个托盘中集成了 256 颗 Groq 3 LPU，拥有每秒 40PB 的 SRAM 带宽，旨在实现超低延迟。如果说 NVL72 负责以最高吞吐量生成 Token，那么 Groq LPX 则负责以最低延迟进行生成。Vera BlueField 4 STX 实现了 AI 内存存储处理的加速，通过 BlueField 4 连接内存、存储及硅片内安全技术。此外，Nvidia Spectrum X 以太网光子技术作为全球首款 200Gbps 共封装光学以太网交换机也已集成。我们还采用了台积电的 COUPE 芯片规模封装工艺，利用磷化铟材料制造超大功率激光染料。

Vera Rubin 5 系统连接了机架级计算，是专为 AI 智能体打造的超级计算机。这背后是 150 家中国台湾供应链合作伙伴、数百万平方英尺的工厂空间、以及数百个站点，芯片、封装、系统与数据中心被共同推向了尺寸、功耗和规模的极限。这就是我们所说的“极端协同设计”。我们与中国台湾共同完成了这一切，共同为 AI 时代重塑了计算。Vera Rubin 不仅仅是为了 AI 而生，也不仅仅是为了运行 AI。Rubin 是为运行“智能体”而生的，这是一个智能体系统。想象一下它的复杂性，这正是智能体作为计算机科学领域最新突破的原因。智能体在实现其潜力和应用价值的过程中，花费了多年时间。顺理成章，驱动它的计算机必须是世界上最先进的。这就是 Vera Rubin NVLink 72，这是 Groq LPX，在下一次 GTC 大

会上，我将向大家分享更多相关信息，这是 Vera CPU 机架，集成 256 颗 CPU，全部采用液冷技术。这是 Vera BlueField 存储处理系统以及安全系统，当然，还有我们的 Mellanox 网络技术，这是全球首个 CPO 方案。这就是 Vera Rubin，令人难以置信的技术集大成者。

回顾过去，当我们构建 Hopper 架构时，预训练是当时最重要的应用和工作负载。后来当我们着手研发 Grace Blackwell 时，每个人都对我说："Jensen，你知道，英伟达在预训练方面非常出色，但推理其实很简单。"人们过去常说推理很简单，他们也能做到。但正如你们所知，推理直接等同于金钱，而专家混合模型（MoE）非常复杂。要在保持极高响应速度、快速交互性的同时，还要兼顾高吞吐量，是一项极难的任务。这正是我们创建 NVLink 72 的原因。如今，英伟达的 Token 生成成本是世界上最低的，不仅降低了 10%，而是实现了数量级的跨越。这一切都是因为我们进行了"极端协同设计"，是因为我们深刻理解了计算模式和推理的计算范式，从而成功创造了 NVLink 72。现在，Vera Rubin 已经超越了推理本身，它进入了智能体系统的推理阶段。Vera Rubin 设计中不再有杂乱的线缆、软管或风扇。回想上一次我向大家展示时，系统内到处是线缆，尽管那看起来很酷，但现在我们采用了 PCB 中板设计，将两侧直接连接。过去需要两个小时的装配时间，现在仅需五分钟。Vera Rubin 在可靠性和韧性方面表现将是超乎寻常的。

图 17：英伟达 Vera Rubin 全面量产



资料来源：英伟达，招商证券

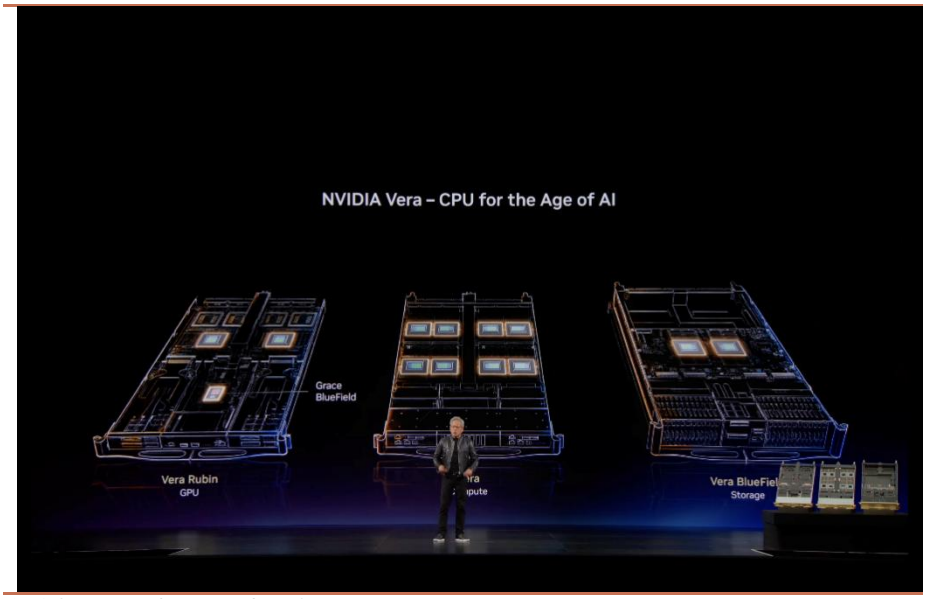
这是我们的 Vera CPU 托盘，这是有史以来建造的最先进的 CPU。这是我们"存储秘籍"：两颗 Vera CPU 与 CX9 芯片及海量的软件。这是我们全新的 LPX LPU30，这套 Groq 系统专为超低延迟推理而设计。Vera Rubin 提供极高的吞吐量，并可通过 NVLink 72 进行扩展，如果你想进一步延伸，还可以添加 Groq LPU。这里是 Vera Rubin NVLink 交换托盘。中间部分的交换机具有革命性意义，这得益于 Vera Rubin、NVLink 72 以及我们发明创造的 NVLink 交换机。此外，还有用于 Scale-out 的以太网交换机。令人惊叹的是，我们为 Grace Blackwell 引入了两个系统，而今天，英伟达已成为全球最大的网络公司。我对我们的网络团队深感自豪，这是支撑我们一切工作的核心驱动力。

3、Vera CPU：专为 Agent AI 设计，性能优于 x86 处理器

现在，我将讨论我们将参与的下一个重要行业——CPU。在 AI 时代，过去所有的 CPU 都是为人类用户而设计的——人类是用户，人类是租赁者。我们生活在以“秒”为单位的节奏里。然而，在云端租赁 CPU 时，核心越多，租赁成本越高。传统的 CPU 架构与应用场景与智能体完全不同，智能体是“迫不及待”的。它们不以秒为计量单位，而是生活在“纳秒”的世界中。当智能体使用工具时，它需要响应时间尽可能快；当它访问数据库时，它必须立即获得反馈。智能体每多等待一刻，都会影响接下来的执行进度。因此，实现 CPU 的极低延迟和高度交互性至关重要。

为此，我们为 AI 时代创造了 Vera CPU。在我们的系统中，它承担着三重角色：第一，在 Vera Rubin 中负责思考和编排，机架内集成了两颗 Vera CPU，负责管理 GPU、处理 KV 缓存以及处理机架内的所有软件调度；第二，它是 Grace BlueField 的核心，用于安全与隔离；第三，Vera 计算单元负责 Harness、AI 模型编排、工具调用及数据库访问。Vera BlueField 拥有全球最快的存储服务器系统。这至关重要，因为智能体对内存的访问速度极快。存储服务器与 CPU 现在构成了数据中心中最关键、也是最昂贵的部分。AI 工厂的经济学本质在于 Token，Token 是在这里生成的，因此你需要制造和生成尽可能多的 Token。Vera CPU 从零开始构建了全新的架构，这是世界上从未见过的 CPU，它是专为智能体打造的。

图 18：英伟达 Vera CPU 托盘架构

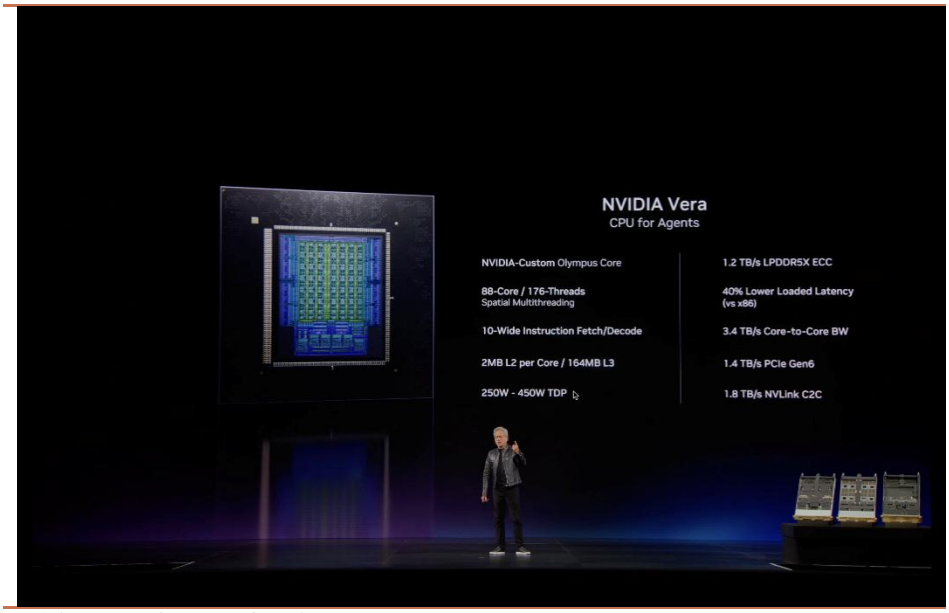


资料来源：英伟达，招商证券

Vera CPU 有四个关键要点。第一，Vera 的“每时钟周期指令数”（IPC）必须极其出色，因为我们需要极短的延迟和处理时间。单线程性能必须达到世界级水平——不仅是吞吐量，而是单线程性能。Vera 每时钟周期可 fetch、解码并执行 10 条指令，这是全球最高的 IPC。第二，带宽至关重要。进出 CPU 的数据移动必须达到顶级。Vera 是长期以来首个达到 reticle limits 的 CPU，其 fabric 连接了所有 CPU 核心，以光速运行，速度达每秒 3.6TB。没有芯片间的边界交叉延迟，因为我们实现了所有核心的高速协同。Vera 的横截面带宽表现惊人。它是首个支持 PCIe Gen 6 的 CPU，也是首个具备 LPDDR5X（DDR5）内存支持的

CPU，带宽高达每秒 1.2TB。其内部带宽是外部高性能 CPU 的两到三倍，且带宽密度均为世界级。我们要记住，智能体数量将远超人类。地球上只有 10 亿人类，但未来会有数十亿个智能体。它们使用 CPU 时极其缺乏耐心，因为它们依赖的 GPU 资源过于昂贵且珍贵。因此，这些 CPU 不仅要高性能，还要具备极高的能效，以便我们能将尽可能多的 CPU 塞进工厂，而不占用用于生产 Token 的电力。Instructions per clock（单线程性能）、带宽密度、总带宽以及能效，这四项特质定义了 Vera。

图 19：英伟达 Vera CPU 性能



资料来源：英伟达，招商证券

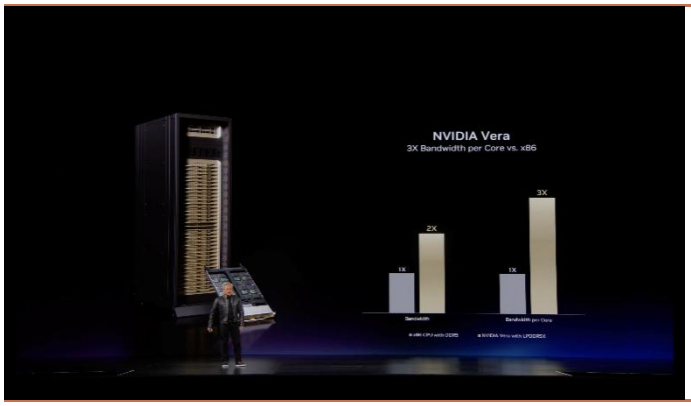
当你将其与最高性能的 x86 处理器进行对比时，Vera 的真实单线程性能表现完全是不同的。通常 CPU 提升 5%或 10%的性能已是了不起，但这种跨越式的性能提升闻所未闻，这就是英伟达 Vera。

Agentic AI 改变了 CPU 的角色。CPU 现在是指挥家，而 GPU 是交响乐团。传统的 CPU 构建于另一个时代，追求单插槽核心最大化，通过每小时租赁的方式进行虚拟化。而在智能体时代，CPU 已成为 GPU 利用率的瓶颈，直接影响 Token 的吞吐量、延迟和用户体验。Nvidia Vera 是专为“智能体循环”打造的 CPU，它结合了英伟达定制的数据中心 CPU 核心与可扩展一致性架构，在性能、内核数与带宽之间实现了完美平衡，从而最大化 AI 工厂的输出效率。Vera 的核心是 Nvidia Olympus Core，专为现代数据中心工作负载、分支密集型 Python 运行时、工具调用及沙盒代码执行而构建。每个内核都经过吞吐量调优。其神经分支预测器每个周期可评估两个分支，10 路解码引擎（10-wide decode engine）在每个周期内引入更多工作，大型乱序执行引擎（out-of-order engine）保证了指令流的顺畅，先进的预取器配合新型图引擎能够预测下一条数据路径。

Vera 是首个使用 LPDDR5X 内存的 CPU，它能在不牺牲带宽的前提下同时纠正多个错误。相比 x86 架构，Vera 将峰值内存延迟降低了 40%，通过检索分析和沙盒执行，确保内核始终处于饱和工作状态。英伟达第二代可扩展一致性架构在单片网格上统一了所有 88 个 Olympus 核心，并为内存和 I/O 设置了独立的晶粒（dyes）。核心并未跨三重片（triplets）分配，这使得核心间通信速度较传统 CPU 快 50%。内存一致性的 NVLink 芯片间互连将 GPU 直接连接到 fabric，并可通过 NVLink 芯片间互连实现 Vera 的多插槽扩展。Vera 的 agentic sandbox

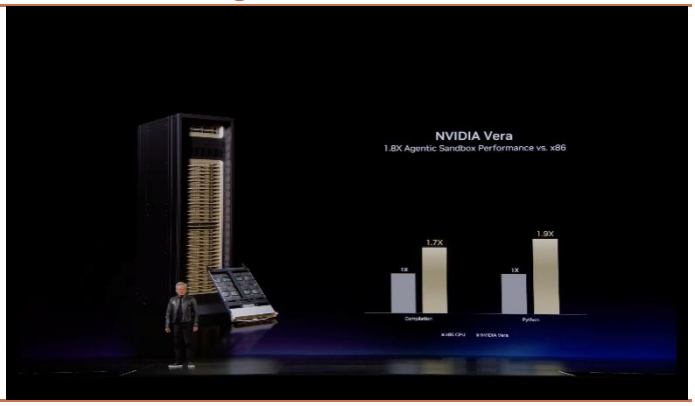
性能达到了 x86 CPU 的 1.8 倍。

图 20：Vera 的每核心带宽对比 x86 架构



资料来源：英伟达，招商证券

图 21：Vera 的 Agentic Sandbox 性能对比 x86 架构



资料来源：英伟达，招商证券

图 22：英伟达 Vera 对标 AMD x86-64

资料来源：英伟达，招商证券

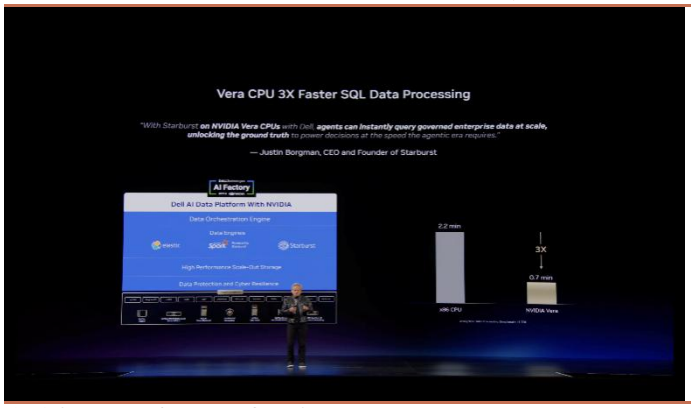
Vera BlueField 4 STX 为上下文内存、AI 存储、计算及网络存储提供了动力。Vera 是智能体时代的 CPU，它将成为我们新的主要增长驱动力。目前的初步评估结果非常出色。此外，Grace 和 Vera 是目前全球 AI 领域“资质最深”的 CPU。因为每一个数据中心、每一个云、每一个与英伟达合作 AI 项目的企业，都已经通过了 Grace 的验证，整个软件栈也已为 Grace 进行了优化。随着企业继续推进，它们自然会转向 Vera。Vera 将成为全球优化程度最高的智能体 CPU，因为它将与 Vera Rubin 系统协同工作。回顾 Grace Blackwell 的过渡期，最大的风险是从外部 x86 CPU 转向 Grace Blackwell，尽管那次转型极具挑战，但我们凭借卓越的执行力完成了任务。现在，Grace 已与 Grace Blackwell 融为一体。当人们谈论 Blackwell 时，就是在谈论 Grace Blackwell，因为它已无处不在。每个企业的软件栈和安全栈都已为其进行了优化，而现在，Vera 来了。

让我们看看一些性能数据。虽然“加速比”（Speedups）是一个指标，但加速 SQL 是极具挑战性的。在 CUDA 或 OpenGL 出现之前，SQL 是由 IBM 发明的，它是当今地球上最重要的结构化数据库引擎。现在的 SQL 运行速度提升了 3 倍，

不是 10%，不是 25%，而是 3 倍。这不仅是性能提升，更是实时性。

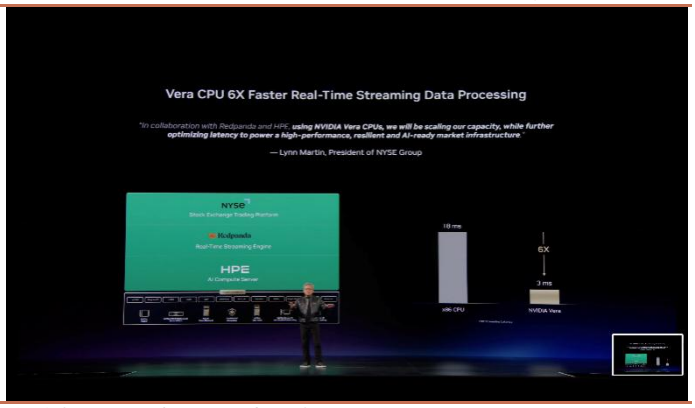
下一项是实时流处理。记住，你的 AI 不仅仅是读取文档。你的 AI 需要时刻关注遥测数据，尤其是在工厂内部或证券交易所。你需要连续监测这些遥测数据。突发的数据流进入 CPU，这就是 Vera CPU 在运行。我们与纽约证券交易（NYSE）总裁 Lynn Martin 合作，为纽交所实现了实时流处理。凭借 Vera 的带宽和单线程指令执行能力，系统速度提升了 6 倍。

图 23：Vera CPU SQL 数据处理速度提升 3 倍



资料来源：英伟达，招商证券

图 24：Vera CPU 实时流数据处理速度提升 6 倍



资料来源：英伟达，招商证券

当你们讨论 GPU 时，常会提到“X 因子”，但在与 CPU 关联的真实工作负载中，提及 X 因子是非常罕见的。我对我们的团队感到无比自豪。我们拥有非凡的路线图，但最令人兴奋的是，几乎每个人都在支持 Vera。

这是 Vera 的开篇，它开启了一个全新的市场——智能体是一个全新的工作负载。我们过去为人类构建 CPU，现在我们需要为智能体构建 CPU。它们的属性不同，为什么要沿用相同的架构？这开启了一个前所未有的新市场。它不会取代旧市场，但 CPU for agents 这个新市场必然会更大。原因在于，智能体的数量将远超人类，而且智能体非常“没耐心”。所以，这就是 Nvidia Vera CPU。

图 25：Vera 打开新市场

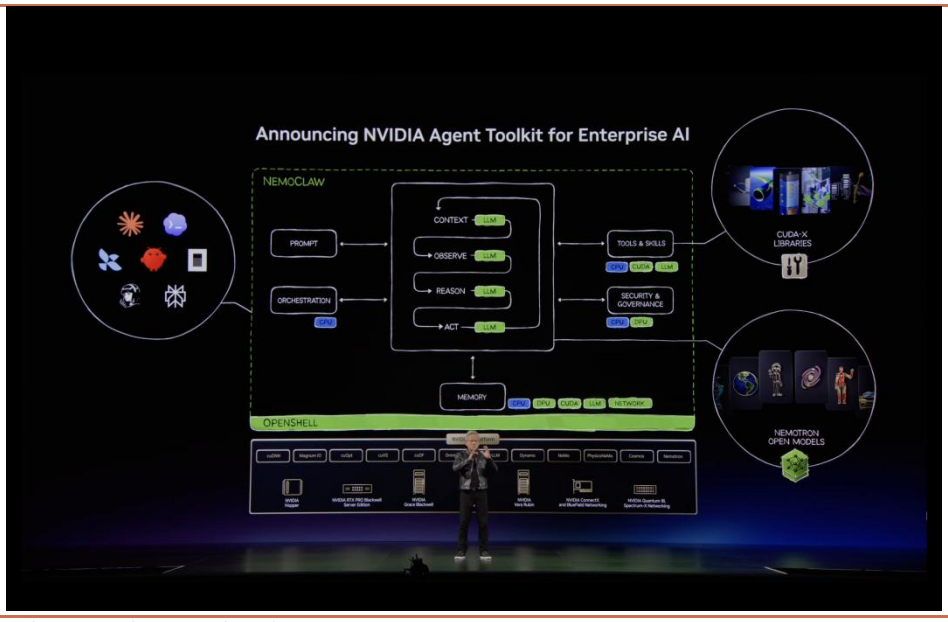
资料来源：英伟达，招商证券

4、AI PC：携手微软重新定义 PC，构建本地化智能体计算终端

在智能体时代，每家公司都将逐步成为智能体公司。企业内部将运行大量智能体，用于软件开发、运营管理、数据分析、客户服务、研发设计和自动化流程。为了让企业安全、高效地构建智能体，NVIDIA 推出了面向企业 AI 的智能体工具包。

黄仁勋将企业构建智能体系统所需的核心要素概括为四类：**1）模型。**企业需要足够聪明、足够快速且成本可控的大语言模型。**2）执行框架。**执行框架负责对模型、工具和任务流程进行编排。**3）工具和技能。**智能体需要调用工具完成具体任务，这些工具可以包括 CUDA-X 库、企业软件、仿真器、数据库、代码工具和行业专用系统。**4）运行时。**运行时相当于智能体时代的操作系统，负责安全、权限、身份、隐私、沙盒隔离和执行环境管理。NVIDIA OpenShell 是其中的重要运行时组件。它为企业运行智能体提供安全沙盒环境，确保智能体的权限、身份、隐私和安全策略得到约束。OpenShell 可运行在云端、本地数据中心、PC 或边缘设备上，并已面向开源生态推进。

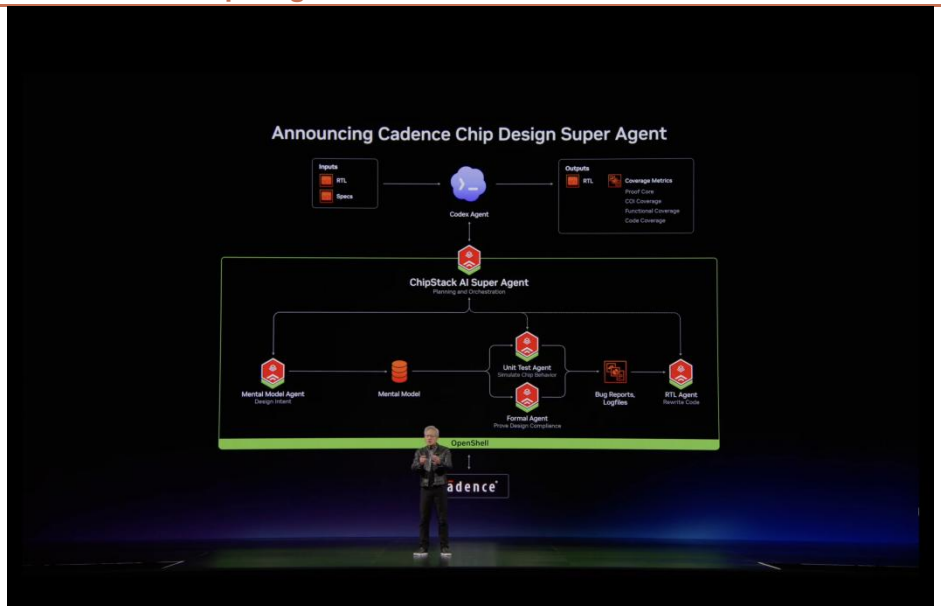
图 27：NVIDIA 企业 AI 工具包



资料来源：英伟达，招商证券

芯片设计是智能体应用的重要场景。NVIDIA 与 Cadence 合作构建芯片设计 SuperAgent，用于提升 RTL 验证、仿真、测试平台生成、回归测试和调试效率。在传统芯片设计流程中，RTL 验证需要大量工程师、计算资源和测试流程，周期通常以周为单位。通过 Cadence ChipStack、NVIDIA Nemotron、OpenShell 以及 Cadence 的仿真和形式验证工具，智能体能够自动完成代码生成、测试创建、仿真执行、缺陷识别和 bug 修复。该系统将部分验证工作从数周压缩至数小时，验证周期加快超过 40 倍。这一案例体现了智能体并非简单的聊天机器人，而是能够嵌入专业工程流程、调用行业工具并完成复杂任务的生产力系统。对于 EDA、工业软件、企业软件和研发工具厂商而言，智能体不是威胁，而是新的增长机会。具备行业知识、专有数据和工具链的企业，可以将其能力封装为 SuperAgent，并以服务形式提供给客户。

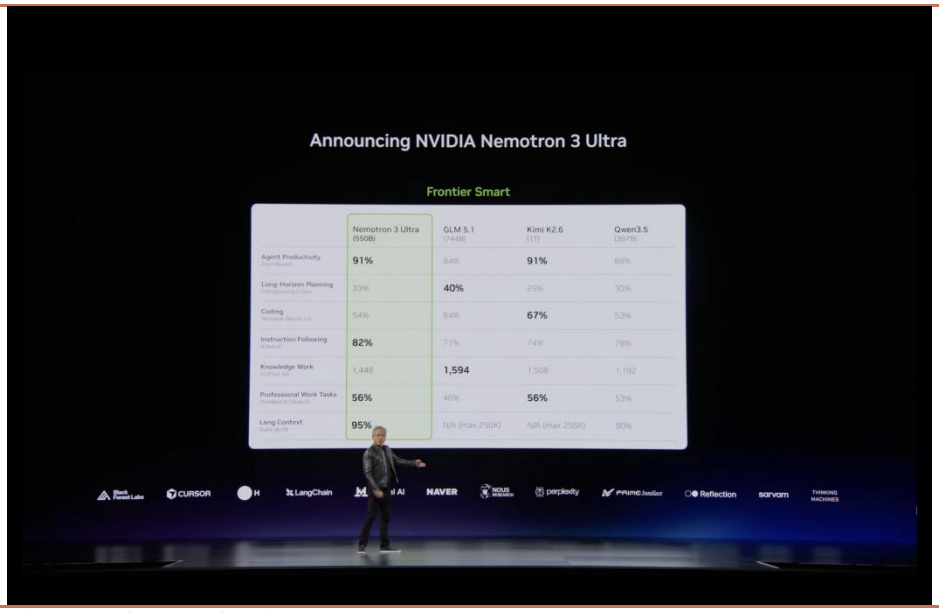
图 28：Cadence SuperAgent：智能体在芯片设计中的典型落地



资料来源：英伟达，招商证券

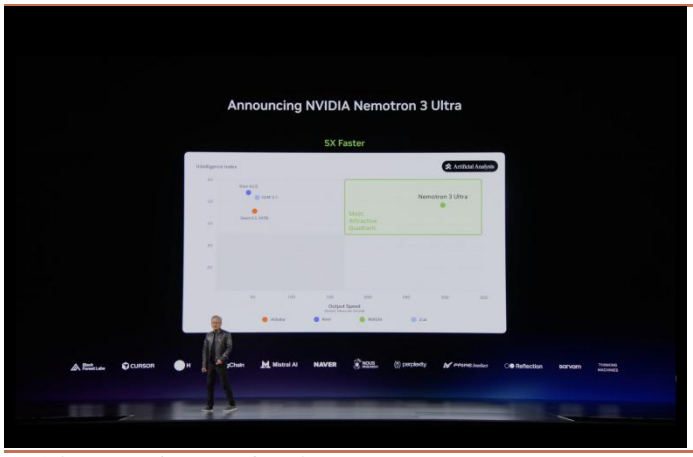
NVIDIA 发布 Nemotron 3 Ultra，作为其下一代开放模型。该模型不仅开放模型本身，还开放训练数据和训练方法，目标是让企业和开发者能够在此基础上进行后训练、增强和专有化，形成适合自身业务流程的智能体模型。Nemotron 3 Ultra 采用混合架构，将 SSM 状态空间模型与 MoE 混合专家架构结合。其特点是推理速度更快、成本更低，并具备较强的长程推理、任务执行和工具调用能力。相比部分领先开放模型，NVIDIA 表示 Nemotron 3 Ultra 速度提升约五倍，成本降低约 30%。这一策略反映出 NVIDIA 不仅在提供硬件，也在构建模型、数据、工具、运行时和生态合作伙伴组成的完整企业 AI 平台。

图 29：Nemotron 3 Ultra：NVIDIA 面向企业智能体的开放模型体系



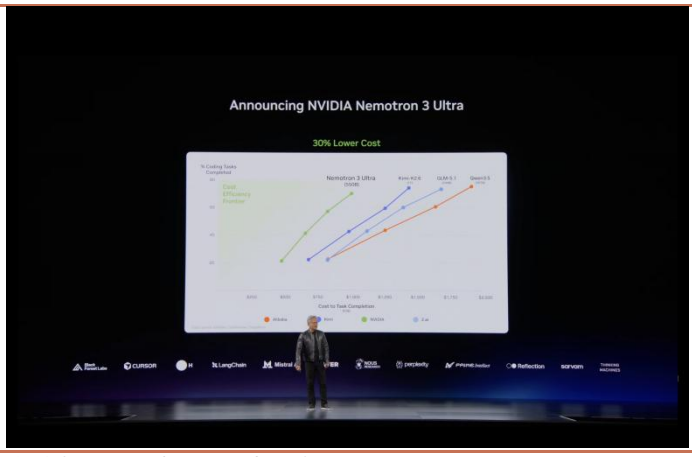
资料来源：英伟达，招商证券

图 30：Nemotron 3 Ultra 速度提升 5 倍



资料来源：英伟达，招商证券

图 31：Nemotron 3 Ultra 推理成本降低 30%

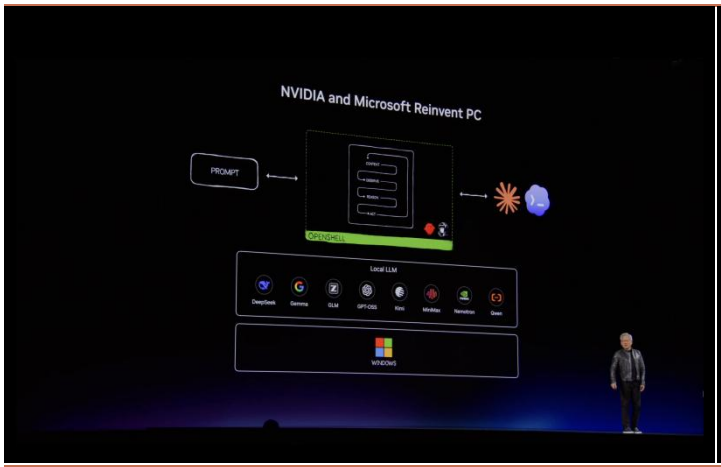


资料来源：英伟达，招商证券

NVIDIA 与 Microsoft 正在重新定义 PC。传统 PC 主要用于启动应用、点击、输入和处理文件；未来 PC 将演变为运行个人智能体的本地计算平台。智能体能够理解用户意图、读取文件、调用应用、进行研究、生成内容，并在本地或云端协同完成任务。

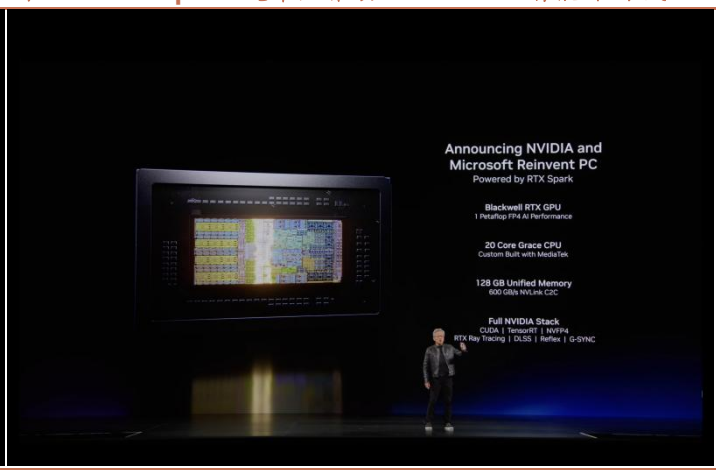
在这一背景下，NVIDIA 发布 RTX Spark 平台。该平台整合 Blackwell RTX GPU、定制 Grace CPU、NVLink、统一内存以及 Windows 智能体运行环境，目标是为创作、游戏、开发和个人智能体提供本地 AI 算力。RTX Spark 的意义在于，它让智能体可以在本地 PC 上运行，并调用本地应用和工具。例如，智能体可以在安全沙盒中调用 Rhino、Blender、Adobe Photoshop、Premiere 等软件，帮助用户完成建筑设计、图像处理、视频编辑和创意工作流自动化。未来 PC 的角色可能类似个人 AI 服务器或家庭 AI 中枢。用户可以在本地部署自己的智能体，持续连接家庭设备、个人文件、摄像头、显示器和应用生态，形成长期运行、不断学习和个性化的 AI 助手。

图 32：NVIDIA 携手 Microsoft 重塑 PC



资料来源：英伟达，招商证券

图 33：RTX Spark 发布，推动 AI PC 迈入智能体时代

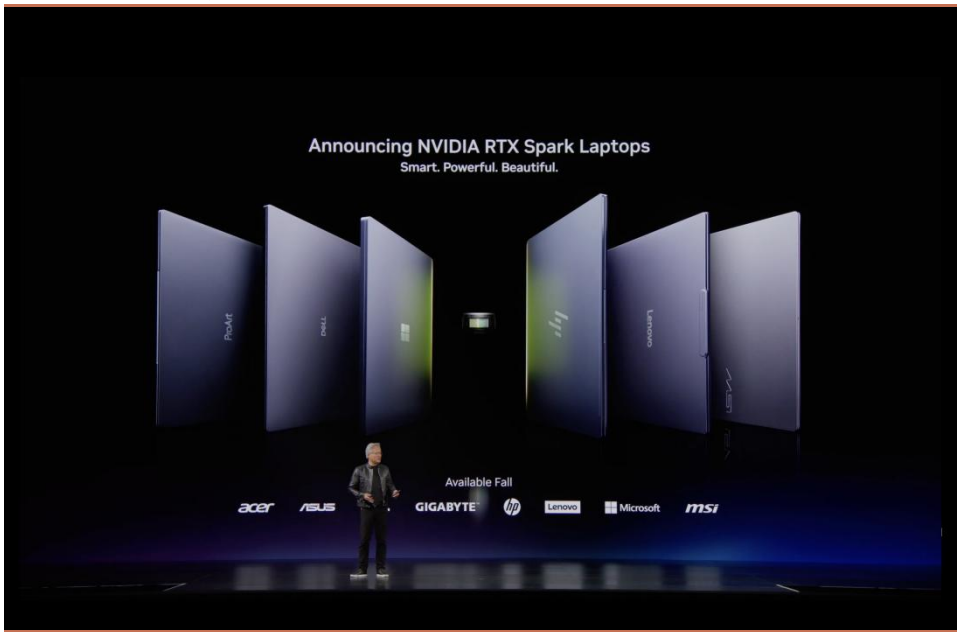


资料来源：英伟达，招商证券

NVIDIA 与 Microsoft 推出面向 Windows 的全新 PC 产品线，覆盖笔记本、台式机和工作站。这些设备具备三个核心特征：兼容 Windows、支持 CUDA、搭载 NVIDIA AI Tensor Core。

其中，RTX Spark 笔记本和桌面设备面向个人智能体和创作场景；DGX Station for Windows 则面向大语言模型开发者和智能体开发者，具备高内存、高带宽和高 AI 算力，可在本地运行更大规模模型。这一产品线的战略意义在于，NVIDIA 试图将数据中心级 AI 能力下沉至个人终端，使 PC 从传统应用入口升级为智能体运行平台。黄仁勋将这一变化类比为手机从“打电话设备”演变为智能手机的过程。未来十年，PC 的定义也可能发生类似变化。

图 34：全新 PC 产品线

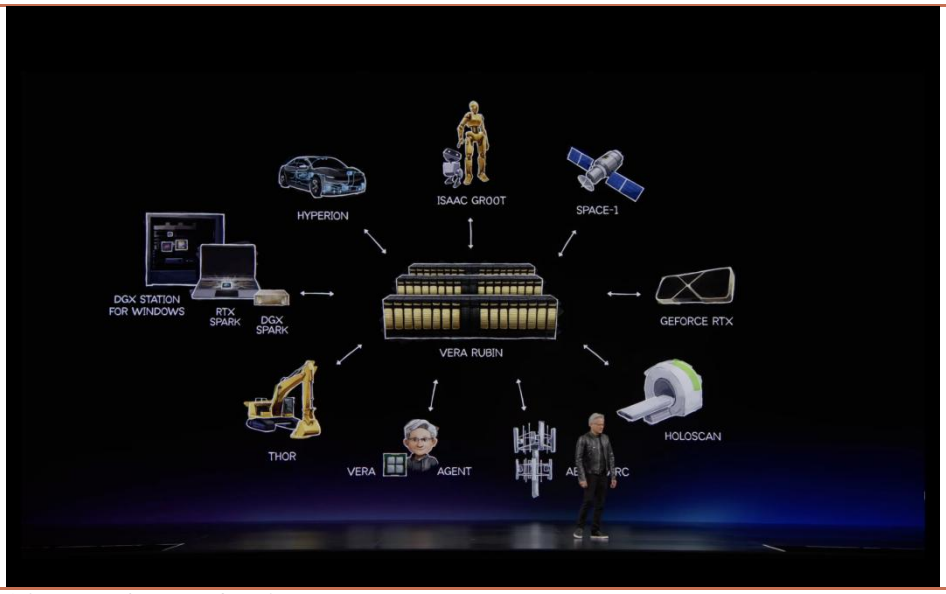


资料来源：英伟达，招商证券

5、Physical AI：从 Cosmos 模型到 Isaac GROOT 机器人开发平台，构建物理 AI 生态底座

智能体 AI 本质上是一种数字机器人。它能够理解、推理、规划、行动并调用工具。因此，同样的智能体计算模式不仅适用于云和 PC，也适用于机器人、自动驾驶汽车、卫星、通信基站、农业设备、制造设备和工业系统。在 NVIDIA 的判断中，未来将有数百亿甚至数千亿个智能体系统运行在不同设备和场景中。数据中心只是智能体计算的起点，真正的长期机会在于智能体向物理世界和边缘设备扩散。

图 35：智能体计算将扩展至机器人、汽车等

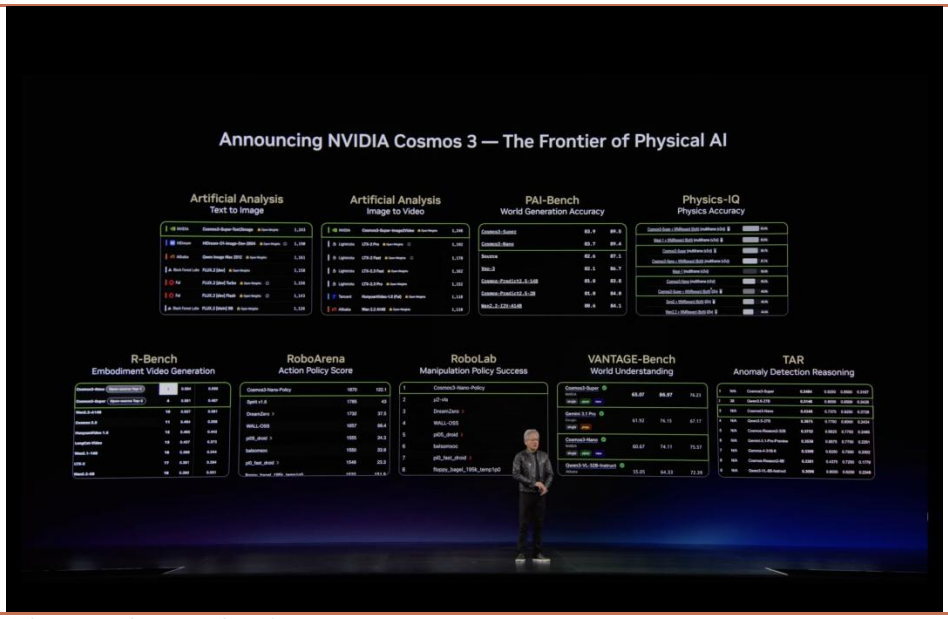


资料来源：英伟达，招商证券

在物理 AI 领域，数据是核心瓶颈。语言模型可以从互联网上学习人类写下的文本，但机器人和自动驾驶系统需要从机器视角理解世界，而现实中大多数视频数据并非第一人称视角。因此，物理 AI 需要新的数据生成方式。

NVIDIA 发布 Cosmos 3，作为面向物理 AI 的开放前沿全模态模型。Cosmos 能够处理像素、动作、声音和语言，并通过自回归 Transformer 与扩散 Transformer 结合，实现物理世界的理解、推理、规划和生成。Cosmos 可以作为视觉语言模型理解场景，也可以作为世界模型生成符合物理规律的视频，还可以作为仿真器为策略训练提供闭环环境。对于机器人、自动驾驶、工业自动化和物理 AI 开发者而言，Cosmos 相当于一个基础模型平台，可用于生成训练数据、评估策略和构建专有模型。NVIDIA 强调，过去 AI 的范式是“数据加计算得到 AI”；而在物理 AI 阶段，有了 AI 之后，计算本身也可以生成数据。Cosmos 代表了这一范式转变。

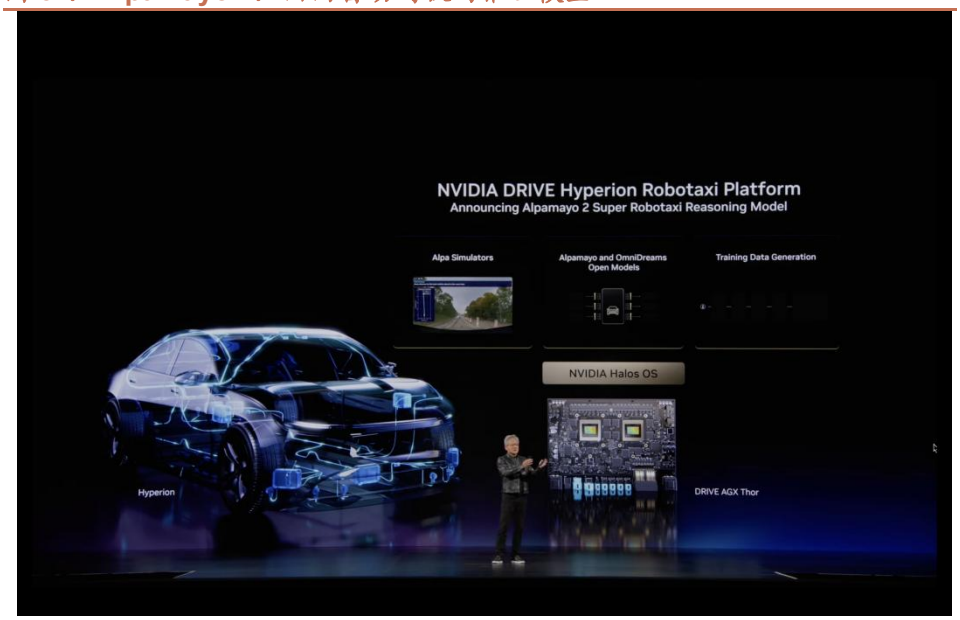
图 36：Cosmos 3：面向物理 AI 的基础模型



资料来源：英伟达，招商证券

NVIDIA 发布 Alpamayo 2，这是面向自动驾驶汽车的开放模型。该模型的目标是让自动驾驶系统具备更强的推理能力，不仅识别道路环境，还能理解交通参与者、规划行驶路径，并解释自身决策过程。Alpamayo 将运行在 NVIDIA Hyperion 平台和 HALOS 操作系统之上。NVIDIA 表示，采用或合作开发 Hyperion 平台的汽车品牌覆盖全球约 80%的汽车产量，同时其出行服务生态覆盖率也非常高。这意味着 NVIDIA 希望通过统一的硬件、运行时和模型平台，成为自动驾驶产业的重要基础设施提供商。

图 37：Alpamayo 2：面向自动驾驶的推理模型

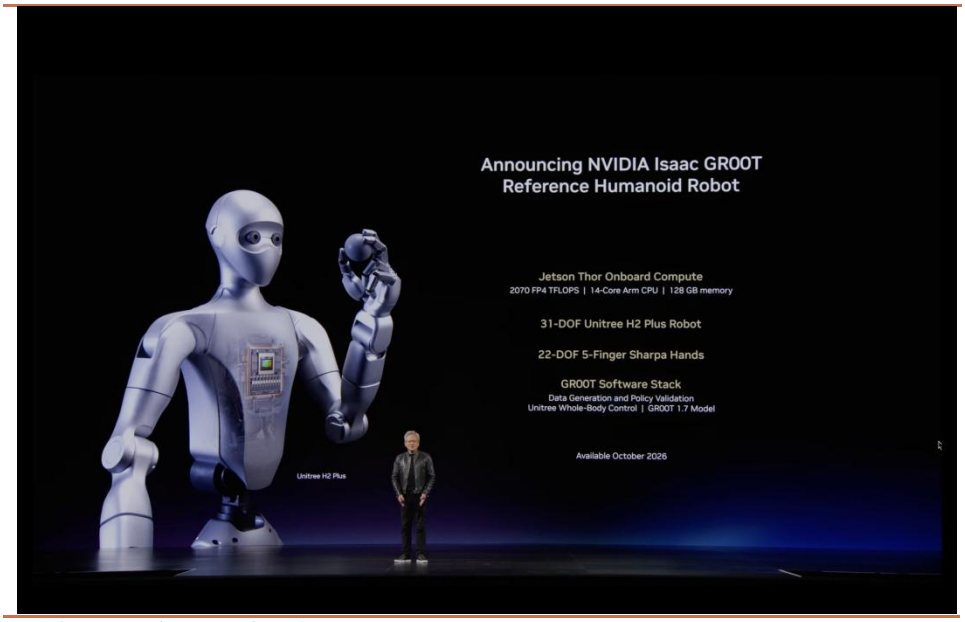


资料来源：英伟达，招商证券

NVIDIA Isaac GR00T 是面向人形机器人的软件和开发平台，覆盖模型、数据生成、仿真、训练、部署、运行时和操作系统。其目标是降低机器人研发门槛，让高校、研究机构和企业不必从零搭建完整机器人系统。Isaac GR00T 平台包括 Isaac Lab、Isaac Teleop、Omniverse、Cosmos、Isaac Lab Arena、Isaac ROS 和 Jetson Thor 等组件。开发者可以在仿真环境中搭建场景，通过遥操作捕捉示范，利用 Omniverse 和 Cosmos 生成合成数据，再进行策略训练、评估和部署。

NVIDIA 同时发布 Isaac GR00T 参考人形机器人。该机器人为完整集成的参考平台，面向高校和前沿研究机构提供。其意义类似于 PC、DGX 和自动驾驶参考平台：通过标准化硬件和软件栈，推动开发者生态形成。

图 38：Isaac GR00T：人形机器人开放平台与参考设计



资料来源：英伟达，招商证券

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

评级说明

报告中所涉及的投资评级采用相对评级体系，基于报告发布日后 6-12 个月内公司股价（或行业指数）相对同期当地市场基准指数的市场表现预期。其中，A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准；美国市场以标普 500 指数为基准。具体标准如下：

股票评级

强烈推荐：预期公司股价涨幅超越基准指数 20%以上

增持：预期公司股价涨幅超越基准指数 5-20%之间

中性：预期公司股价变动幅度相对基准指数介于±5%之间

减持：预期公司股价表现弱于基准指数 5%以上

行业评级

推荐：行业基本面向好，预期行业指数超越基准指数

中性：行业基本面稳定，预期行业指数跟随基准指数

回避：行业基本面转弱，预期行业指数弱于基准指数

重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。**仅对中国境内投资者发布，请自行评估接收相关内容的适当性。本公司不会因收到、阅读相关内容而视为中国境内投资者。**本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价，在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。