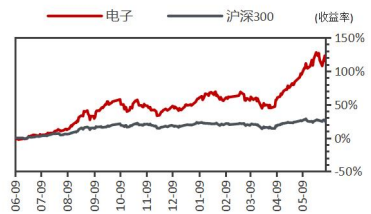


行业及产业
电子

强于大市

一年内行业指数与沪深 300 指数对比走势：



资料来源：聚源数据，爱建证券研究所

相关研究

《电子行业周报：长鑫科技 IPO 迎来关键节点，Intel 加码玻璃基板与硅光子》2026-06-01

《电子行业专题报告：华为发布 τ 定律，助力后摩尔时代半导体产业发展》2026-05-28

《电子行业周报：AI 算力高景气延续国产存储替代加速》2026-05-25

《电子行业跟踪报告：电子布行业具备估值提升潜力》2026-05-25

《电子行业专题报告：AI Glasses 开启智能穿戴时代》2026-05-15

证券分析师

许亮
S0820525010002
0755-83562506
xuliang@ajzq.com

联系人

朱俊宇
S0820125040021
021-32229888-25520
zhujunyu@ajzq.com

Vera Rubin 量产提速，RTX Spark 打开终端 AI 新空间

——电子行业专题报告

投资要点：

- **事件：**2026 年 6 月 1 日，NVIDIA CEO 黄仁勋于台北 Computex GTC 大会发表主题演讲。黄仁勋宣布，下一代 Vera Rubin 平台已全面投产，同时正式推出基于 ARM 架构、适配 Windows 系统的终端产品 RTX Spark，标志着 NVIDIA 正式切入长期由 Intel、AMD、Apple、Qualcomm 主导的 PC 市场。
- **NVIDIA 于 2026 年 1 月 CES 首次公开 Vera Rubin 架构与技术路线，2026 年 3 月 GTC 大会正式发布平台并启动量产，2026 年 6 月 Computex 进一步确认全面量产与交付节奏。**该平台搭载 NVIDIA 自研 Vera CPU 与 Groq 3 LPU 推理加速器，构筑 CPU+GPU+LPU 异构协同计算架构，叠加网络、存储、安全专用芯片，打造完整全栈 AI 算力体系。依托 2025 年底英伟达与 Groq 达成的 200 亿美元级非独家技术授权协议，平台融合定制芯片、整机温水浸没液冷、CPO 光电共封装等核心技术，针对性优化 AI 推理性能，弥补传统 GPU 推理短板，显著提升数据中心散热与网络能效，是适配大规模 AI 工厂的核心算力基础设施。
- **NVIDIA Vera Rubin 综合性能表现优异。**平台依托 Vera CPU、Groq 3 LPU 双专用芯片，搭建 CPU+GPU+LPU 全栈协同算力体系：1) Vera CPU 搭载自研 Olympus 定制核心，在核心规模、算力、内存带宽及容量等关键指标上较前代 Grace CPU 实现全方位升级；2) Groq 3 LPU 主打超低延迟推理能力，有效弥补传统 GPU 在 AI Agent 交互任务中的适配缺陷，配套 Groq 3 LPX 加速器机架可进一步实现高吞吐、低时延的规模化推理；3) 平台采用 100% 全液冷散热架构，大幅提升散热效率与整机柜交付效率；4) 英伟达量产落地 CPO 光电共封装 Spectrum-X 交换机，通过光铜混合扩展方案支撑算力集群规模化扩容。
- **此外，NVIDIA 还推出了面向 PC 终端的全新 SoC 芯片 RTX Spark，持续完善终端 AI 芯片产品线布局。**1) 初代 RTX Spark 由英伟达联合联发科协同研发，采用 TSMC 3nm 工艺、集成 700 亿晶体管，采用 CPU+GPU 一体化 SoC 设计；搭载 Blackwell 架构 GPU 与 20 核 Grace CPU，通过 NVLink-C2C 高速互联实现算力高效互通，硬件规格相较传统 PC 处理器大幅提升。2) 产品迭代路径清晰：当前采用 Grace CPU+Blackwell GPU 组合，2027 年将升级为 Vera CPU+Rubin GPU、LPDDR6 内存与 CX9 高速网卡；2029-2030 年新一代产品将换装 Rosa CPU、Feynman 架构 GPU 与 CX10 网卡。3) 芯片最高支持 128GB 统一内存、300GB/s 内存带宽，稀疏 FP4 AI 算力可达 1PFLOPS，能够在终端本地承载大模型推理运算。4) 依托 NVIDIA 跨端统一架构，RTX Spark 打通了云端 Vera Rubin 与终端 PC 的软硬件链路，进一步完善英伟达从数据中心到个人终端的全栈 AI 产品矩阵。
- **投资建议：**NVIDIA 在 Computex GTC2026 上敲定 Vera Rubin 全面量产交付节奏，同步发布全新 SoC 芯片 RTX Spark，有望拉动 SRAM、浸没式液冷、CPO 光模块、PC 端 AI SoC 产业链迎来增量机遇。1) Groq 3 LPU 与 Rubin GPU 异构协同，妥善解决传统 GPU 推理适配不佳、端到端时延偏高的问题，LPU 内置大容量片上 SRAM 直接拉动上游存储需求；2) 平台全栈采用 45℃温水浸没液冷，相较风冷与传统液冷能效优势显著，加速液冷方案在 AI 数据中心落地替代；3) Spectrum-X CPO 交换机落地，光铜混用的集群扩容模式拓宽 CPO 行业成长空间，带动光芯片、先进封装、高速光模块产业链景气度提升；4) RTX Spark 的发布标志 NVIDIA 正式切入 PC 处理器市场，产品分代迭代落地将持续拉动 PC 端 AI 芯片、高速配套内存需求扩容。建议重点关注国产 SRAM、液冷设备、CPO 光器件、PC AI 芯片相关上市公司的投资机会。
- **风险提示：**1) 技术迭代不及预期；2) 行业竞争加剧风险；3) 下游需求波动风险。

目录

1. 黄仁勋于 NVIDIA GTC 台北大会发布主题演讲	4
1.1 NVIDIA Vera Rubin 宣布大批量量产	4
1.2 NVIDIA 开辟 RTX Spark 产品线	7
2. 风险提示	9

图表目录

图表 1 : NVIDIA Vera Rubin 核心芯片与组件示意图	4
图表 2 : NVIDIA 为 Rubin 平台开发了全新 Vera CPU	5
图表 3 : Vera CPU 结构示意图	5
图表 4 : Vera CPU 较 Grace CPU 性能提升显著	5
图表 5 : Rubin GPU 与 Groq 3 LPU	6
图表 6 : NVIDIA Groq 3 LPU 芯片结构示意图	6
图表 7 : NVIDIA Groq 3 LPX 参数梳理	6
图表 8 : NVIDIA 风冷、传统液冷降温性能梳理	7
图表 9 : 黄仁勋宣布 Vera Rubin 全面生产	7
图表 10 : NVIDIA 持续拓展 RTX Spark 产品线	8
图表 11 : RTX-Spark 性能示意图	8
图表 12 : RTX Spark 未来生产产线	9

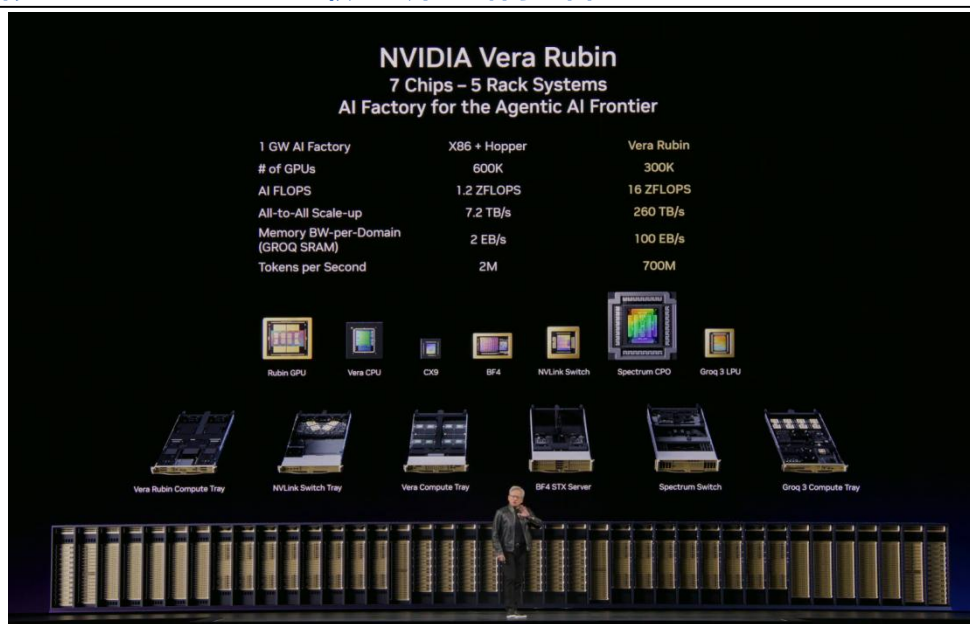
1. 黄仁勋于 NVIDIA GTC 台北大会发布主题演讲

事件：2026 年 6 月 1 日，NVIDIA CEO 黄仁勋于台北 Computex GTC 大会发表主题演讲。黄仁勋宣布，下一代 Vera Rubin 平台已全面投产，同时正式推出基于 ARM 架构、适配 Windows 系统的终端产品 RTX Spark，标志着 NVIDIA 正式切入长期由 Intel、AMD、Apple、Qualcomm 主导的 PC 市场。

1.1 NVIDIA Vera Rubin 宣布大批量量产

NVIDIA 于 2026 年 1 月 CES 首次公开 Vera Rubin 架构与技术路线，2026 年 3 月 GTC 大会正式发布平台并宣布量产，2026 年 6 月 Computex 进一步确认全面量产与交付节奏。该平台搭载 NVIDIA 自研 Vera CPU 与 Groq 3 LPU 推理加速器两大专用芯片，形成“CPU+GPU+LPU”核心异构协同计算架构，并整合网络/存储/安全专用芯片，构建全栈 AI 算力平台。依托 2025 年底英伟达与 Groq 达成的 200 亿美元级非独家技术授权协议的产业基础，平台整合专用定制芯片、整机全温水浸没液冷、CPO 光电共封装等核心技术协同升级，有效优化 AI 推理性能、弥补传统 GPU 推理场景短板，大幅提升数据中心散热与网络能效，是适配大规模 AI 工厂的核心高效算力基础设施。

图表 1：NVIDIA Vera Rubin 核心芯片与组件示意图

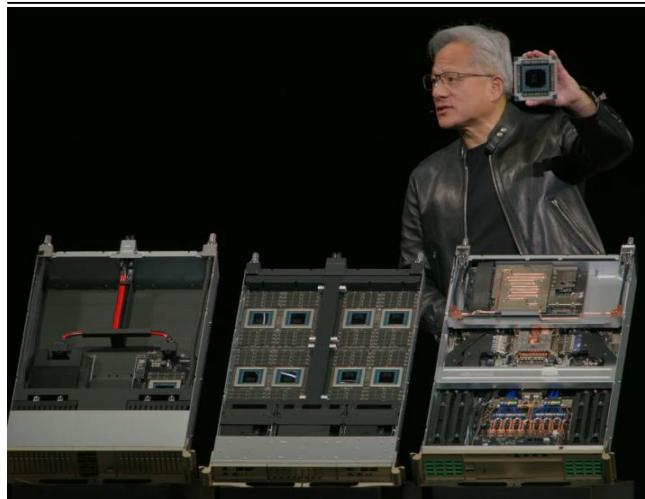


资料来源：NVIDIA GTC 2026 发布会视频，爱建证券研究所

■ Vera CPU:面向智能体推理的专用 CPU

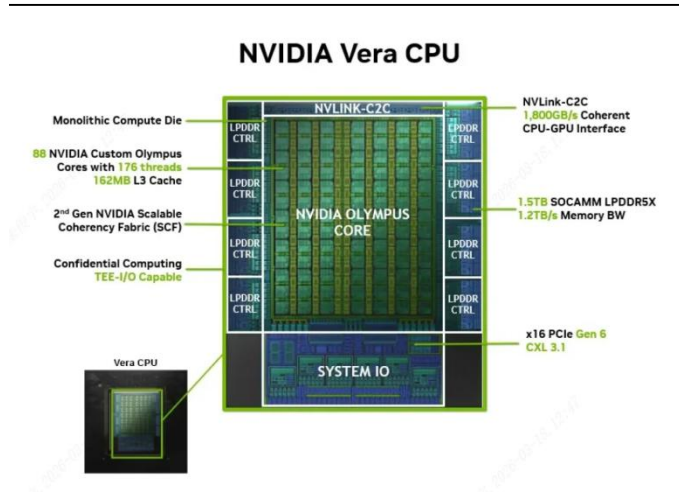
Vera CPU 是 NVIDIA 专为 Vera Rubin 平台打造的数据中心级专用芯片，针对大规模 AI 智能体推理场景深度优化，是当前 AI 工厂推理场景下能效最优的专用 CPU 产品。

图表 2: NVIDIA 为 Rubin 平台开发了全新 Vera CPU



资料来源: NVIDIA GTC 2026 发布会视频, 爱建证券研究所

图表 3: Vera CPU 结构示意图



资料来源: NVIDIA, 爱建证券研究所

硬件配置上, Vera CPU 搭载 88 颗 NVIDIA 自研 Olympus 定制核心, 采用 176 线程空间多线程架构; 具备 1.8 TB/s 的 NVLink-C2C 一致性内存互联带宽、1.2 TB/s 的 LPDDR5X 内存带宽, 最高支持 1.5 TB 系统内存, 内存容量为前代 Grace CPU 的 3 倍, 同时原生支持机架级机密计算。整体来看, Vera CPU 的核心算力与数据压缩性能较 Grace CPU 实现翻倍升级。

图表 4: Vera CPU 较 Grace CPU 性能提升显著

	Grace CPU	Vera CPU
核心	72 Neoverse V2 cores	88 NVIDIA Custom Olympus cores
线程	72	176 Spatial Multithreading
L2 Cache per core	1MB	2MB
Unified L3 Cache	114MB	162MB
内存带宽 (BW)	Up to 512GB/s	Up to 1.2TB/s
内存容量	Up to 480GB LPDDR5X	Up to 1.5TB LPDDR5X
SIMD	4x 128b SVE2	6x 128b SVE2 FP8
NVLINK-C2C	900GB/s	1.8TB/s
PCIe/CXL	Gen5	Gen6/CXL 3.1
机密计算	NA	支持

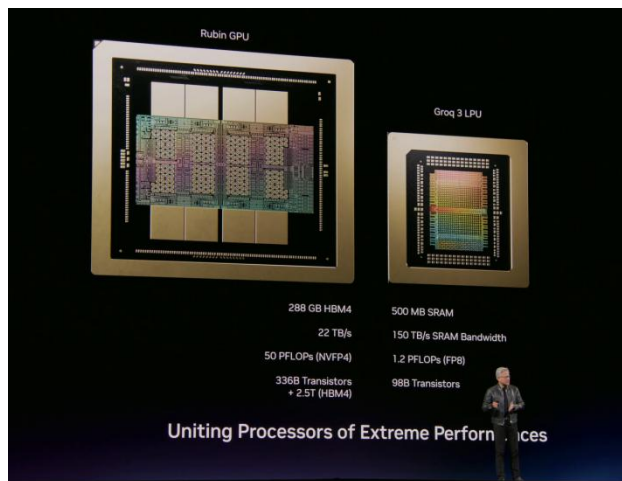
资料来源: NVIDIA DEVELOPER, 爱建证券研究所

■ Groq 3 LPX/LPU:填补了 GPU 推理短板

AI 智能体时代下, 大模型训练依赖 GPU 的高吞吐算力, 但在高交互、低延迟的智能体推理任务中, 传统 GPU 架构存在适配性差、端到端延迟过高的痛点。为此 NVIDIA 为 Vera Rubin 平台定制 Groq 3 LPU 专用处理器 (单芯片), 主打极致低延迟推理, 与 Rubin GPU 形成高效协同的异构算力组合, 补齐传统 GPU 推理短板。

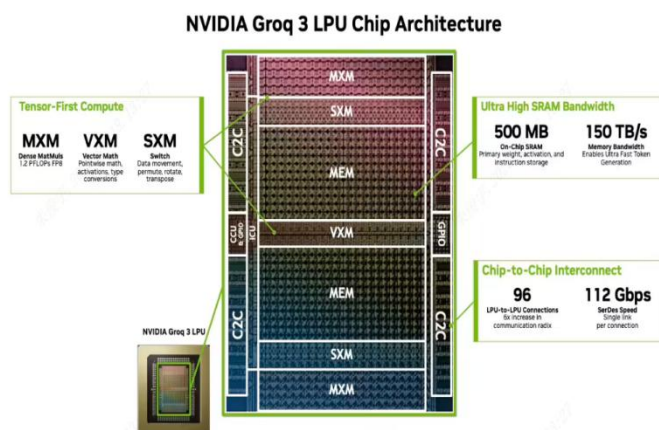
核心参数方面, Groq 3 LPU 内置 500 MB 片上 SRAM, 片上带宽高达 150 TB/s, 远超传统 GPU 依赖外部 HBM 内存的架构, 可大幅降低数据访问延迟; 芯片 FP8 算力达 1.2PFLOPs, 晶体管规模 980 亿。

图表 5: Rubin GPU 与 Groq 3 LPU



资料来源: NVIDIA GTC 2026 发布会视频, 爱建证券研究所

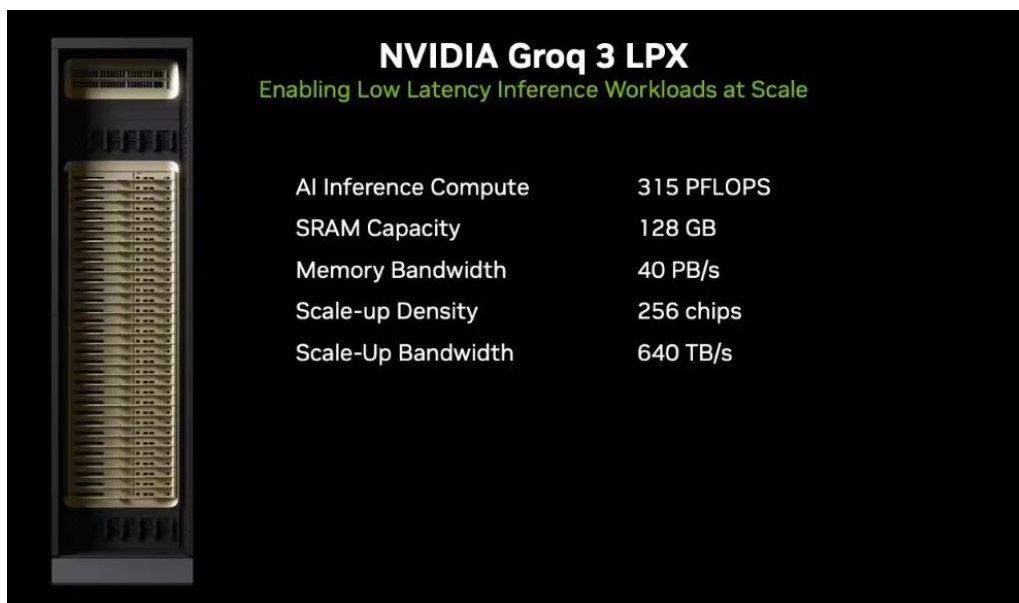
图表 6: NVIDIA Groq 3 LPU 芯片结构示意图



资料来源: NVIDIA, IT 之家, 爱建证券研究所

同时 NVIDIA 推出 Groq 3 LPX 推理加速器机架, 单机架集成 256 颗 LPU, 可提供 128 GB 片上 SRAM 与 640 TB/s 机架级带宽。该架构将多颗 LPU 整合为统一逻辑处理器, 通过确定性数据流执行推理任务, 有效提升大模型 token 生成的吞吐效率, 实现超低延迟推理。

图表 7: NVIDIA Groq 3 LPX 参数梳理



资料来源: NVIDIA, IT 之家, 爱建证券研究所

■ 新型液冷技术逐步替代传统风冷与常规液冷

Vera Rubin 采用 100%全液冷架构, 采用 45°C温水直接浸没液冷 (WDC) 方案实现高效散热, 大幅降低数据中心制冷成本; 同时内部布线简化, 使整机柜现场部署时间由传统 2 天压缩至 2 小时, 显著提升交付效率。据 NVIDIA 《Comparison of Air-Cooled versus Liquid-Cooled NVIDIA GPU Systems》数据, 传统液冷仅能将 GPU 温度控制在 46–54°C, 风冷则为 55–71°C。

图表 8：NVIDIA 风冷、传统液冷降温性能梳理

PERFORMANCE SUMMARY TABLE			
	Air-Cooled	Liquid-Cooled	Improvement
GPU Temperature Range	55-71°C	46-54°C	9-17°C Reduction
Stress Test Throughput	Baseline	+17%	17% Increase
Average Training Time	Baseline	5.2kW	1.4% Faster
Node Power (Production Workload)	6.2kW	5.2kW	1.0kW (16%) Reduction
Power Savings Range	-	100W-1.5kW	Scales with Workload
GPU Power Reduction	-	~150W Average	Per 8-GPU node

资料来源：NVIDIA 《Comparison of Air-Cooled versus Liquid-Cooled NVIDIA GPU Systems》，爱建证券研究所

■ CPO 光电共封装技术的持续突破

NVIDIA 还推出了全球首款 CPO（光电共封装）Spectrum-X 以太网交换机。它通过将光模块直接集成到芯片封装内，实现电信号与光信号的直接转换，可显著提升数据中心网络带宽与能效。该技术由 NVIDIA 与台积电联合开发，目前已进入量产阶段。

黄仁勋表示，在 Rubin 平台中，现有 Oberon 机架架构除铜线互联外，还将搭配支持 CPO 的 Spectrum 6 交换机实现光互联 Scale-out（水平扩展），可将 NVL72 系统规模拓展至 NVL576；借助 Spectrum-X 共封装光学技术进行集群扩展，可大幅提升数据中心基础设施的能效与系统稳定性。2028 年推出的 Feynman 平台中，Kyber 机架将同时支持铜线与 CPO 两种 Scale-up（垂直扩展）方式，系统规模可达 NVL1152。

本次 Computex GTC 大会上，黄仁勋宣布 Vera Rubin 已实现大规模量产。凭借芯片与机柜的一体化设计，NVIDIA 目前拥有全球最低的 AI 推理 Token 成本。

图表 9：黄仁勋宣布 Vera Rubin 全面生产



资料来源：NVIDIA Computex GTC 主题演讲，爱建证券研究所

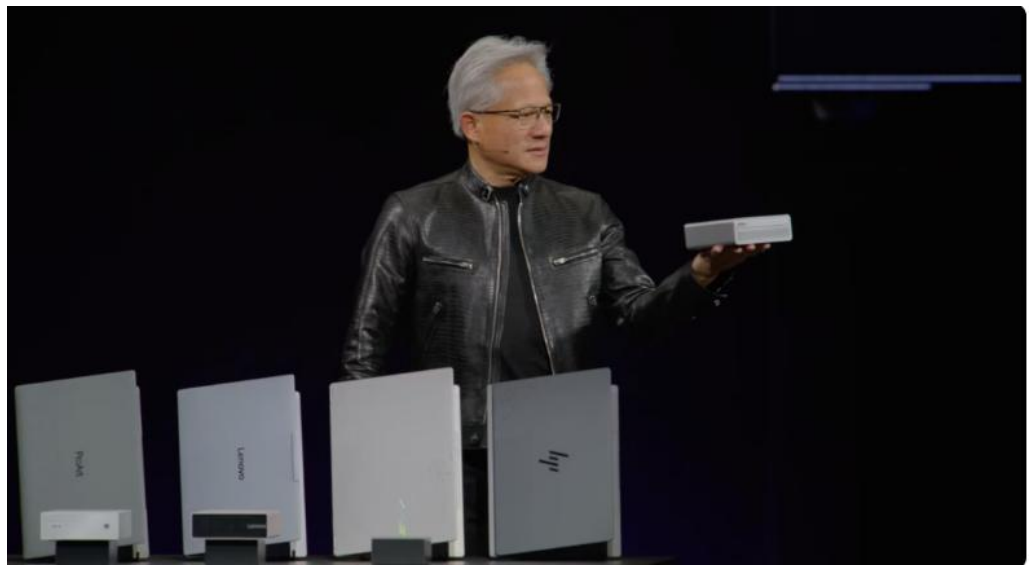
作为面向特定工作负载打造的芯片，Vera CPU 的首批客户包括 Anthropic、OpenAI、SpaceX 和 Oracle。

1.2 NVIDIA 开辟 RTX Spark 产品线

除 Vera Rubin 官宣量产外，NVIDIA 于 GTC 2026 发布面向 PC 市场的全新芯片 RTX

Spark，并持续拓展该产品线。

图表 10: NVIDIA 持续拓展 RTX Spark 产品线



资料来源：NVIDIA Computex GTC 主题演讲，爱建证券研究所

RTX Spark 由 NVIDIA 联合联发科共同研发，芯片采用台积电 3nm 制程、集成 700 亿晶体管，为 CPU 与 GPU 整合一体的 SoC 方案。GPU 搭载 Blackwell 架构，配备 6144 个 CUDA 核心与第五代 Tensor Core，支持 FP4 算力规格，稀疏 FP4 人工智能算力可达 1 PFLOPS；GPU 依托 NVLink-C2C 互连链路与 20 核 Grace CPU 对接，互连带宽 600GB/s。存储端最高可搭载 128GB 统一内存，内存带宽 300GB/s。

图表 11: RTX-Spark 性能示意图

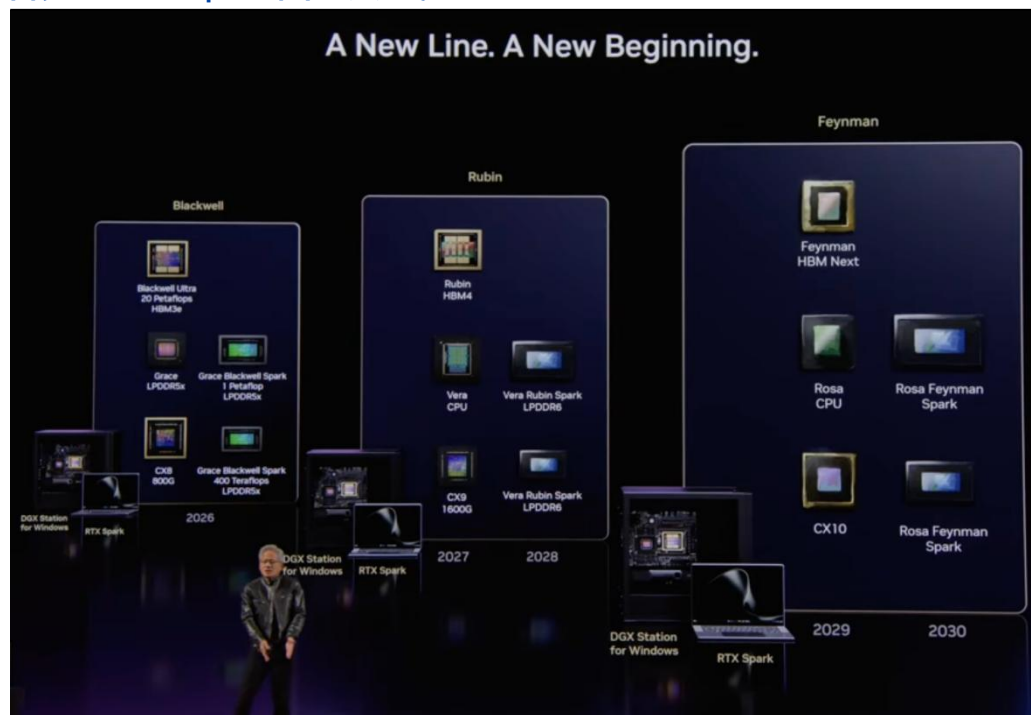


资料来源：NVIDIA，Storage Review，爱建证券研究所

初代 RTX Spark 采用 Grace CPU+Blackwell GPU 组合；产品明年迭代将换装 Vera CPU、Rubin GPU，内存升级至 LPDDR6，配套 CX9 网卡、实现 1600G 带宽。2029-2030 年公司将推出下一代 RTX Spark，CPU 换用 Rosa 架构、GPU 升级为新

近发布的 Feynman 架构，网卡迭代至 CX10。

图表 12: RTX Spark 未来生产产线



资料来源：NVIDIA，Storage Review，爱建证券研究所

2. 风险提示

- 1) 技术迭代不及预期风险：**Vera Rubin、RTX Spark 相关架构迭代、CPO 光电共封装、全浸没液冷产业化落地进度或低于英伟达规划，若技术研发受阻、良率爬坡缓慢，产业链上游零部件需求释放节奏延后。
- 2) 行业竞争加剧风险：**SRAM、新型液冷、CPO 光芯片领域行业竞争持续加剧，国产产业链相关公司或将面临订单流失、产品价格下滑、毛利率承压的风险。
- 3) 下游需求波动风险：**AI 算力、数据中心等下游行业需求受宏观经济、政策调整、技术路线变更等因素影响，若需求不及预期，将直接影响 SRAM、液冷、CPO 等相关产品的市场需求。

爱建证券有限责任公司

上海市浦东新区前滩大道 199 弄 5 号

电话: 021-32229888

传真: 021-68728700

服务热线: 956021

邮政编码: 200124

邮箱: ajzq@ajzq.com

网址: <http://www.ajzq.com>

评级说明

投资建议的评级标准

报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 个月内的相对市场表现，也即以报告发布日后的 6 个月内的公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A 股市场：沪深 300 指数（000300.SH）；新三板市场：三板成指（899001.CSI）（针对协议转让标的）或三板做市指数（899002.CSI）（针对做市转让标的）；北交所市场：北证 50 指数（899050.BJ）；香港市场：恒生指数（HIS.HI）；美国市场：标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）。

股票评级

买入	相对同期相关证券市场代表性指数涨幅大于 15%
增持	相对同期相关证券市场代表性指数涨幅在 5% ~ 15% 之间
持有	相对同期相关证券市场代表性指数涨幅在 -5% ~ 5% 之间
卖出	相对同期相关证券市场代表性指数涨幅小于 -5%

行业评级

强于大市	相对表现优于同期相关证券市场代表性指数
中性	相对表现与同期相关证券市场代表性指数持平
弱于大市	相对表现弱于同期相关证券市场代表性指数

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告采用信息和数据来自公开、合规渠道，所表述的观点均准确地反映了我们对标的证券和发行人的独立看法。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法可能存在局限性，请谨慎参考。

法律主体声明

本报告由爱建证券有限责任公司（以下统称为“爱建证券”）证券研究所制作，爱建证券具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管。

本报告是机密的，仅供我们的签约客户使用，爱建证券不因收件人收到本报告而视其为爱建证券的签约客户。本报告中的信息均来源于我们认为可靠的已公开资料，但爱建证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供签约客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户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、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，爱建证券及其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测后续可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，爱建证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

版权声明

本报告版权归属爱建证券所有，未经爱建证券事先书面许可，任何机构或个人不得以任何形式翻版、复制、转载、刊登和引用。否则由此造成的一切不良后果及法律责任由私自翻版、复制、转载、刊登和引用者承担。版权所有，违者必究。