东吴证券 SOOCHOW SECURITIES

电子行业点评报告

百万 Token 时代来临，Rubin CPX 重塑推理架构与产业链

2025 年 09 月 10 日

增持（维持）

证券分析师 陈海进
执业证书：S0600525020001
chenhj@dwzq.com.cn
研究助理 解承堯
执业证书：S0600125020001
xiechy@dwzq.com.cn

投资要点

- **Rubin CPX 切入百万 Token 痛点，重塑推理架构基础。**过去一年，随着生成式 AI 进入规模化落地阶段，行业对“长上下文”的需求快速上升。无论是企业级知识库问答、代码生成，还是多模态长视频生成，均需要模型在极大输入序列下保持推理准确性与计算效率。然而，现有 GPU 在应对超长上下文时普遍存在内存带宽瓶颈与计算冗余，导致算力利用率不足。英伟达于 2025 年 9 月发布的 Rubin CPX，正是为解决这一痛点而生，标志着 NVIDIA 将推理场景的架构优化推向新高度。Rubin CPX 是一款专为“百万级上下文”推理场景设计的专用加速处理器（Contextual Processing eXtension）。其设计目标是通过硬件与架构优化，提升对超长上下文场景的吞吐与能效，并在机架级系统，如 Vera Rubin NVL144 CPX 中与 Rubin GPU 及 Vera CPU 协同构成面向大规模推理的整体平台。

- **上下文与生成任务分工，实现算力利用率与效率提升。**在大模型推理过程中，可大体分为两类任务：一是“上下文分析”，即对超长输入序列进行编码、筛选与压缩，以便后续生成环节调用；二是“生成任务”，即基于上下文信息进行逐 Token 的预测输出。二者在计算负载和性能需求上差异显著：上下文分析更依赖并行化处理和带宽利用，而生成任务则要求对计算延迟与单步性能进行极致优化。英伟达 Rubin CPX 定位为“上下文处理加速器”，负责对海量输入做高通量注意力与前置计算；而 Rubin 系列通用 GPU 则负责生成/输出阶段的持续带宽密集型任务。官方展示的 Vera Rubin NVL144 CPX 机架中，144 个 Rubin CPX（context）配合 144 个 Rubin GPU（generation）与 36 个 Vera CPU（调度/通用），共同提供完整服务能力，进而实现资源的高效利用，推理成本降低以及推理响应加速。

- **Rubin CPX 30PFLOPS，机架 8EFLOPS 算力，2026 年落地路径明确。**从技术参数看，Rubin CPX 单卡提供约 30 PFLOPS（NVFP4 精度）的算力，并搭载 128GB GDDR7 显存，同时内置视频编解码能力，可满足多模态场景需求。官方展示的 Vera Rubin NVL144 CPX 系统，由 144 张 CPX、144 张 Rubin GPU 和 36 个 Vera CPU 共同构成，其整体性能指标达到 8 ExaFLOPS 算力、100TB 高速内存与 1.7PB/s 内存带宽，相比上一代 GB300 NVL72 系统在上下文处理效率上实现数倍提升。根据公司规划，Rubin CPX 预计将在 2026 年底上市，与 Rubin GPU 与 Dynamo、TensorRT-LLM 等软件工具链一体化部署。

- **海外算力链受益加速，长上下文推理带来新增长动能。**从产业角度来看，Rubin CPX 的推出不仅是 NVIDIA 产品线的升级，更意味着海外算力基础设施进入“上下文与生成分工协作”的新阶段。随着百万 Token 推理与长视频生成成为 AI 应用的标配需求，硬件和软件的耦合度显著提高，算力产业链的价值量同步上升。无论是 GPU、存储、网络，还是配套的高速 PCB、光模块与封装工艺，相关厂商都有望深度受益。我们认为 Rubin CPX 的量产与落地，将成为全球算力需求加速释放的重要信号，产业链公司中长期成长空间将更加清晰。

- **产业链相关公司：** **PCB/CCL：**沪电股份、胜宏科技、生益电子、深南电路、景旺电子、广合科技、生益科技、南亚新材；**铜缆：**沃尔核材、博创科技、华丰科技；**光芯片/光器件：**博创科技、仕佳光子、太辰光、长光华芯、源杰科技；**服务器代工：**工业富联、华勤技术

- **风险提示：**供应链波动风险，下游需求不及预期，行业竞争加剧。

行业走势

相关研究

《端侧 AI 散热机遇，微泵液冷关注艾为/南芯》

2025-09-02

《苹果秋季发布会前瞻：AI 战略落地、iPhone 硬件自主化与可穿戴健康升级》

2025-08-31

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。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，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn