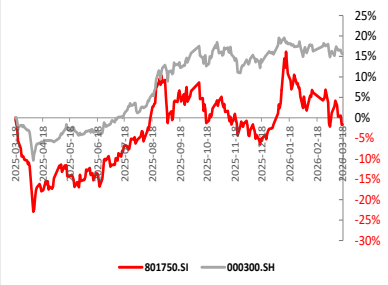
市场研究部 2026 年 3 月 18 日

英伟达 GTC：芯片提升算力 全栈推动产业

看好

市场表现截至 2026.3.17



数据来源：Wind，国新证券整理

事件

2026 年 3 月 16 日，英伟达创始人兼 CEO 黄仁勋在 GTC 2026 大会上发表主题演讲，核心议题涵盖 CUDA 平台 20 周年、推理拐点与算力需求爆发、Vera Rubin 系统架构、Groq 集成、OpenClaw 代理革命及物理 AI 与机器人。

核心观点

在 GTC 2026 大会上，英伟达将 2025-2027 年累计收入指引提升至至少 1 万亿美元，其中 60%业务将来自超大规模云，数据中心转型为生产 Token 的“AI 工厂”，Token 成为核心数字商品，并给出其分层定价体系。

大会主要是推出 Vera Rubin 全栈 AI 算力平台。它包含七大芯片：Rubin GPU、Vera CPU、NVLink 6 交换机、ConnectX-9 超级网卡、BlueField-4 DPU、Spectrum-6 以太网交换机和 Groq 3 LPU，覆盖计算、推理、网络与存储全链路。五类标准化机架，采用统一 MGX 模块化架构与 100%液冷设计，将整机安装时间从两天缩短至两小时，极大提升了部署效率。

Vera CPU 采用 88 核 Arm 架构，单线程性能提升 50%，成为智能体任务调度的核心。通过与 Groq 3 LPU 基于 Dynamo 框架的异构协同，系统实现了推理流程解耦：GPU 处理高吞吐任务，LPU 专攻低延迟解码。这使得每兆瓦推理吞吐量最高提升 35 倍，万亿参数模型收益潜力提升 10 倍，单 Token 成本降至 Blackwell 平台的十分之一。

软件层面，CUDA 生态装机量达数亿，形成坚固壁垒。针对开源的 OpenClaw 智能体框架，英伟达推出企业级安全方案 NemoClaw，支持混合调用本地与云端模型，推动智能体进入企业核心场景，并断言 SaaS 将全面转向 AaaS。

物理 AI 进入爆发期，现场展示 110 台机器人，比亚迪、现代等主流车企加入其自动驾驶生态，计划 2028 年在全球 28 城部署 L4 车队。同时，英伟达发布用于太空的 Vera Rubin Space-1 轨道计算模块，将算力拓展至近地轨道。

技术路线上，英伟达明确光铜协同，其量产 CPO 交换机将成为扩展核心。Vera Rubin 平台的全系统设计也带动了高端 PCB、光模块、导热材料等产业链的升级需求。

投资线索

聚焦智能体算力基础设施、推理加速与异构计算、AI 基础软件与向量数据库、大模型与行业智能体、数据中心运营。优先布局能紧跟技术迭代、在推理加速、高速互联、液冷、智能体调度与安全等关键环节有核心壁垒的企业，把握国产替代红利。

分析师：钟哲元
登记编码：S1490523030001
邮箱：zhongzheyuan@crsec.com.cn

证券研究报告

风险提示

1、技术发展不及预期；2、市场竞争加剧；3、地缘政治影响。

目录

一、推理拐点确立，1 万亿美元市场开启

英伟达在 GTC 2026 大会上，将 2025-2027 年累计收入指引提升至至少 1 万亿美元， AI 发展从实验室主导的训练阶段，全面迈入规模化推理、生成与智能体执行的新阶段。黄仁勋在演讲中明确，这一万亿市场中 60%业务将来自超大规模云计算，背后是过去两年 AI 计算需求约 10000 倍的指数级增长，传统以存储文件为核心的数据中心，彻底转型为生产 Token 的 “AI 工厂”，Token 成为 AI 时代核心数字商品。在 Token 工厂的商业模式中，以每瓦生成 Token 数为纵轴、交互速度为横轴的效率模型成为重要评判标准，纵轴代表固定电力下的运营效率，决定 AI 工厂的成本控制能力与盈利基底，横轴则直接对应终端用户体验与服务分层定价，二者共同构成 Token 分层定价体系的底层逻辑，从免费层到 Premium 层的价值分化，对应这一效率模型实现，也成为企业 AI 基础设施投入的决策依据之一。

二、Vera Rubin 全栈落地，七大芯片+五类机架定型新架构

本次大会的核心成果是 Vera Rubin 全栈 AI 算力平台，英伟达摒弃单一芯片迭代思路，以全链路系统级协同设计重构 AI 算力底座，使其成为智能体 AI 时代的标准。该平台包含七大全新量产芯片，分别为 Rubin GPU、Vera CPU、NVLink 6 交换机、ConnectX-9 超级网卡、BlueField-4 DPU、Spectrum-6 以太网交换机、Groq 3 LPU，全面覆盖计算、专用推理、高速网络、AI 原生存储等环节，实现从 AI 预训练、后训练、推理全流程优化。同步推出五类标准化机架，包括 Vera Rubin NVL72 GPU 机架、Vera CPU 机架、Groq 3 LPX 推理机架、BlueField-4 STX 存储机架、Spectrum-6 SPX 以太网机架，采用统一 MGX 模块化架构可自由组合部署，搭配 100%液冷设计与 45℃温水冷却方案，彻底简化内部布线，将整机安装时间从传统两天大幅缩短至两小时，显著降低 AI 集群的部署难度、冷却压力与总体运营成本，推动超大规模 AI 工厂的规模化落地效率实现飞跃。

三、CPU+GPU+LPU 异构协同，推理性能量级突破

Vera CPU 的正式发布，标志着英伟达强势进军数据中心 CPU 市场，该芯片采用 88 核 Arm v9.2-A Olympus 核心架构，单线程性能提升 50%，通过 NVLink-C2C 互连技术与 GPU 实现 1.8TB/s 相干带宽高速互联，成为 AI 智能体任务调度、工具调用、代码编译等场景的核心中枢。英伟达整合 2025 年 12 月收购的 Groq 核心技术，推出 Groq 3 LPU 专用推理芯片，构建起 GPU+LPU 深度协同的异构计算架构，二者基于 Dynamo 推理调度框架实现推理流程解聚分工：Rubin GPU 凭借 HBM4 海量带宽与强大算力，专注处理高吞吐的预填充环节与注意力机制解码任务，承载大模型上下文输入与 KV Cache 管理；Groq 3 LPU 依托 500MB 片上 SRAM 带来的 150TB/s 超高带宽，专攻低延迟敏感的前馈网络解码环节，解决纯 GPU 在

推理解码阶段利用率低、延迟高的行业痛点。这种专业化分工协同，使系统每兆瓦推理吞吐量最高提升 35 倍，万亿参数模型的商业收益潜力提升 10 倍，完美破解大模型推理长期存在的低延迟与高吞吐量不可兼得的矛盾，同时将单 Token 生成成本降至 Blackwell 平台的十分之一，大幅改变了 AI 推理的成本与效率。

四、CUDA 壁垒加固，NemoClaw 推动企业智能体规模化

今年恰逢 CUDA 诞生 20 周年，全球数亿级的 CUDA 装机量形成难以复制的生态飞轮，持续吸引开发者、催生算法创新与新市场，成为英伟达算力生态的核心壁垒。针对当下爆火的 OpenClaw 智能体框架，英伟达推出 NemoClaw 企业级安全方案，在开源 OpenClaw 基础上叠加隔离沙箱、隐私路由、企业级安全防护多层机制，通过 NVIDIA Agent Toolkit 实现一键部署优化，支持本地开源模型与云端前沿模型混合调用，解决智能体在企业场景中访问敏感数据、执行外部通信的安全隐患，推动 AI 智能体从消费级场景快速渗透至企业核心业务场景。黄仁勋在演讲中断言，未来传统 SaaS 将全面转向 AaaS（智能体即服务），企业需制定专属 OpenClaw 战略；同时英伟达发布 Nemotron 系列开源模型并组建 Nemotron 联盟，联合全球顶尖 AI 技术公司完善语言、物理、自动驾驶等多场景开放模型生态，进一步夯实全栈软件与模型布局。

五、物理 AI 升级，太空计算拓展算力版图

物理 AI 进入规模化落地爆发期，本次大会现场展示 110 台机器人，比亚迪、现代、日产、吉利等主流车企正式加入 DRIVE Hyperion 生态，计划 2028 年在全球四大洲 28 座城市部署 L4 级自动驾驶车队，与 Uber 的合作也将在 2027 年率先落地洛杉矶与旧金山湾区。英伟达同步推出全新 Isaac 仿真框架、Cosmos 世界基础模型与 Isaac GR00T 开放模型，打造物理 AI 数据工厂 Blueprint，实现机器人、自动驾驶车辆物理 AI 模型的大规模数据处理、合成数据生成与强化学习，全面赋能工业机器人与人形机器人的研发、训练与部署。与此同时，英伟达发布 Vera Rubin Space-1 轨道计算模块，针对太空无对流、无传导仅辐射散热的工程难题，将数据中心级 AI 算力部署至近地轨道，支撑轨道数据中心、高级地理空间智能处理与自主太空操作任务，实现从地面数据中心到太空的全域算力覆盖，拓展 AI 算力的全新应用疆域。

六、光铜协同演进，高端元器件需求提升

英伟达明确光铜协同的长期技术路线，打破单一技术路线争论：Spectrum-6 成为全球首款量产 CPO（共封装光学）交换机，由英伟达与台积电合作开发，

2027 年起 CPO 正式成为 Scale-Up 核心方案，2028 年 Feynman 架构将采用 CPO 与铜缆混合互联模式，铜缆在短距高速互联中仍具备不可替代的价值，光学扩展则负责长距大规模集群连接。Vera Rubin 平台的全液冷设计、Kyber 机架垂直插入与中板连接架构、正交背板技术，叠加 LPU 专用机架的高密度集成需求，直接带动高阶 PCB、高速互连芯片、800G/1.6T 光模块、先进接口组件、高端导热材料等算力硬件产业链需求快速提升，算力硬件从单一芯片竞争转向系统级配套升级，相关高端元器件迎来量价齐升的结构性机遇。

七、投资线索

本次 GTC 2026 进一步印证，Agent 爆发将驱动推理算力进入高速增长期，英伟达围绕 Vera Rubin 全栈平台、Groq 3 LPU 异构推理、NemoClaw 企业级智能体生态完成系统性升级，同时强化 Dynamo、cuVS 等软件栈对智能体工作流的适配，标志着 AI 算力正式从训练主导转向智能体+推理双轮驱动。结合国内产业环境，一方面英伟达高端算力产品在国内落地仍存在政策不确定性，另一方面智能体普及带来的推理算力需求门槛更适配国产芯片突破方向，使得 AI 算力硬件、基础软件、模型与数据中心服务迎来明确的国产替代与产业升级机遇。

建议围绕智能体算力基础设施、推理加速与异构计算、AI 基础软件与向量数据库、大模型与行业智能体、数据中心运营五大核心方向布局，优先布局能够紧跟英伟达架构迭代、在推理加速、高速互联、液冷与高密度机架、智能体调度与安全、向量检索等环节具备核心技术与客户壁垒的企业，同时把握国内 AI 芯片、服务器、基础软件与云服务的替代红利，充分受益于万亿级 AI 新基建与智能体时代带来的长期投资机遇。

八、风险提示

1、技术发展不及预期;

2、市场竞争加剧;

3、地缘政治影响。

投资评级定义

公司评级		行业评级	
强烈推荐	预期未来 6 个月内股价相对市场基准指数升幅在 15%以上	看好	预期未来 6 个月内行业指数优于市场指数 5%以上
推荐	预期未来 6 个月内股价相对市场基准指数升幅在 5%到 15%	中性	预期未来 6 个月内行业指数相对市场指数持平
中性	预期未来 6 个月内股价相对市场基准指数变动在-5%到 5%内	看淡	预期未来 6 个月内行业指数弱于市场指数 5%以上
卖出	预期未来 6 个月内股价相对市场基准指数跌幅在 15%以上		

免责声明

国新证券股份有限公司市场研究部
地址：北京市朝阳区朝阳门北大街 18 号中国人保寿险大厦 11 层（100020）
传真：010－85556155 网址：**www.crsec.com.cn**