

行业点评

计算机

英伟达发布 NIM 软件包，简化 AI 部署

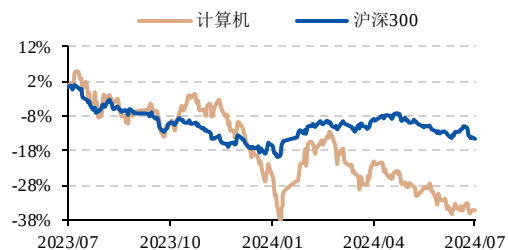
计算机行业点评

2024 年 07 月 30 日

评级 同步大市

评级变动： 维持

行业涨跌幅比较



%	1M	3M	12M
计算机	-4.38	-15.38	-35.42
沪深 300	-2.05	-5.40	-15.08

何晨

分析师

执业证书编号:S0530513080001
hechen@hncasing.com

黄奕景

研究助理

huangyijing@hncasing.com

相关报告

- 计算机行业 6 月月报：AI 和自主可控为大势所趋 2024-07-02
- 计算机行业 5 月月报：信创产业有望边际回暖，算力资本开支延续高景气 2024-06-03
- 计算机行业点评：安可测评结果出炉，信创订单有望释放 2024-05-21

重点股票	2023A		2024E		2025E		评级
	EPS (元)	PE (倍)	EPS (元)	PE (倍)	EPS (元)	PE (倍)	
浪潮信息	1.21	30.52	1.45	25.45	1.73	21.34	增持

资料来源：iFinD，财信证券

注：按照 7-29 收盘价计算

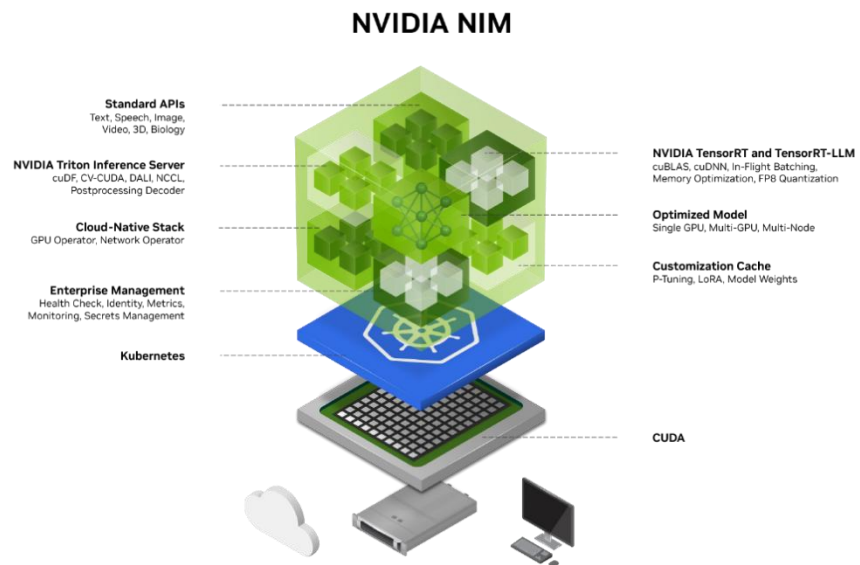
投资要点：

- **英伟达发布全新软件包 NIM。** 财联社 7 月 30 日电，英伟达公司正式发布了 Nvidia inference micro services(NIM)，这个软件包可以解决将 AI 用于特定目的所需的许多后勤问题。生成式 AI 是聊天机器人、语音识别和其他自动化人机交互背后的技术，通常需要协调众多硬件、软件和信息检索。许多公司不具备这方面的专业知识，所以英伟达试图就此提供服务并收取费用。英伟达创始人兼首席执行官黄仁勋周一在丹佛举行的 Siggraph 大会上展示并宣传了该最新产品。
- **什么是 NIM：NVIDIA NIM 是一套经过优化的云原生微服务，旨在缩短 AI 上市时间，并简化生成式 AI 模型在云、数据中心和 GPU 加速工作站的任何位置的部署。借助行业标准 API 和抽象化 AI 模型开发，开发者可以使用几行代码借助 NIM 部署 AI 模型。**
- **NIM 的核心优势：** 1) 随时随地部署；2) 使用行业标准 API 进行开发；3) 利用特定领域的模型；4) 在优化的推理引擎上运行；5) 支持企业级 AI；6) 加速 AI 模型随时可供部署。
- **投资建议：** NIM 作为英伟达在 AI 应用层的全新软件工具包，进一步巩固了英伟达在 AI 领域的竞争优势和护城河。国内的企业级 AI 开发者也有望通过英伟达中国的解决方案合作伙伴使用 NIM 服务进行 AI 模型部署。 建议关注：亚康股份、紫光股份、浪潮信息。
- **风险提示：** AI 技术发展不及预期；宏观环境不及预期；政策不及预期。

1 什么是 NIM?

NVIDIA NIM 是一套经过优化的云原生微服务，旨在缩短 AI 上市时间，并简化生成式 AI 模型在云、数据中心和 GPU 加速工作站的任何位置的部署。借助行业标准 API 和抽象化 AI 模型开发，开发者可以使用几行代码借助 NIM 部署 AI 模型。NIM 容器可与 Kubernetes(K8s)生态系统无缝集成，从而实现容器化 AI 应用程序的高效编排和管理。

图 1：NVIDIA NIM 工作原理



资料来源：英伟达官网，财信证券

2 NIM 核心优势

1) 随时随地部署

NIM 专为可移植性和可控性而构建，支持跨各种基础设施（从本地工作站到云再到本地数据中心）进行模型部署。其中包括 NVIDIA DGX、NVIDIA DGX 云、NVIDIA 认证系统、NVIDIA RTX 工作站和 PC。

NIM 通过预构建的容器和 Helm Chart 打包了优化模型，并在不同的 NVIDIA 硬件平台、云服务提供商和 Kubernetes 发行版中进行了严格验证和基准测试。这支持所有 NVIDIA 驱动的环境，并确保组织可以在任何地方部署其生成式 AI 应用。

2) 使用行业标准 API 进行开发

开发者可以通过符合每个领域行业标准的 API 访问 AI 模型，从而简化 AI 应用的开发。这些 API 与生态系统中的标准部署流程兼容，使开发者能够快速更新其 AI 应用(通常只需 3 行代码)。这种无缝集成和易用性有助于在企业环境中快速部署和扩展 AI 解决方案。

3) 利用特定领域的模型

NIM 包含特定于领域的 NVIDIA CUDA 库，以及为语言、语音、视频处理、医疗健康等各个领域量身定制的专用代码，可确保应用程序准确无误并与其特定用例相关。

4) 在优化的推理引擎上运行

NIM 针对每个模型和硬件设置利用经过优化的推理引擎，在加速基础设施上提供尽可能好的延迟和吞吐量。这降低了在扩展推理工作负载时运行推理工作负载的成本，并改善了最终用户体验。除了支持优化的社区模型外，开发者还可以通过使用从未离开数据中心边界的专有数据源对模型进行对齐和微调，从而实现更高的准确性和性能。

5) 支持企业级 AI

作为 NVIDIA AI Enterprise 的一部分，NIM 采用企业级基础容器构建，通过功能分支、严格的验证、通过服务级别协议提供的企业级支持以及针对 CVE 的定期安全更新，为企业 AI 软件提供坚实的基础。全面的支持结构和优化功能突出了 NIM 作为在生产环境中部署高效、可扩展和定制的 AI 应用的关键工具的作用。

6) 加速 AI 模型随时可供部署

支持多种 AI 模型，包括社区模型 NVIDIA AI 基础模型和 NVIDIA 合作伙伴提供的定制 AI 模型。NIM 支持跨多个领域的 AI 用例，包括大型语言模型 (LLM)、视觉语言模型 (VLM)，以及用于语音、图像、视频、3D、药物研发、医学成像等的模型。

开发者可以使用 NVIDIA 托管的云 API 测试新的生成式 AI 模型，或者通过下载 NIM 来自行托管模型，并在主要云提供商或本地使用 Kubernetes 快速部署，以减少开发时间、复杂性和成本。

NIM 微服务通过打包算法、系统和运行时优化并添加行业标准 API 来简化 AI 模型部署流程。这使开发者能够将 NIM 集成到其现有应用程序和基础设施中，而无需大量定制或专业知识。

借助 NIM，企业可以优化其 AI 基础架构，以更大限度地提高效率和成本效益，而无需担心 AI 模型开发的复杂性和容器化。在加速 AI 基础架构的基础上，NIM 有助于提高性能和可扩展性，同时降低硬件和运营成本。

对于希望为企业应用程序定制模型的企业，NVIDIA 提供微服务，以便跨不同领域定制模型。NVIDIA NeMo 使用专有数据为 LLM、语音 AI 和多模态模型提供微调功能。NVIDIA BioNeMo 通过越来越多的生成生物学化学和分子预测模型集合加速药物研发。NVIDIA Picasso 通过 Edify 模型实现更快的创意工作流程。这些模型在视觉内容提供商的许可库中进行训练，从而能够部署自定义的生成式 AI 模型，以创建视觉内容。

3 投资建议

NIM 作为英伟达在 AI 应用层的全新软件工具包，进一步巩固了英伟达在 AI 领域的

竞争优势和护城河。国内的企业级 AI 开发者也有望通过英伟达中国的解决方案合作伙伴使用 NIM 服务进行 AI 模型部署。建议关注：亚康股份、紫光股份、浪潮信息。

4 风险提示

AI 技术发展不及预期；行业需求不及预期；宏观经济状况不及预期；政策不及预期。

评级系统说明

以报告发布日后的 6—12 个月内，所评股票/行业涨跌幅相对于同期市场指数的涨跌幅度为基准。

类别	投资评级	评级说明
股票投资评级	买入	投资收益率超越沪深 300 指数 15%以上
	增持	投资收益率相对沪深 300 指数变动幅度为 5%—15%
	持有	投资收益率相对沪深 300 指数变动幅度为-10%—5%
	卖出	投资收益率落后沪深 300 指数 10%以上
行业投资评级	领先大市	行业指数涨跌幅超越沪深 300 指数 5%以上
	同步大市	行业指数涨跌幅相对沪深 300 指数变动幅度为-5%—5%
	落后大市	行业指数涨跌幅落后沪深 300 指数 5%以上

免责声明

本公司具有中国证监会核准的证券投资咨询业务资格，作者具有中国证券业协会注册分析师执业资格或相当的专业胜任能力。

本报告仅供财信证券股份有限公司客户及员工使用。本公司不会因接收人收到本报告而视其为本公司当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发送，概不构成任何广告。

本报告信息来源于公开资料，本公司对该信息的准确性、完整性或可靠性不作任何保证。本公司对已发报告无更新义务，若报告中所含信息发生变化，本公司可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告中所指投资及服务可能不适合个别客户，不构成客户私人咨询建议。任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司及本公司员工或者关联机构不承诺投资者一定获利，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。投资者务必注意，其据此作出的任何投资决策与本公司及本公司员工或者关联机构无关。

市场有风险，投资需谨慎。投资者不应将本报告作为投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

本报告版权仅为本公司所有，未经书面许可，任何机构和个人（包括本公司客户及员工）不得以任何形式复制、发表、引用或传播。

本报告由财信证券研究发展中心对许可范围内人员统一发送，任何人不得在公众媒体或其它渠道对外公开发布。任何机构和个人（包括本公司内部客户及员工）对外散发本报告的，则该机构和个人独自为此发送行为负责，本公司保留对该机构和个人追究相应法律责任的权利。

财信证券研究发展中心

网址：stock.hnchasing.com

地址：湖南省长沙市芙蓉中路二段 80 号顺天国际财富中心 28 层

邮编：410005

电话：0731-84403360

传真：0731-84403438