国金证券 SINOLINK SECURITIES

2025 年 03 月 19 日

GTC Keynote 点评

数据专题研究

证券研究报告

分析师：刘道明（执业 S1130520020004） liudaoming@gjzq.com.cn
联系人：黄晓军（执业 S1130122050092） huangxiaojun@gjzq.com.cn
联系人：麦世学（执业 S1130123100111） maishixue@gjzq.com.cn

内存加强版 GB300 正式发布，后续产品迭代节奏不及预期

摘要

- NVIDIA 发布了内存增强版 Blackwell 产品 Blackwell Ultra。NVL7 规格显示，单卡 Dense FP4 算力较 B200 提升 50%，HBM 配置升级至 288GB HBM3e。网络层面，采用 ConnectX 8 网卡替代 ConnectX 7，进一步提升性能。Blackwell Ultra 预计于 2025 年下半年出货。此前市场预期的 GB300 采用“SXM Puck”形式及 BGA 封装的设计未在本次发布会上得到确认，整体发布符合市场预期。
- NVIDIA 发布了 Blackwell 继代产品 Rubin 和 Rubin Ultra，并统一了 GPU die 计数标准。以 Rubin NVL144 为例，内含 144 颗 die，实际由 72 颗 Rubin 芯片组成。Rubin Ultra 单颗芯片整合 4 颗 die，NVL576 内仅有 144 颗 Rubin Ultra 芯片，低于此前预期的 NVL288。Rubin Ultra 采用纵向 Tray 结构，优化了机柜空间，预计将在大规模 Scale Up 场景中成为主流。值得注意的是，相较于 GTC 2024 发布的路线图，Rubin 系列的产品节奏出现约半年延迟，首款 Rubin 产品推迟至 2026 年下半年上市，而 Rubin Ultra 则需等待至 2027 年下半年，整体进度不及市场预期。此次延迟或与制程、封装及机柜层面的技术挑战有关，成为市场反应不佳的主要原因之一。
- NVIDIA 发布了 Spectrum-X Photonics 和 Quantum-X Photonics 硅光交换机平台，单端口速率达 1.6 Tb/s，总带宽最高 400 Tb/s，显著提升数据中心传输性能。Spectrum-X Photonics 提供最高 512 个 800 Gb/s 端口，Quantum-X Photonics 提供 144 个 800 Gb/s InfiniBand 端口，采用 200 Gb/s SerDes 技术，进一步提升传输效率。该系列交换机提升了 AI 集群的扩展性，为超大规模数据中心提供更优解决方案。
- NVIDIA 发布了 DGX Spark 和 DGX Station，进一步推动 AI 超算向个人桌面端普及。DGX Spark 采用 GB10 Blackwell Superchip，具备 128GB 统一内存和最高 4TB SSD，算力达 1,000 TOPS。DGX Station 搭载更强的 GB300 Blackwell Ultra Superchip，AI 性能达 20 PFLOPS，配备 784GB 统一内存，满足更高强度的 AI 训练和推理任务，进一步拓展 AI 计算的应用场景。
- NVIDIA 发布了开源推理服务框架 NVIDIA Dynamo，旨在优化大规模 AI 模型的推理部署。Dynamo 在运行 DeepSeek-R1 模型时，将请求处理能力提升多达 30 倍。其性能提升得益于解耦的 Prefill/Decode 阶段、动态 GPU 调度、LLM 感知请求路由、加速 GPU 间异步数据传输及 KV Cache 跨内存层级卸载机制。Dynamo 已在 GitHub 开源，并集成至 NVIDIA AI Enterprise 的 NVIDIA NIM 微服务，助力企业更高效部署 AI 推理模型。

风险提示

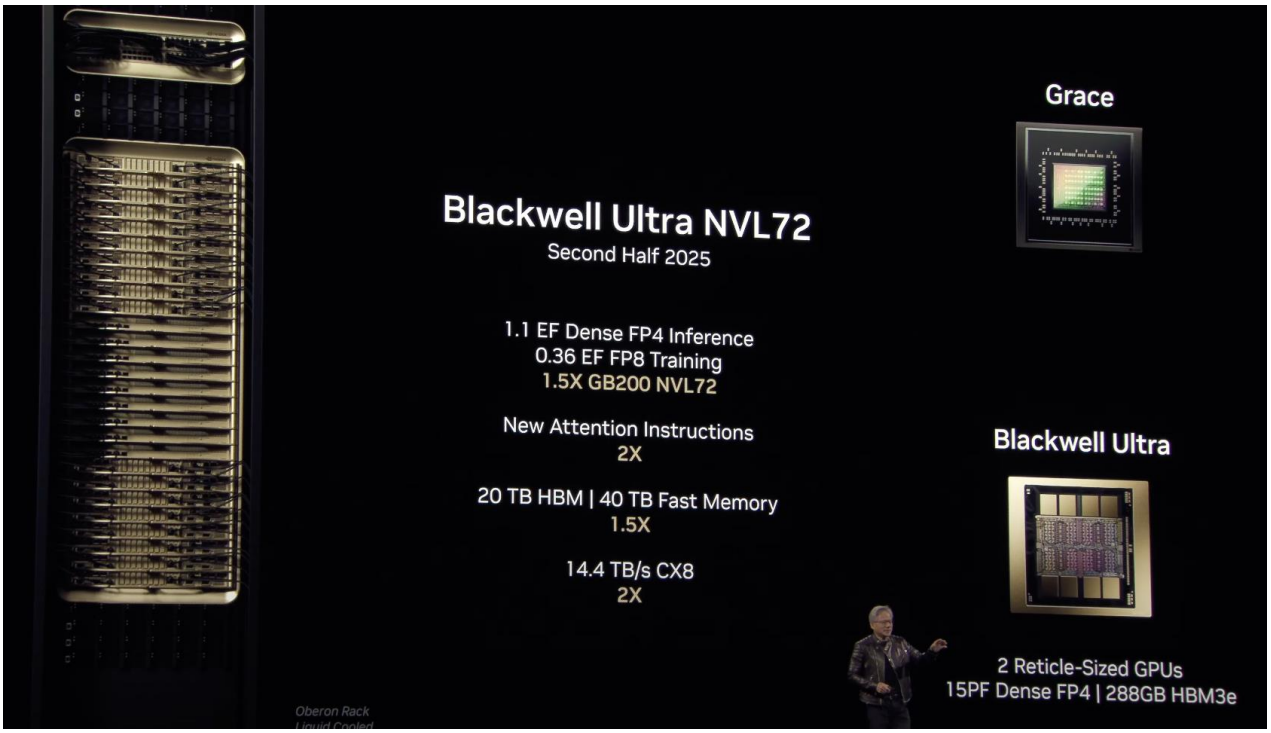
芯片制程发展与良率不及预期
中美科技领域政策恶化
智能手机销量不及预期

内容目录

内存加强版 Blackwell 产品 Blackwell Ultra 正式发布

英伟达正式发布其内存增强版 Blackwell 产品 Blackwell Ultra，从 NVL72 总体参数上来看，单卡 Dense FP4 算力相较 B200 提升 50%，单卡 HBM 配置提升至 288GB HBM3e，网络层面，ConnectX 8 网卡取代了之前的 ConnectX 7，预计将于 2025 年下半年出货。

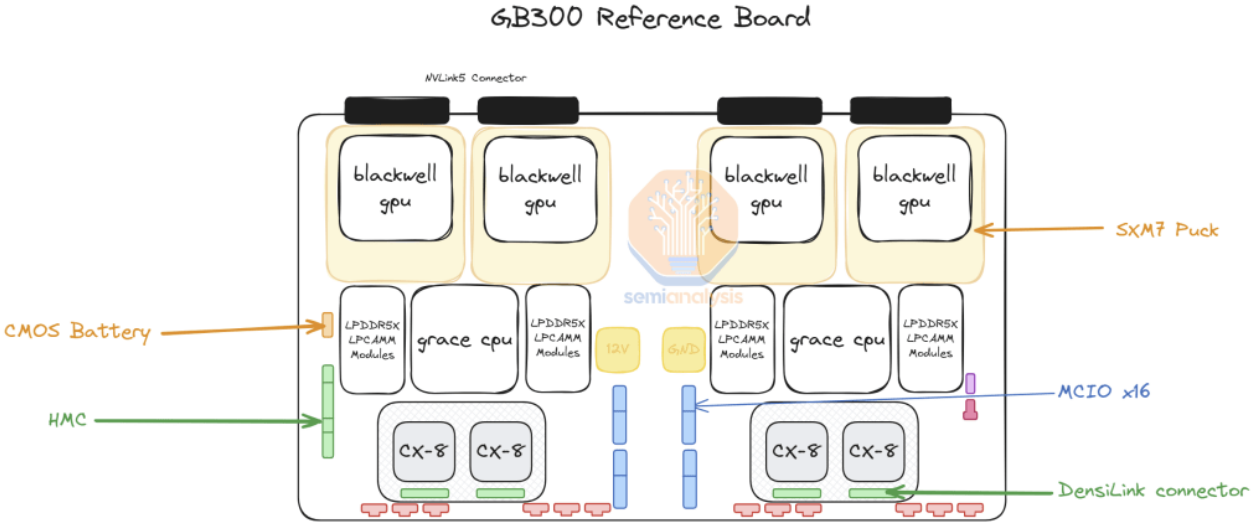
图表1：Blackwell Ultra 参数



来源：英伟达、国金证券研究所

此前市场预期，与 B200 系列产品所采用的整套 Bianca 主板设计不同，GB300 将以“SXM Puck”的形式提供，而 Grace CPU 将采用 BGA 封装形式。这一设计意味着 B300 可更快速地从主板上拆卸或更换，从而提升后期维护的便捷性。然而，这一消息并未在本次发布会上得到确认。

图表2：先前市场预期 B300 仅以 SXM Puck 的形式提供，给终端客户更大的定制空间



来源：Semianalysis、国金证券研究所

总体上，本次 Blackwell Ultra 产品发布符合市场预期。

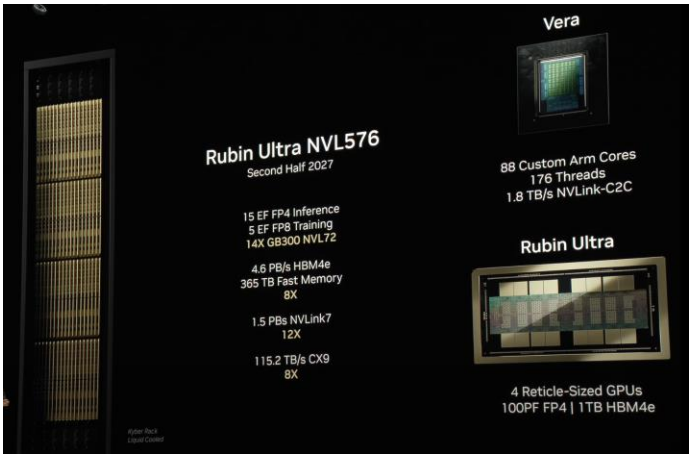
Vera Rubin 及后续产品 Roadmap 发布

在本次 GTC 大会上，NVIDIA 同步发布了 Blackwell 继代产品——Rubin 和 Rubin Ultra。值得注意的是，公司在此次发布会上对 GPU die 的计数标准进行了统一，明确将 die 数量作为衡量机柜内互联范围的单位。以 Vera Rubin NVL144 为例，NVL144 机柜内共包含 144 颗 die。鉴于单颗 Rubin 芯片由两颗 die 组成，这意味着 NVL144 实际上由 72 颗 Rubin 芯片构成。相比之下，Rubin Ultra 单颗芯片内集成了 4 颗 die。因此，尽管 NVL576 机柜内共有 576 颗 die，但实际上仅由 144 颗 Rubin Ultra 芯片组成。这和先前市场预期的芯片层面而非 die 层面的 NVL288（由 288 颗芯片组成）尚有差距。

图表3：Rubin NVL144 参数

来源：英伟达、国金证券研究所

图表4：Rubin Ultra NVL576 参数

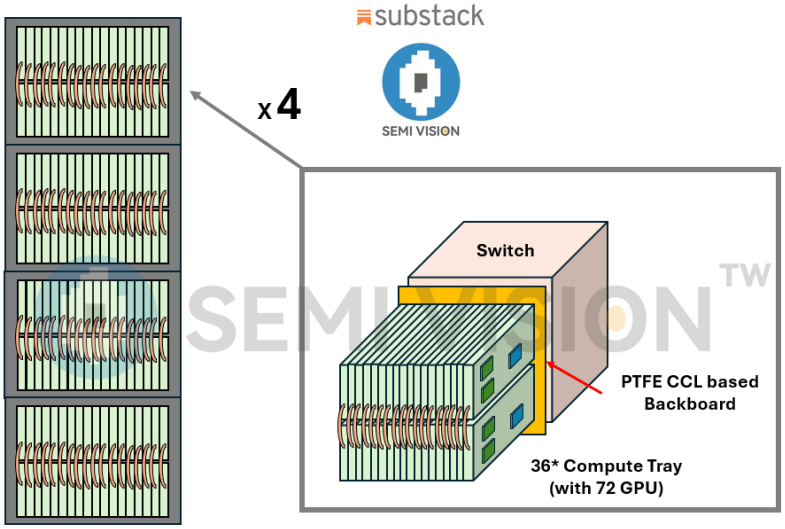


来源：英伟达、国金证券研究所

在单卡算力方面，Rubin 和 Blackwell Ultra 的芯片面积均接近两倍光罩极限。然而，Rubin 的算力超过 Blackwell Ultra 的三倍，这表明 Rubin 可能采用了比 N4 更先进的制程工艺。而 Rubin Ultra 的芯片面积接近四倍光罩极限，算力相应翻倍，意味着 Rubin Ultra 的算力提升更多依赖于更大的芯片面积。这与我们此前的观点一致，即在晶体管制程迭代放缓的背景下，GPGPU 的算力提升将更多依赖于更大的芯片和封装面积。

由于 Rubin NVL144 仍由 72 颗接近两倍光罩极限面积的芯片组成，其机柜结构与 Blackwell NVL72 类似。而 Rubin Ultra NVL576 由于芯片数量翻倍，机柜结构有所变化，可以看到机柜中所有的 Tray 均由横向放置改为纵向放置。此前市场对这一放置形式已有预期，从发布会的设计图来看，我们认为这一设计将在大规模 Scale Up 场景中成为主流方案。

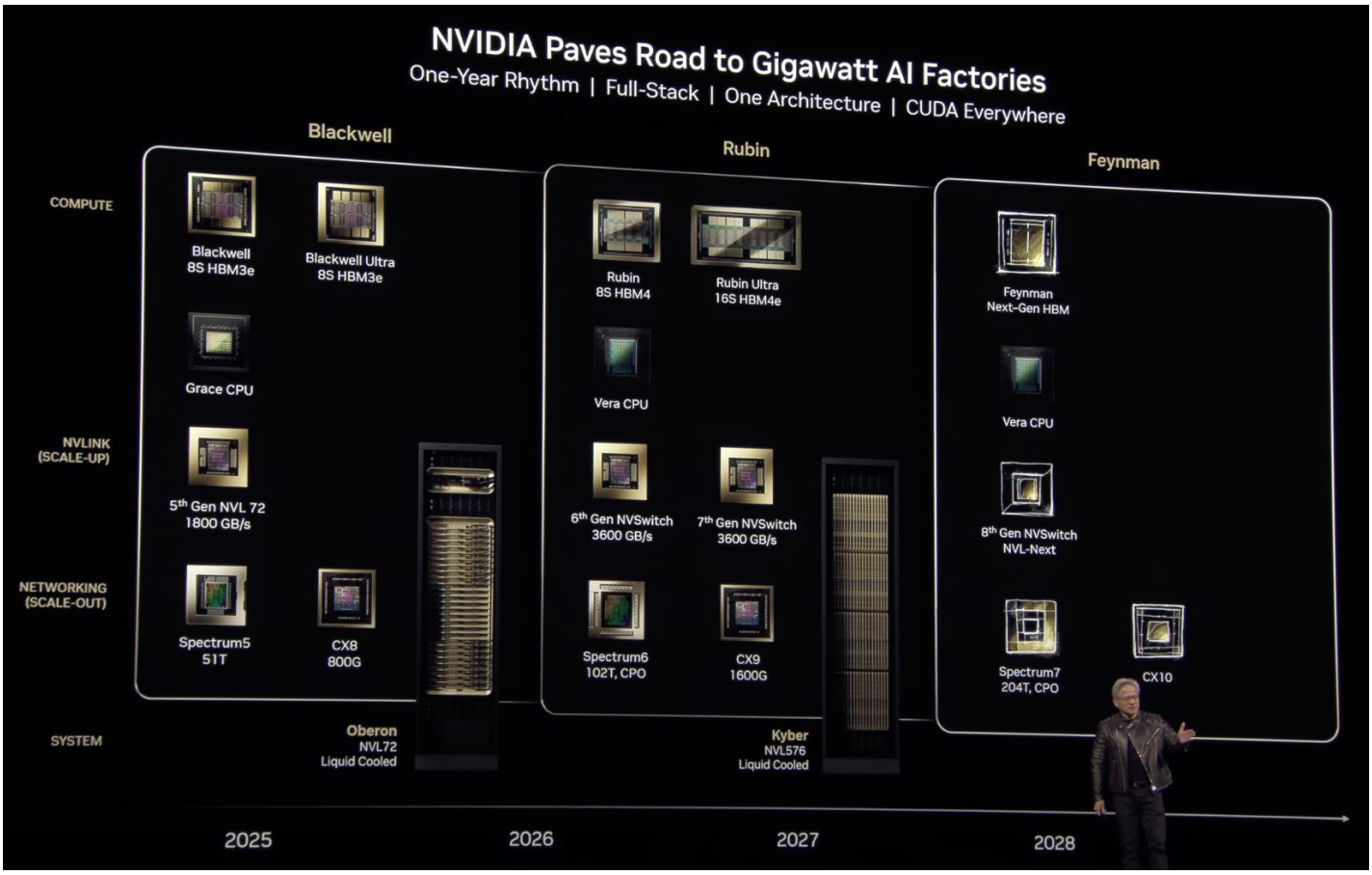
图表5：纵向 Compute Tray 设计概念图



来源：Semi Vision、国金证券研究所

公司在本次大会上同步发布了后续数据中心产品的路线图，披露了 Rubin 的继任产品代号为 Feynman。值得注意的是，与 GTC 2024 发布的路线图相比，本次 GTC 2025 发布的版本在产品迭代节奏上出现了约半年的延迟。根据 GTC 2024 的规划，Blackwell 的产品周期原定为 2024 至 2025 年，而 Rubin 则覆盖 2026 至 2027 年。然而，从本次 GTC 公布的时间线来看，首款 Rubin 产品的上市时间已推迟至 2026 年下半年，而 Rubin Ultra 则需等待至 2027 年下半年。我们认为，公司的产品迭代节奏放缓主要受限于制程、封装及机柜层面的挑战。这一进度延迟正是 GTC 发布会后市场反应不佳的核心原因。

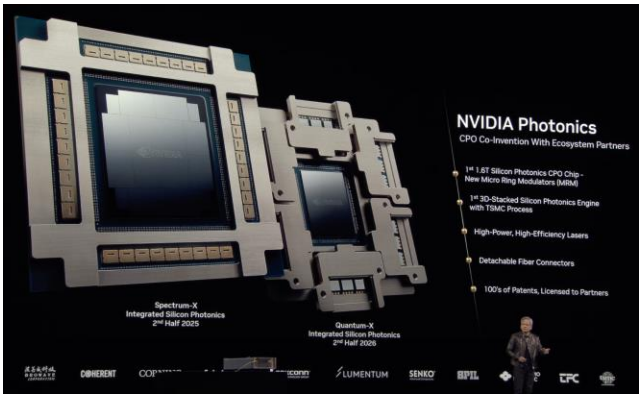
图表6：公司后续数据中心产品线路线图



来源：英伟达、国金证券研究所

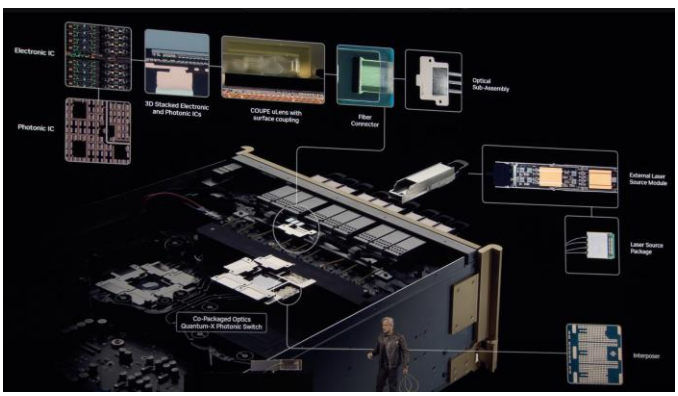
推出首个硅光交换机产品

图表7：Spectrum-X 和 Quantum-X



来源：英伟达、国金证券研究所

图表8：Quantum-X 结构图



来源：英伟达、国金证券研究所

在 GTC 2025 大会上，NVIDIA 推出了面向超大规模数据中心的全新网络交换机平台——Spectrum-X Photonics 和 Quantum-X Photonics，两者均采用硅光子（Silicon Photonics）技术。该系列新品将数据传输速率提升至每端口 1.6 Tb/s，总带宽最高可达 400 Tb/s，从而支持数百万颗 GPU 的高效协同运作。NVIDIA 表示，相较于传统网络解决方案，新的交换机平台具备更高带宽、更低功耗损耗以及更优异的可靠性。

Spectrum-X Photonics 以太网平台和 Quantum-X Photonics InfiniBand 平台均可实现每端口 1.6 Tb/s 的速率，达到当前顶级铜缆以太网解决方案的两倍。两者通过不同端口配置可实现高达 400 Tb/s 的总带宽。Spectrum-X Photonics 交换机提供多种配置选项，基础型号支持 128 个 800 Gb/s 端口或 512 个 200 Gb/s 端口，总带宽可达 100 Tb/s。更高规格的机型则提供 512 个 800 Gb/s 端口或 2,048 个 200 Gb/s 端口，总带宽高达 400 Tb/s。

Quantum-X Photonics 系列则采用 144 个 800 Gb/s InfiniBand 端口，并配备 200 Gb/s SerDes 技术，以进一步优化数据传输效率。与上一代网络解决方案相比，Quantum-X 平台的性能提升至两倍，且 AI 计算集群的可扩展性提升了五倍，使其成为应对高强度工作负载和构建超大规模 AI 集群的理想选择。

NVIDIA 的 Spectrum-X Photonics 以太网平台和 Quantum-X Photonics InfiniBand 平台采用了台积电（TSMC）的硅光子平台——Compact Universal Photonic Engine（COUPE）。该平台将基于 65nm 工艺的电子集成电路（EIC）与光子集成电路（PIC）相结合，并采用台积电的 SoIC-X 封装技术，实现高度集成。

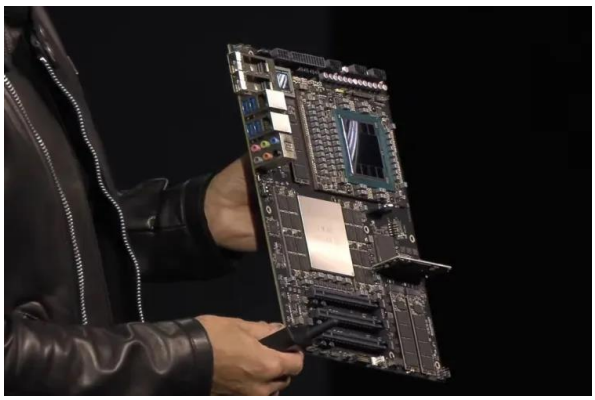
个人 AI 电脑 DGX Spark 与 DGX Station

在本次 GTC 2025 大会上，NVIDIA 发布了两款面向个人 AI 计算的新产品——DGX Spark 和 DGX Station，标志着 AI 超算能力正进一步向个人桌面端渗透。凭借更强大的计算性能与更便捷的部署方式，这两款设备有望在 AI 开发者、研究人员及数据科学家群体中引发广泛关注。

图表9：DGX Spark 结构图

来源：The Verge、国金证券研究所

图表10：DGX Station 母板



来源：The Verge、国金证券研究所

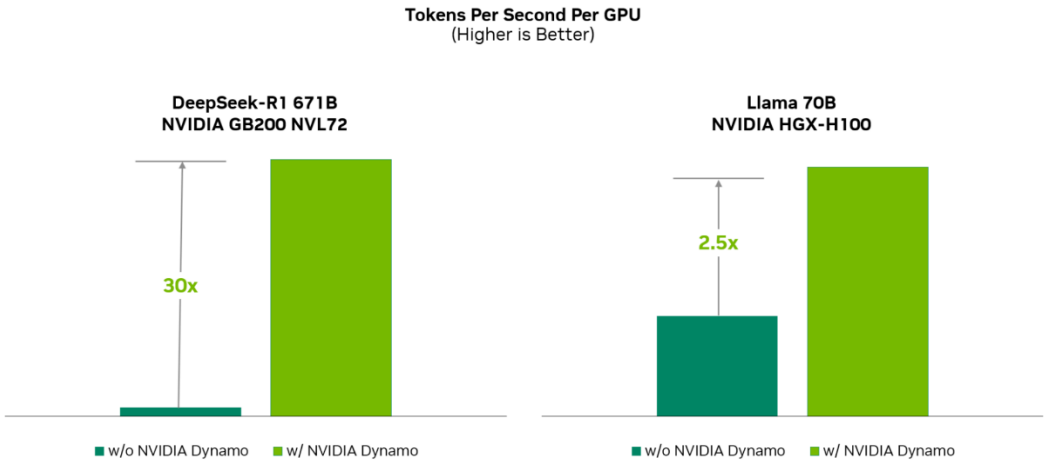
公司将 DGX Spark 定位为全球最小的 AI 超级计算机，外观类似 Mac Mini，售价为 3,000 美元。该设备搭载了基于 Grace Blackwell 平台的 GB10 Blackwell Superchip，集成第五代 Tensor Core，并支持 FP4 格式计算，专为桌面小型化设计而优化。尽管体积小巧，DGX Spark 依然具备强大的 AI 推理与微调能力，最高可实现 1,000 TOPS（每秒万亿次运算）的算力，支持 NVIDIA 最新的 Cosmos Reason 世界大模型及 GR00T N1 机器人基础模型。此外，Spark 配备 128GB 统一内存和最高 4TB 的 NVMe SSD，进一步增强其本地 AI 训练与推理性能，满足中小型企业、研究机构及独立开发者的日常 AI 任务需求。

相比之下，DGX Station 则以更强性能面向高强度 AI 开发和推理场景。其搭载的全新 GB300 Blackwell Ultra Superchip 具备高达 20 PFLOPS（每秒千万亿次运算）的 AI 性能，并搭配高达 784GB 统一内存，充分满足大型模型训练、推理及复杂 AI 工作负载的需求。凭借强大的性能配置，DGX Station 无疑将成为企业、科研机构及 AI 专业人员在本地 AI 开发中的理想选择。

开源分布式推理服务库 Dynamo

公司发布了全新的开源推理服务框架——**NVIDIA Dynamo**，专为生成式 AI 和推理模型的大规模部署而设计。Dynamo 在运行开源 DeepSeek-R1 模型时，将请求处理能力提升**多达 30 倍**，显著优化了推理性能和计算成本，尤其在 NVIDIA Blackwell 平台上表现突出。

图表11：Dynamo 使 Deepseek-R1 和 Llama 70B 推理性能显著提升



来源：英伟达、国金证券研究所

Dynamo 的性能突破源于多项创新设计，包括解耦的 Prefill 与 Decode 推理阶段以提升 GPU 吞吐量、动态 GPU 调度以优化资源利用、LLM 感知的请求路由避免 KV Cache 重复计算、加速 GPU 间异步数据传输缩短响应时间，以及 KV Cache 跨内存层级卸载机制以进一步提高系统吞吐量。Dynamo 已在 GitHub（ai-dynamo/dynamo）开源，并将集成至 NVIDIA AI Enterprise 的 NVIDIA NIM 微服务中，为企业用户提供更快速、更稳定的生产环境部署方案。Dynamo 的推出为 AI 推理性能提升提供了更具成本效益的新选择，进一步强化了 NVIDIA 在生成式 AI 推理领域的领先地位。

风险提示

1. 芯片制程发展与良率不及预期：半导体工艺的发展面临诸多挑战，主要包括技术瓶颈、良率提升难度、研发成本高企以及供应链不确定性等问题。随着工艺节点微缩变得愈发复杂，先进制程的实现难度和成本不断攀升，可能导致量产延迟，甚至影响产品性能和成本控制。此外，地缘政治风险和出口管制可能扰乱供应链，进一步拖累产能扩张。

2. 中美科技领域政策恶化：中美在 AI 领域竞争激烈，美国限制先进芯片和半导体对中国的出口，随着竞争的加剧，未来可能会推出更严格的限制政策，限制国内 AI 模型的发展。

3. 智能手机销量不及预期：智能手机销量与产品本身质量关系紧密，若产品本身有缺陷则智能手机销量可能收到影响。同时宏观经济变化也有可能导致消费者消费意愿发生变化从而影响智能手机销量。

特别声明：

国金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。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的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，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究