信达证券 CINDA SECURITIES

Research and Development Center

CES 2026：英伟达六大芯片协同升级，算力+存力迈入新纪元

2026 年 1 月 11 日

证券研究报告

行业研究

行业专题研究（普通）

电子

投资评级 看好

上次评级 看好

莫文宇 电子行业首席分析师
执业编号：S1500522090001
邮 箱：mowenyu@cindasc.com

王义夫 电子行业分析师
执业编号：S1500525090001
邮 箱：wangyifu@cindasc.com

信达证券股份有限公司
CINDA SECURITIES CO.,LTD
北京市西城区宣武门西大街甲 127 号金隅大厦 B 座
邮编：100031

CES 2026：英伟达六大芯片协同升级，算力+存力迈入新纪元

2026 年 01 月 11 日

本期内容提要：

- **Vera Rubin：六大芯片协同升级，推理性能大幅提升。**美国拉斯维加斯时间 1 月 5 日，英伟达创始人兼 CEO 黄仁勋在国际消费电子展 CES 2026 上发表主题演讲，并正式发布 Rubin 平台。该平台由六款专为打造超凡 AI 超级计算机而设计的全新芯片组成，包括 Vera CPU、Rubin GPU、NVLink 6 交换机、ConnectX-9 SuperNIC、BlueField-4 DPU 和 NVIDIA Spectrum-6 以太网交换机，6 款芯片协同设计，从而大幅缩短训练时间并降低推理 token 成本。**（1）算力方面**，Rubin GPU 引入 Transformer 引擎，NVFP4 推理性能高达 50PFLOPS，是 Blackwell GPU 的 5 倍，向后兼容，在保持推理精度的同时提升 BF16/FP4 级别的性能；NVFP4 训练性能达到 35PFLOPS，是 Blackwell 的 3.5 倍。**（2）存储方面**，针对大量 KV Cache 问题，Rubin 平台引入了由 BlueField-4 驱动的 NVIDIA 推理上下文内存存储，构建了一个 AI 原生的基础设施层，用于存储 KV Cache，通过扩展 GPU 内存容量为上下文记忆提供基础设施。**（3）PCB 方面**，从 Cable Tray 到 Cableless 互联，Rubin NVL72 PCB 价值量显著提升，这种机械简化使得组装速度比上一代托盘架构快 18 倍，并且显著缩短了现场维护期间的服务时间，降低了部署时间和持续的运营成本。**（4）Rack 方面**，系统协同设计显著提升，ODM 环节核心受益。由于在 6 款芯片、托盘、机架、Pod、数据中心与软件层面实现了极致协同设计，Rubin 平台在训练与推理成本上实现了大幅下降。
- **开源模型：从数据到代码全覆盖，Nemotron 3 重构 Agentic AI 范式。**本次 CES 上，黄仁勋宣布扩展其开源模型生态（Open Model Universe），新增和更新了一系列模型、数据集、代码库和工具。这个生态覆盖六大领域：生物医学 AI（Clara）、AI 物理模拟（Earth-2）、Agentic AI（Nemotron）、物理 AI（Cosmos）、机器人（GR00T）和自动驾驶（Alpamayo）。Nemotron 系列是此次更新的重点，覆盖了推理、RAG、安全和语音四个应用方向。
- **物理 AI：从 Cosmos 到 Alpamayo，走向商业化落地。**在物理 AI 领域，英伟达更新了用于理解和生成物理世界视频的 Cosmos，机器人通用基础模型 Isaac GR00T、自动驾驶视觉-语言-行动模型 Alpamayo。Cosmos 是一个开源的物理 AI 世界基础模型，目前已经用海量视频、真实驾驶与机器人数据，以及 3D 模拟做过预训练。

Alpamayo 是一个面向自动驾驶领域的开源工具链，也是首个开源的视觉-语言-行动（VLA）推理模型。

- **投资建议：英伟达 Rubin 平台正式发布开启 AI 算力新纪元，全球算力设施向“AI 工厂”范式全面转型。**我们认为，Rubin 通过协同设计实现性能飞跃，算力、存储、PCB、机架等多环节价值量或显著提升，产业链核心环节在基建投资化趋势下仍具备较大的成长空间，建议关注：【海外 AI】工业富联/沪电股份/鹏鼎控股/胜宏科技/生益科技/生益电子等；【国产 AI】寒武纪/芯原股份/海光信息/中芯国际/深南电路等；【存储】德明利/江波龙/兆易创新/聚辰股份/普冉股份等；【SoC】瑞芯微/乐鑫科技/恒玄科技/晶晨股份/中科蓝讯等。
- **风险提示：**AI 发展进程不及预期，下游需求发展不及预期。

目 录

Vera Rubin：六大芯片协同升级，推理性能大幅提升 5
算力：NVFP4 推理性能提升 5 倍，推理成本显著下降 6
存储：针对大量 KV Cache 问题，推出推理上下文内存存储平台 6
PCB：从 Cable Tray 到 Cableless 互联，PCB 价值量显著提升 8
Rack：系统协同设计显著提升，ODM 环节核心受益 9
开源模型：从数据到代码全覆盖，Nemotron 3 重构 Agentic AI 范式 10
物理 AI：从 Cosmos 到 Alpamayo，走向商业化落地 11
投资建议 12
风险因素 12

图 目 录

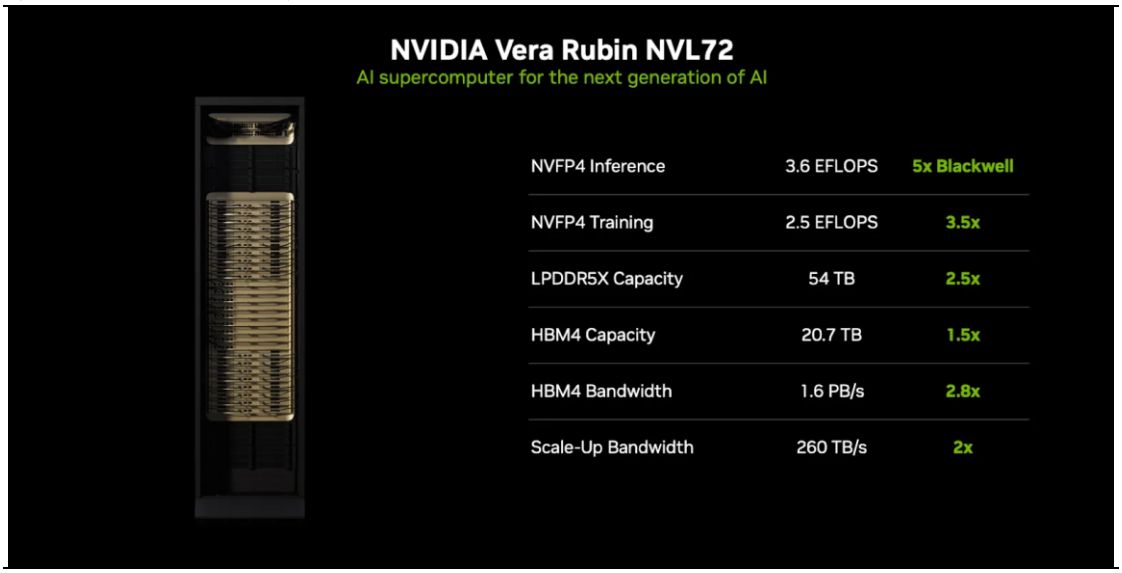
图 1：Vera Rubin NVL72 概述 5
图 2：Vera Rubin 平台六款新芯片 5
图 3：Vera CPU 主要参数 6
图 4：Rubin GPU 相较于 Blackwell 算力提升 5 倍 6
图 5：Rubin HBM4 总内存带宽提升至 22TB/s 7
图 6：NVLink-C2C 一致性内存架构 7
图 7：BlueField-4 DPU 7
图 8：Rubin 平台的存储结构 8
图 9：Vera Rubin 计算托盘采用 cableless 无缆互联架构 8
图 10：基于 Vera Rubin 构建的 NVIDIA DGX SuperPOD 9
图 11：英伟达开源模型生态系统 10
图 12：英伟达全栈物理 AI 平台 11
图 13：英伟达推出开源世界基础模型 Cosmos 11
图 14：面向自动驾驶领域的开源工具链 Alpamayo 11

Vera Rubin：六大芯片协同升级，推理性能大幅提升

Rubin 平台正式发布，六款全新芯片协同增强推理性能。美国拉斯维加斯时间 1 月 5 日，英伟达创始人兼 CEO 黄仁勋在国际消费电子展 CES 2026 上发表主题演讲，并正式发布 Rubin 平台。该平台由六款专为打造超凡 AI 超级计算机而设计的全新芯片组成，包括 Vera CPU、Rubin GPU、NVLink 6 交换机、ConnectX-9 SuperNIC、BlueField-4 DPU 和 NVIDIA Spectrum-6 以太网交换机，6 款芯片协同设计，从而大幅缩短训练时间并降低推理 token 成本。

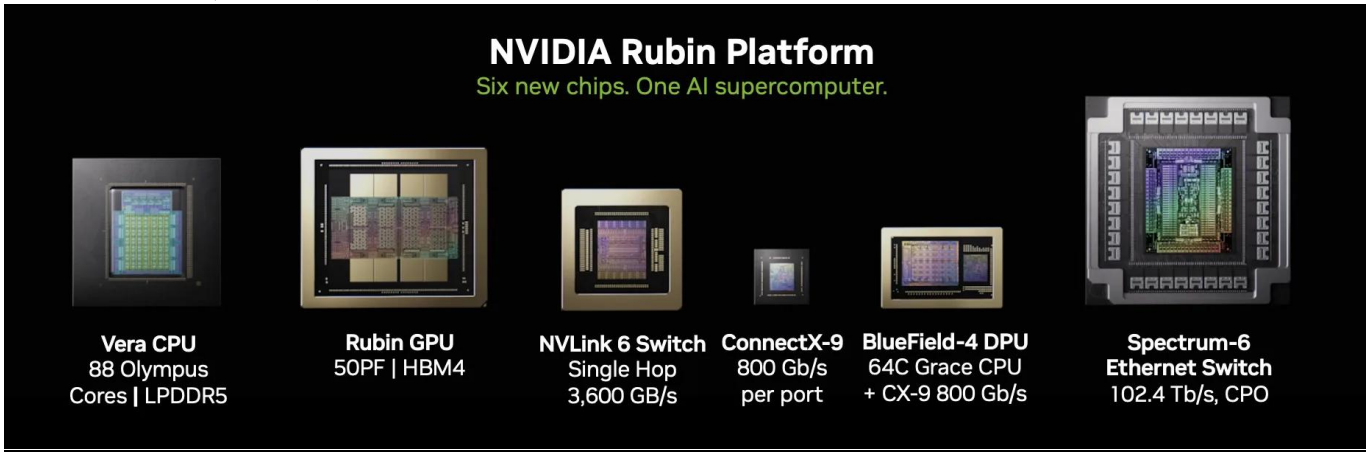
Rubin 平台的旗舰产品是 Vera Rubin NVL72 机架级系统，其设计使得整个机架在更大的 AI 工厂中作为一个协调运作的机器发挥作用。NVL72 系统不仅针对峰值性能进行了优化，还致力于实现持续的智能生成：可预测的延迟、在异构执行阶段的高利用率，以及将电力高效转化为可用智能。

图 1：Vera Rubin NVL72 概述



资料来源：Nvidia 官网，信达证券研发中心

图 2：Vera Rubin 平台六款新芯片

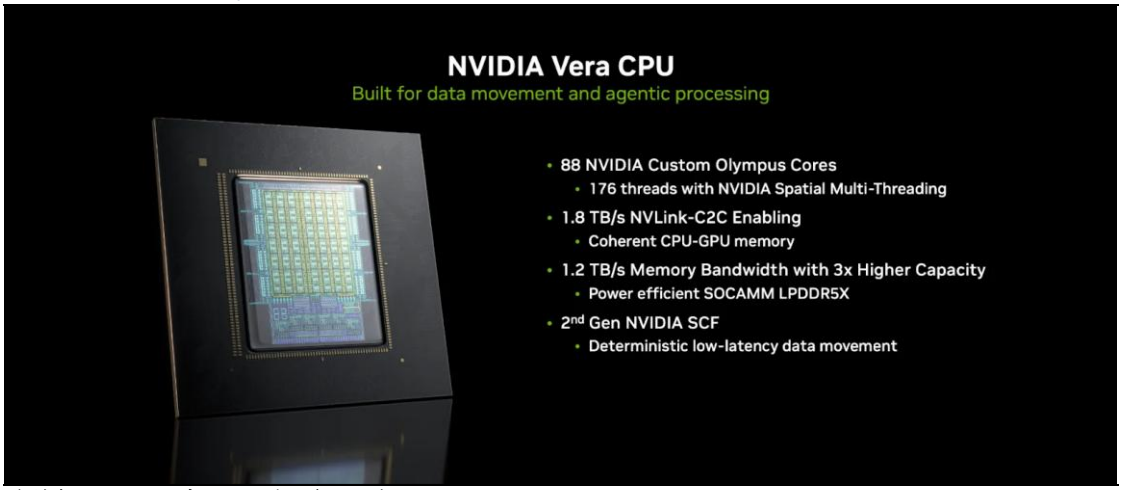


资料来源：Nvidia 官网，信达证券研发中心

算力：NVFP4 推理性能提升 5 倍，推理成本显著下降

Vera CPU: 为数据移动和智能体处理而设计，拥有 88 个英伟达定制 Olympus 核心、176 线程的英伟达空间多线程，1.8TB/sNVLink-C2C 支持 CPU:GPU 统一内存，系统内存达 1.5TB（是 Grace CPU 的 3 倍），SOCAMM LPDDR5X 内存带宽为 1.2TB/s，并支持机架级机密计算，数据处理性能翻倍提升。

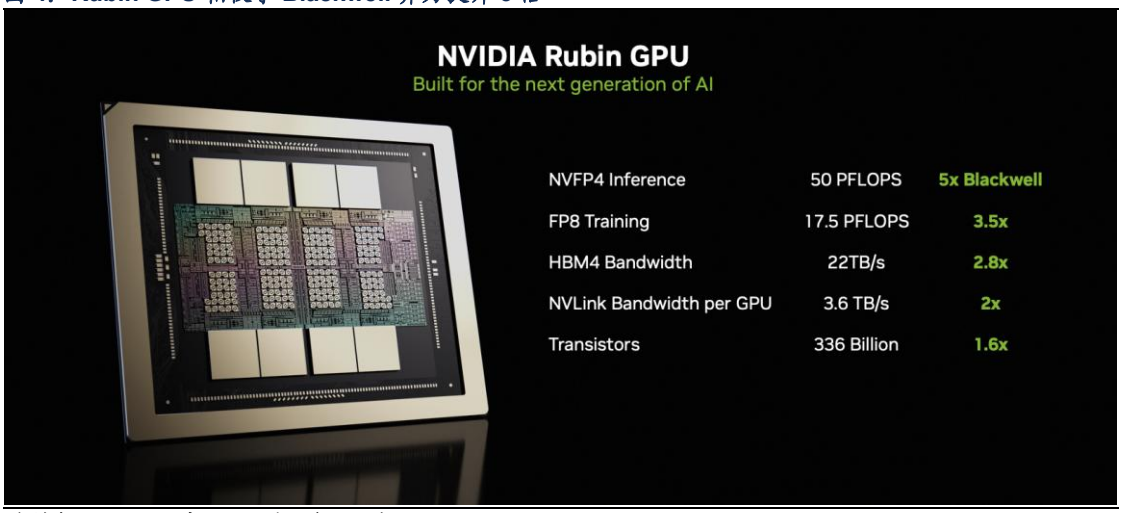
图 3：Vera CPU 主要参数



资料来源：Nvidia 官网，信达证券研发中心

Rubin GPU: 引入 Transformer 引擎，NVFP4 推理性能高达 50PFLOPS，是 Blackwell GPU 的 5 倍，向后兼容，在保持推理精度的同时提升 BF16/FP4 级别的性能；NVFP4 训练性能达到 35PFLOPS，是 Blackwell 的 3.5 倍。

图 4：Rubin GPU 相较于 Blackwell 算力提升 5 倍



资料来源：Nvidia 官网，信达证券研发中心

存储：针对大量 KV Cache 问题，推出推理上下文内存存储平台

HBM: 随着上下文长度的增加以及推理过程变得越来越具交互性，内存性能的表现已成为影响整体效率的主导因素。Rubin GPU 集成了新一代的 HBM4，每个 GPU 最高配备 288GB HBM4，总内存带宽提升至 22TB/s，是 Blackwell 的 2.8 倍。这使得 Rubin GPU 能

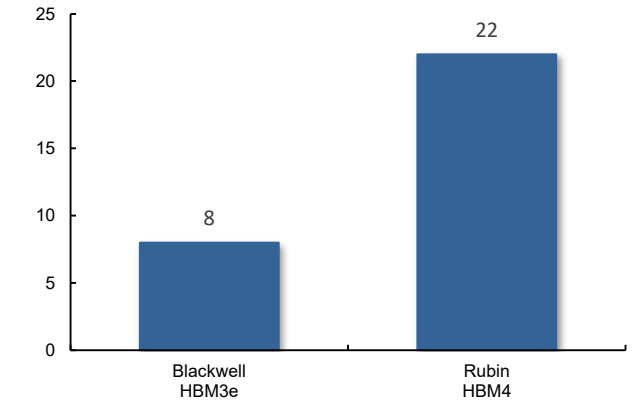
够在不牺牲并发性或者利用率的情况下，支持长上下文推理、高批次 MoE 执行和交互式推理。

DRAM: 每个 Vera CPU 还将搭载最高 1.5TB 的 LPDDR5X 内存子系统，能在低功耗下提供高达 1.2TB/s 的带宽，采用 LPDDR5X SOCAMM 形式。第二代 NVLink-C2C 在 Vera CPU 和 Rubin GPU 之间提供 1.8 TB/s 的一致性带宽，实现了 CPU 和 GPU 内存间的统一地址空间。应用可将 LPDDR5X 和 HBM4 视为一个单一的一致性内存池，减少数据移动开销，并支持 KV Cache 和高效多模型执行等技术。

NAND: 随着 AI 模型扩展到数万亿参数和多步骤推理，它们会生成大量的上下文数据，并以 KV Cache 表示，这对准确性、用户体验和连续性至关重要。KV Cache 无法长期存储在 GPU 上，因为这会成为多智能体系统中实时推理的瓶颈，AI 原生应用需要一种新型的可扩展基础设施来存储和共享这些数据。

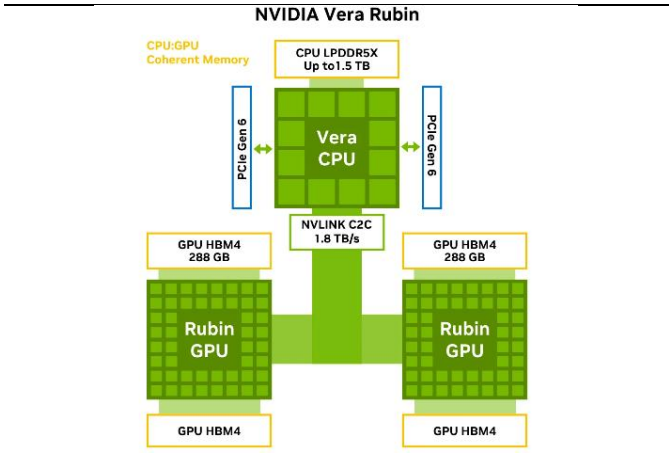
Rubin 平台引入了由 BlueField-4 驱动的 NVIDIA 推理上下文内存存储，构建了一个 AI 原生的基础设施层，用于存储 KV Cache，通过扩展 GPU 内存容量为上下文记忆提供基础设施，支持跨节点高速共享，与传统存储相比，可将每秒 tokens 数提升高达 5 倍，并实现高达 5 倍的能效提升。每个节点配 4 个 BlueField-4，每个背后有 150TB 上下文 NAND，分配到 GPU 上，每块 GPU 额外获得 16TB NAND，同时带宽保持 200Gbps。我们认为这种新型存储架构或将显著提升 NAND 需求。

图 5：Rubin HBM4 总内存带宽提升至 22TB/s



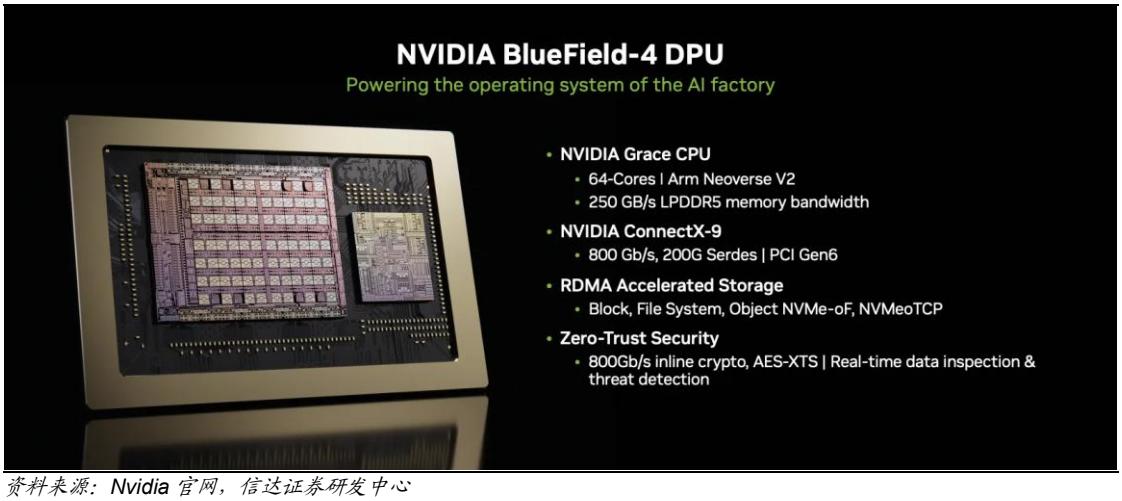
资料来源：Nvidia 官网，信达证券研发中心

图 6：NVLink-C2C 一致性内存架构



资料来源：Nvidia 官网，信达证券研发中心

图 7：BlueField-4 DPU



资料来源：Nvidia 官网，信达证券研发中心

图 8：Rubin 平台的存储结构

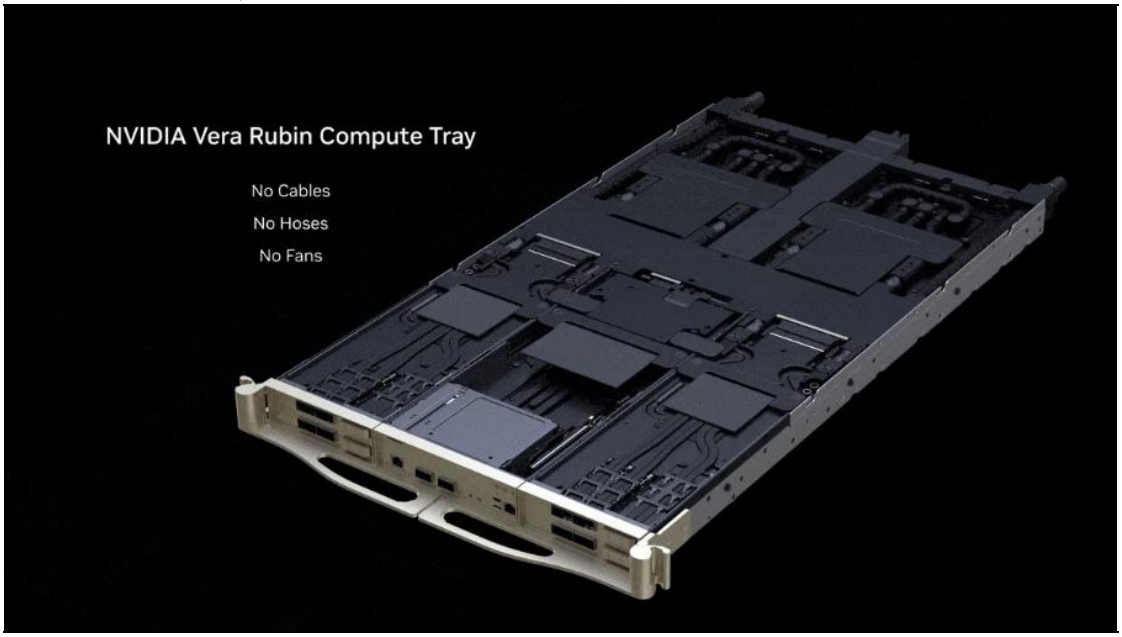


资料来源：智东西，信达证券研发中心

PCB：从 Cable Tray 到 Cableless 互联，PCB 价值量显著提升

NVIDIA 上一代 Blackwell NVL72 采用传统 cable tray 架构，线缆连接复杂不仅增加了部署难度，更成为故障高发点和服务效率的瓶颈。而 Rubin NVL72 实现了核心架构的重大革新——全面转向 cableless 无缆互联架构，通过 PCB 板直接实现组件间的信号传输，彻底告别了繁杂的线缆布局。这种机械简化使得组装速度比上一代托盘架构快 18 倍，并且显著缩短了现场维护期间的服务时间，降低了部署时间和持续的运营成本。

图 9：Vera Rubin 计算托盘采用 cableless 无缆互联架构

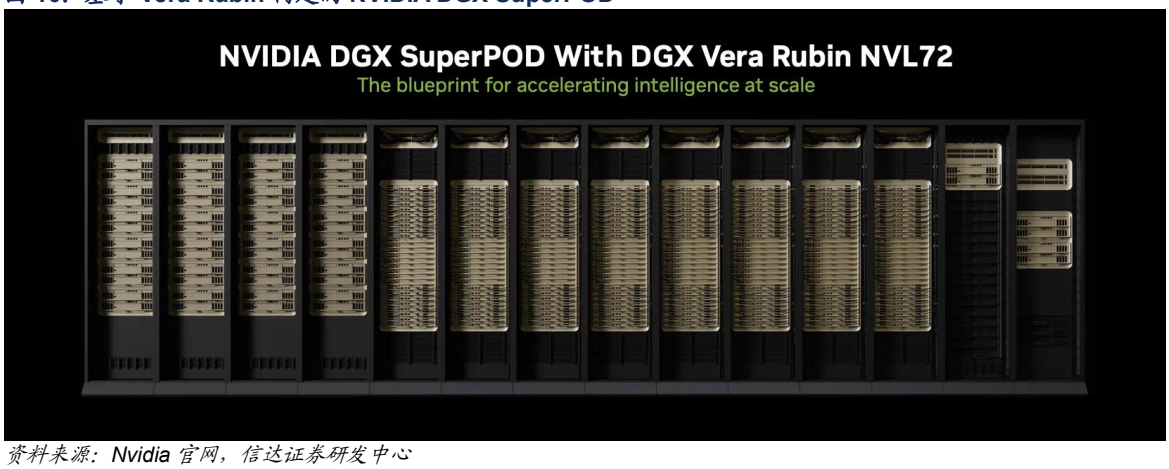


资料来源：Nvidia 官网，信达证券研发中心

Rack：系统协同设计显著提升，ODM 环节核心受益

在系统层面，NVIDIA DGX SuperPOD 作为大规模 AI 工厂部署蓝图，采用 8 套 DGX Vera Rubin NVL72 系统，用 NVLink 6 纵向扩展网络，用 Spectrum-X Ethernet 横向扩展网络，内置 NVIDIA 推理上下文内存存储平台，并经过工程化验证。整个系统由 NVIDIA Mission Control 软件管理，实现极致效率。客户可将其作为交钥匙平台部署，用更少 GPU 完成训练与推理任务。由于在 6 款芯片、托盘、机架、Pod、数据中心与软件层面实现了极致协同设计，Rubin 平台在训练与推理成本上实现了大幅下降。与上一代 Blackwell 相比，训练相同规模的 MoE 模型，仅需 1/4 的 GPU 数量；在相同延迟下，大型 MoE 模型的 token 成本降低至 1/10。

图 10：基于 Vera Rubin 构建的 NVIDIA DGX SuperPOD

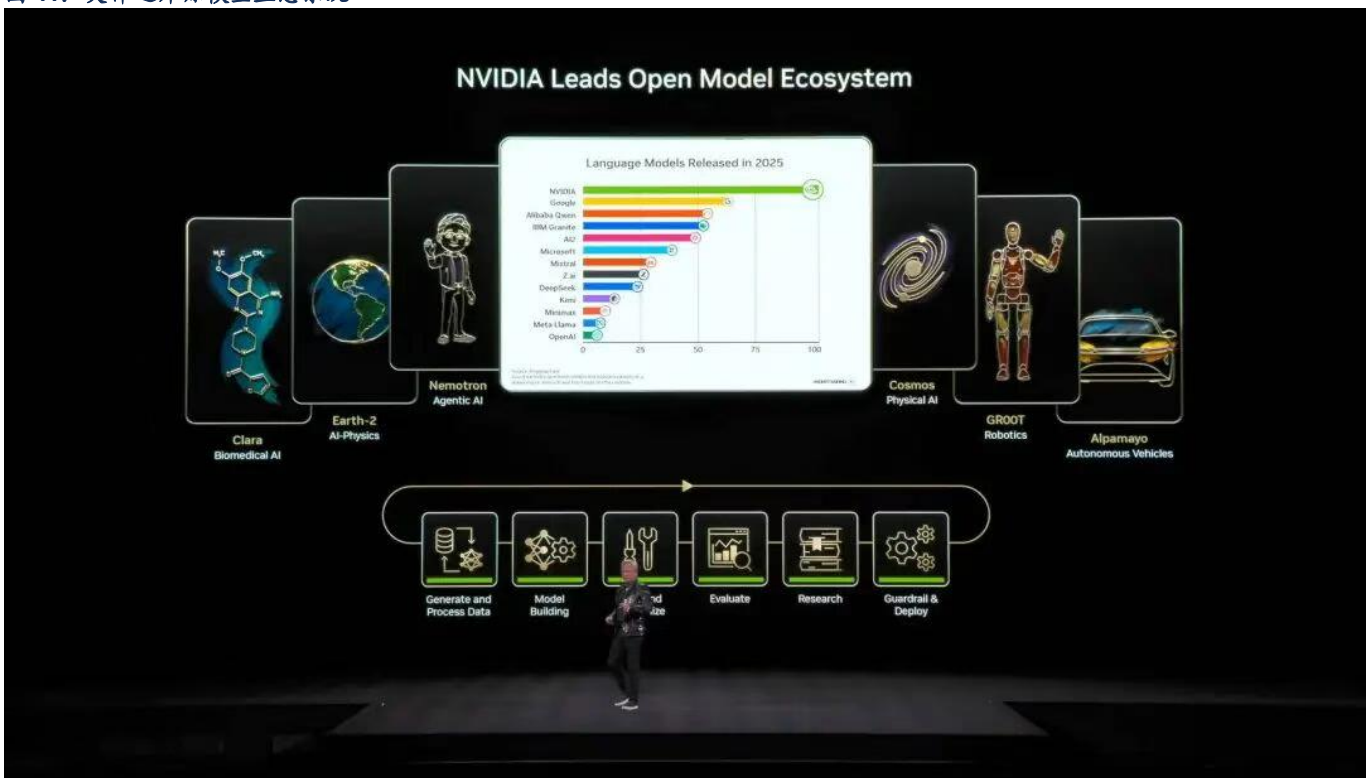


资料来源：Nvidia 官网，信达证券研发中心

开源模型：从数据到代码全覆盖，Nemotron 3 重构 Agentic AI 范式

本次 CES 上，黄仁勋宣布扩展其开源模型生态（Open Model Universe），新增和更新了一系列模型、数据集、代码库和工具。这个生态覆盖六大领域：生物医学 AI（Clara）、AI 物理模拟（Earth-2）、Agentic AI（Nemotron）、物理 AI（Cosmos）、机器人（GR00T）和自动驾驶（Alpamayo）。

图 11：英伟达开源模型生态系统



资料来源：爱范儿，信达证券研发中

Nemotron 系列是此次更新的重点，覆盖了四个应用方向。1）在推理方向，包括 Nemotron 3 Nano、Nemotron 2 Nano VL 等小型化推理模型，以及 NeMo RL、NeMo Gym 等强化学习训练工具。2）在 RAG（检索增强生成）方向，提供了 Nemotron Embed VL（向量嵌入模型）、Nemotron Rerank VL（重排序模型）、相关数据集和 NeMo Retriever Library（检索库）。3）在安全方向，有 Nemotron Content Safety 内容安全模型及配套数据集、NeMo Guardrails 护栏库。4）在语音方向，则包含 Nemotron ASR 自动语音识别、Granary Dataset 语音数据集和 NeMo Library 语音处理库。

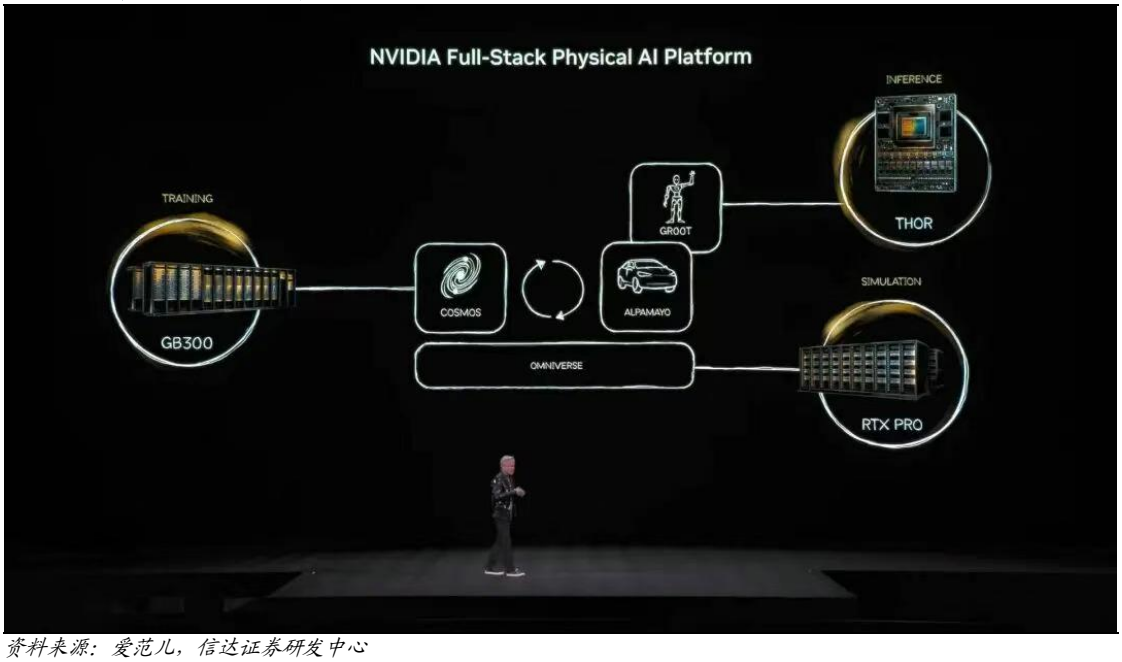
物理 AI：从 Cosmos 到 Alpamayo，走向商业化落地

在物理 AI 领域，英伟达更新了用于理解和生成物理世界视频的 Cosmos，机器人通用基础模型 Isaac GR00T、自动驾驶视觉-语言-行动模型 Alpamayo。

Cosmos 是一个开源的物理 AI 世界基础模型，目前已经用海量视频、真实驾驶与机器人数据，以及 3D 模拟做过预训练。它能理解世界是怎么运行的，可以把语言、图像、3D 和动作联系起来。黄仁勋表示，Cosmos 能实现不少物理 AI 技能，比如生成内容、做推理、预测轨迹。它可以依据 3D 场景生成逼真的视频，根据驾驶数据生成符合物理规律的运动，还能从模拟器、多摄像头画面或文字描述生成全景视频。

Alpamayo 是一个面向自动驾驶领域的开源工具链，也是首个开源的视觉-语言-行动（VLA）推理模型。Alpamayo 最大的突破在于它是“推理型”自动驾驶模型。传统自动驾驶系统是“感知-规划-控制”的流水线架构，看到红灯就刹车，看到行人就减速，遵循预设规则。而 Alpamayo 引入了“推理”能力，理解复杂场景中的因果关系，预测其他车辆和行人的意图，甚至能处理需要多步思考的决策。

图 12：英伟达全栈物理 AI 平台



资料来源：爱范儿，信达证券研发中心

图 13：英伟达推出开源世界基础模型 Cosmos



资料来源：腾讯科技，信达证券研发中心

图 14：面向自动驾驶领域的开源工具链 Alpamayo



资料来源：腾讯科技，信达证券研发中心

投资建议

英伟达 Rubin 平台正式发布开启 AI 算力新纪元，全球算力设施向“AI 工厂”范式全面转型。我们认为，Rubin 通过协同设计实现性能飞跃，算力、存储、PCB、机架等多环节价值量或显著提升，产业链核心环节在基建投资化趋势下仍具备较大的成长空间，建议关注：

- 海外 AI：工业富联/沪电股份/鹏鼎控股/胜宏科技/生益科技/生益电子等；
- 国产 AI：寒武纪/芯原股份/海光信息/中芯国际/深南电路等；
- 存储德：明利/江波龙/兆易创新/聚辰股份/普冉股份等；
- SoC：瑞芯微/乐鑫科技/恒玄科技/晶晨股份/中科蓝讯等。

风险因素

（1）AI 发展进程不及预期：AI 前期投资规模庞大，且大模型开发存在一定难度，AI 发展进展可能不及预期；

（2）下游需求发展不及预期：目前 AI 应用普及程度仍然有限，若未来需求增长幅度不及预期，可能影响上游开发进程。

研究团队简介

莫文宇，电子行业分析师，S1500522090001。毕业于美国佛罗里达大学，电子工程硕士，2012-2022 年就职于长江证券研究所，2022 年入职信达证券研发中心，任电子行业首席分析师。

郭一江，电子行业研究员。本科兰州大学，研究生就读于北京大学化学专业。2020 年 8 月入职华创证券电子组，后于 2022 年 11 月加入信达证券电子组，研究方向为光学、消费电子、汽车电子等。

杨宇轩，电子行业分析师，华北电力大学本科，清华大学硕士，曾就职于东方证券、首创证券、赛迪智库，2025 年 1 月加入信达证券电子组，研究方向为半导体等。

王义夫，电子行业研究员。西南财经大学金融学士，复旦大学金融硕士，2023 年加入信达证券电子组，研究方向为存储芯片、模拟芯片等。

分析师声明

负责本报告全部或部分内容的每一位分析师在此申明，本人具有证券投资咨询执业资格，并在中国证券业协会注册登记为证券分析师,以勤勉的职业态度,独立、客观地出具本报告；本报告所表述的所有观点准确反映了分析师本人的研究观点；本人薪酬的任何组成部分不曾与，不与，也将不会与本报告中的具体分析意见或观点直接或间接相关。

免责声明

信达证券股份有限公司(以下简称“信达证券”)具有中国证监会批复的证券投资咨询业务资格。本报告由信达证券制作并发布。

本报告是针对与信达证券签署服务协议的签约客户的专属研究产品，为该类客户进行投资决策时提供辅助和参考，双方对权利与义务均有严格约定。本报告仅提供给上述特定客户，并不面向公众发布。信达证券不会因接收人收到本报告而视其为本公司的当然客户。客户应当认识到有关本报告的电话、短信、邮件提示仅为研究观点的简要沟通，对本报告的参考使用须以本报告的完整版本为准。

本报告是基于信达证券认为可靠的已公开信息编制，但信达证券不保证所载信息的准确性和完整性。本报告所载的意见、评估及预测仅为本报告最初出具日的观点和判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会出现不同程度的波动，涉及证券或投资标的的历史表现不应作为日后表现的保证。在不同时期，或因使用不同假设和标准，采用不同观点和分析方法，致使信达证券发出与本报告所载意见、评估及预测不一致的研究报告，对此信达证券可不发出特别通知。

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，也没有考虑到客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况，若有必要应寻求专家意见。本报告所载的资料、工具、意见及推测仅供参考，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人做出邀请。

在法律允许的情况下，信达证券或其关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能会为这些公司正在提供或争取提供投资银行业务服务。

本报告版权仅为信达证券所有。未经信达证券书面同意，任何机构和个人不得以任何形式翻版、复制、发布、转发或引用本报告的任何部分。若信达证券以外的机构向其客户发放本报告，则由该机构独自为此发送行为负责，信达证券对此等行为不承担任何责任。本报告同时不构成信达证券向发送本报告的机构之客户提供的投资建议。

如未经信达证券授权，私自转载或者转发本报告，所引起的一切后果及法律责任由私自转载或转发者承担。信达证券将保留随时追究其法律责任的权利。

评级说明

投资建议的比较标准	股票投资评级	行业投资评级
本报告采用的基准指数 ：沪深 300 指数（以下简称基准）； 时间段：报告发布之日起 6 个月内。	**买入：**股价相对强于基准 15%以上；	**看好：**行业指数超越基准；
	增持：股价相对强于基准 5%～15%；	**中性：**行业指数与基准基本持平；
	持有：股价相对基准波动在±5% 之间；	**看淡：**行业指数弱于基准。
	卖出：股价相对弱于基准 5%以下。	

风险提示

证券市场是一个风险无时不在的市场。投资者在进行证券交易时存在赢利的可能，也存在亏损的风险。建议投资者应当充分深入地了解证券市场蕴含的各项风险并谨慎行事。

本报告中所述证券不一定能在所有的国家和地区向所有类型的投资者销售，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专业顾问的意见。在任何情况下，信达证券不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任，投资者需自行承担风险。