

市场研究部

2025 年 7 月 21 日

通信行业周报（7 月 14 日 - 7 月 20 日）

通信行业市场回顾

2025 年 7 月 14 日至 7 月 18 日，通信（中信）板块上涨了 7.02%，沪深 300 指数上涨了 1.09%，通信板块跑赢沪深 300 指数 5.91 个百分点。期间通信业区间涨幅位列中信一级行业第 1 位，在 30 个中信一级行业中表现领先。今年以来通信行业累计上涨 17.46%，在中信一级行业中排第 4 位，表现处于中上。截至 7 月 18 日，通信行业 PE TTM 为 23.91 倍，处于 31.57% 的分位数。

中信通信行业包括上市公司共 121 家，期间 80 家公司收涨，41 家公司收跌，总体涨多跌少。涨幅前 3 名分别为新易盛、中际旭创及仕佳光子，而跌幅前 3 名则分别为恒宝股份、通鼎互联及元道通信。

周度关注：CloudMatrix 384

2025 年 4 月 10 日，华为公司推出 CloudMatrix 384 超节点，并宣布已在芜湖数据中心规模上线。5 月 27 日，华为在昇腾 AI 开发者峰会上表示，多个昇腾 384 超节点可以组成十万卡级的超节点集群，以支持更大规模的模型训练需求。华为在《Serving Large Language Models on Huawei CloudMatrix384》论文中详细描述了其体系架构、资源调度和算子优化等多个方面的基数创新。

论文中测评结果显示，CloudMatrix-Infe 在 Prefill 阶段实现了每个 NPU 6,688 tokens/s 的吞吐量，在 decode 阶段实现了每个 NPU 1,943 tokens/s 的吞吐量（在 <50 ms TPOT 下）。此外，Ascend 910C 上的 INT8 量化在 16 个不同的基准测试中保持了与官方 DeepSeek-R1 API 相当的模型准确性。并且节点间带宽退化小于 3%，节点间延迟增加不到 1 微秒，对 AI 任务的端到端性能影响可以忽略不计。

我们认为，CloudMatrix 的创新突破，为方兴未艾的 AI 大模型进一步应用提供了全新的解决方案。尽管其前期投入和运维费用更高，但未来可望进一步优化提升。不仅为自主可控带来底气，同时也提供了全新的生态链演进方向，并对于光模块、AI 服务器、云服务等相关业务有较强的推动作用。

投资建议

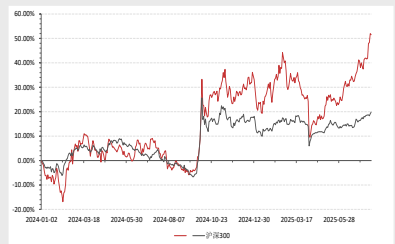
通信行业传统业务稳健发展，创新应用积极布局。经过近期市场大涨，估值水平有所回升。通信企业在当前业绩增长的基础上，积极探索未来新的增长点。近期市场波动加大，普涨后调整，此后或将走势分化。建议保持谨慎乐观，关注盈利增长持续，网络价值提升的运营商；受益于流量增长和算力网络的光通信公司；以及技术创新持续投入，核心竞争力突出的优质企业。

风险提示

1、产业发展不及预期；2、技术创新进展缓慢；3、大国博弈升级。

市场表现截至

2025.7.18



数据来源：Wind，国新证券整理

分析师：彭竑

登记编码：S1490520090001

邮箱：penghong@crsec.com.cn

目录

一、通信行业市场回顾.....	4
二、行业要闻.....	5
1. 行业动态	5
2. 企业龙头	7
3. 技术前沿	8
4. 终端	10
三、本周关注：华为 CLOUD MATRIX 384	11
四、投资建议.....	13
五、风险提示.....	13

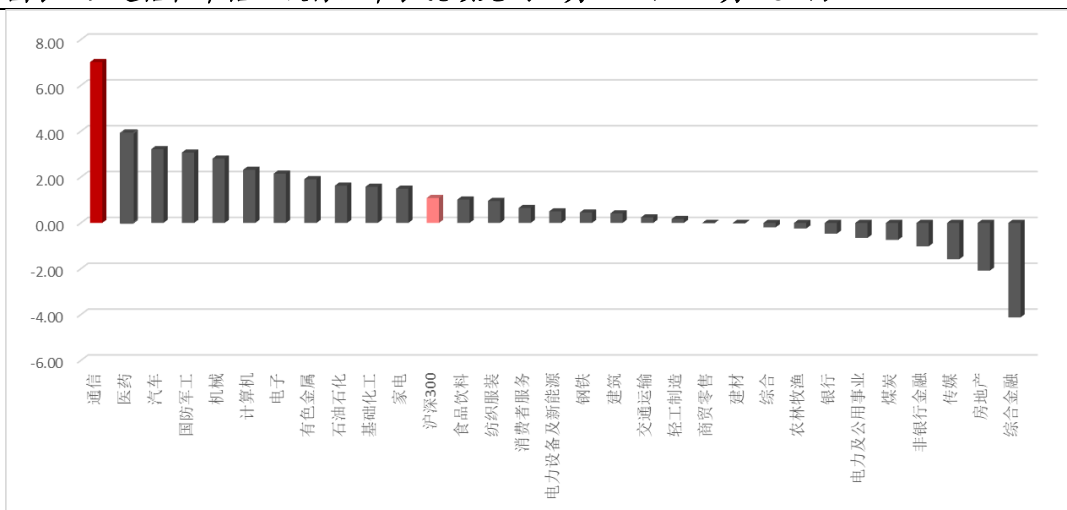
图表目录

图表 1：通信在中信一级行业中表现领先（7月14日 - 7月18日）	4
图表 2：通信行业个股区间涨跌幅前三名（7月14日 - 7月18日）	4
图表 3：通信行业估值水平略有回升	5

一、通信行业市场回顾

2025 年 7 月 14 日至 7 月 18 日，通信（中信）板块上涨了 7.02%，沪深 300 指数上涨了 1.09%，通信板块跑赢沪深 300 指数 5.91 个百分点。期间通信业区间涨幅位列中信一级行业第 1 位，在 30 个中信一级行业中表现领先。今年以来通信行业累计上涨 17.46%，在中信一级行业中排第 4 位，表现处于中上。

图表 1：通信在中信一级行业中表现领先（7 月 14 日 - 7 月 18 日）



数据来源：Wind，国新证券整理

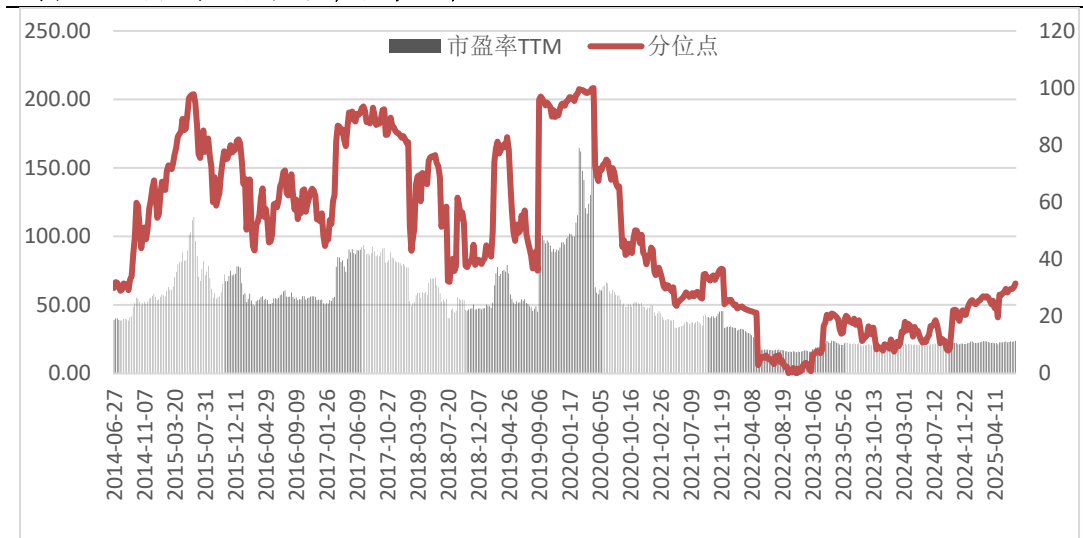
中信通信行业包括上市公司共 121 家，期间 80 家公司收涨，41 家公司收跌，总体涨多跌少。涨幅前 3 名分别为新易盛、中际旭创及仕佳光子，而跌幅前 3 名则分别为恒宝股份、通鼎互联及元道通信。

图表 2：通信行业个股区间涨跌幅前三名（7 月 14 日 - 7 月 18 日）

行业涨幅前三名			板块跌幅前三名		
证券代码	证券简称	涨跌幅	证券代码	证券简称	涨跌幅
300502.SZ	新易盛	39.01	002104.SZ	恒宝股份	-8.53
300308.SZ	中际旭创	24.33	002491.SZ	通鼎互联	-14.48
688313.SH	仕佳光子	22.43	301139.SZ	元道通信	-33.44

数据来源：Wind，国新证券整理

上周通信行业震荡攀升，行业估值水平也有所回升。2023 年以来，通信行业上涨主要为盈利驱动，受近年来系统性波动影响，估值水平提升相对不多。年报和一季报披露后分化明显，部分前期涨幅过大，缺乏业绩支撑的公司调整持续，优秀企业业绩有利于支撑业绩和股价表现。截至 7 月 18 日，通信行业 PE TTM 为 23.91 倍，处于 31.57% 的分位数。

图表 3：通信行业估值水平略有回升


数据来源：Wind，国新证券整理

二、行业要闻

1. 行业动态

我国“灯塔工厂”数量全球第一：AI 独角兽企业 71 家占全球近 1/3

商务部部长王文涛在《求是》刊文称：我国坚持科技自立自强、人才引领驱动，为高水平对外开放提供强大动力。

他表示，我国拥有大量高素质劳动者和企业家，科学家和工程师数量全球第一，研发人员全时当量稳居世界首位，人才红利正在形成。

他还提到，我国是世界第二大研发经费投入国、第一大专利申请国，在 2024 年全球创新指数中排名第 11 位。我国“灯塔工厂”数量全球第一，占比超 40%。人工智能独角兽企业 71 家，占全球近 1/3。新能源、新材料、新一代信息技术等新兴产业达到领先水平，带动国内经济转型升级，助力全球数字化绿色化发展。

注：“灯塔工厂”项目由达沃斯世界经济论坛与管理咨询公司麦肯锡合作开展遴选，被誉为“世界上最先进的工厂”，代表当今全球制造业领域智能制造和数字化最高水平，于 2018 年开始评选。

他认为，高水平对外开放要落脚在优化资源配置、激发创新活力、提高生产效率上，为高质量发展提供更强支撑。推动货物贸易优化升级，推进市场多元化，促进贸易数字化绿色化发展，营造有利于新业态新模式发展的制度环境，推进跨境电

商综合试验区建设。创新提升服务贸易，全面实施跨境服务贸易负面清单，建设国家服务贸易创新发展示范区。

此外，创新发展数字贸易，深化数字贸易国际合作。放宽外商投资市场准入，推动扩大电信、医疗领域的开放试点，稳妥有序扩大教育、文化领域自主开放。落实好外商投资法及其实施条例，打造“投资中国”品牌。深化国家服务业扩大开放综合试点示范。完善国家级经开区开放政策体系，更好发挥稳外资主阵地作用。发挥外资企业圆桌会议制度作用，为外资企业提供优质服务。推动产业链供应链国际合作，引导产业链合理有序跨境布局，拓展境外经贸合作区功能，完善海外综合服务体系，促进投资合作。（IT之家）

英伟达宣布重启 H20 对华出口 并推出一款面向中国的全新合规 GPU

据英伟达官网消息显示，英伟达将恢复向中国销售 H20，并宣布推出一款全新、完全合规的面向中国的 GPU。

这份官方新闻稿中指出，黄仁勋向客户透露，英伟达已正式提交申请，计划恢复面向中国市场销售 H20 GPU。据其介绍，美国政府已明确表示将批准相关许可，预计不久后即可启动交付。

此外，黄仁勋还宣布推出一款全新的、完全符合出口管制要求的 GPU 产品——NVIDIA RTX PRO。该产品“非常适合用于智能制造工厂和物流系统的数字孪生 AI 应用”。

黄仁勋指出，当前世界正处于一个转折点，人工智能已经成为像能源、水和互联网一样的基础资源。他强调，NVIDIA 将致力于支持开源研究、基础模型及应用，这将实现人工智能的民主化，赋能包括拉丁美洲、欧洲、亚洲及其他地区的新兴经济体。

黄仁勋曾公开批评美国政府的出口管制政策是“失败”的做法，这些限制措施激发了中国企业的创新动力。英伟达首席科学家 Bill Dally 也曾公开表示，美国对中国实施的高端芯片与技术出口限制，非但未能遏制中国人工智能产业发展，反而让中国获得很大的发展空间。

7 月 9 日，英伟达股价刷新记录，市值突破 4 万亿美元大关，成为全球历史上首家达到这一里程碑的公司，超越苹果公司此前于 2024 年 12 月 26 日创下的 3.915 万亿美元的历史纪录。（C114）

Canalys: Q1 阿里云继续稳居中国大陆云服务市场首位 华为云排名第二

7 月 14 日消息 据 Canalys 数据显示，2025 年第一季度，中国大陆在云基础设施服务上的支出达到 116 亿美元，同比增长 16%。AI 相关需求已成为推动企业向云端迁移的主要动力。

中国云服务市场竞争格局持续演进，阿里云、华为云、腾讯云这三大主要云厂

商正围绕模型开源、智能体平台、生态合作等路径加速推进 AI 商业化进程。

2025 年第一季度，阿里云继续稳居中国大陆云服务市场首位，市场份额达 33%，同比增长 15%，其 AI 相关工作负载已连续七个季度实现三位数增长，成为其云业务的核心引擎。今年 4 月，阿里云发布具备“思考”与“非思考”模式动态切换能力的混合推理模型 Qwen3，并将该系列八个模型全面开源，进一步夯实 AI 生态基础。同时，阿里云持续推进国际化战略，宣布在新加坡设立首个全球 AI 能力中心，并在马来西亚和菲律宾扩展数据中心布局。

华为云以 18% 的市场份额位列第二，营收同比增长同样达到 18%。6 月，华为云推出“盘古 5.5”大模型套件，覆盖五大核心领域，推理效率提升八倍。同时上线企业级智能体开发平台“ModelArts Versatile”，通过场景化模板简化 AI 智能体的开发流程。为加强生态建设，华为云还发布了一项总额 1.5 亿元人民币的生态激励计划，并通过其云市场 KooGallery 实现年交易笔数突破百万。

腾讯云以 10% 的市场份额位列第三。尽管短期内受内部资源调配影响，GPU 供给紧张，但其 AI 布局步伐加快。6 月，腾讯云对其混元大模型 Turbo S 与 T1 进行升级，Turbo S 已进入全球 Chatbot Arena 评测前八，并推出“腾讯云智能体开发平台”（TCADP），支持多类高性能模型接入。在国际市场，腾讯云于日本新增一个云区域，设立该国第三个可用区，进一步强化亚太服务能力。（C114）

2. 企业龙头

AMD MI308 AI 加速卡将恢复出口中国

7 月 15 日 在英伟达宣布其 H20 芯片在中国的销售获得美国政府批准仅数小时后，AMD 也确认将恢复向中国出口其 MI308 芯片。

AMD 表示，“我们最近收到特朗普政府的通知，向中国出口 MI308 产品的许可证申请将被推进至审核流程。我们计划在许可证获批后恢复出货。”

与英伟达一样，AMD 此前一直受到拜登和特朗普两届政府提出的 AI 硬件出口限制的困扰。拜登政府曾引入全面的 AI 扩散规则，而特朗普政府虽然撤销了 AI 扩散规则，但对 H20 和 MI308 芯片实施了限制。AMD 曾确认，这些措施可能使其在库存、采购承诺及相关储备方面损失约 8 亿美元。（C114）

中国移动丁海煜：频谱是低空信息基础设施发展的基石

在日前举办的 2025 无线电技术与应用发展论坛上，中国移动通信研究院副院长丁海煜表示，低空经济的发展离不开安全、高效的低空信息基础设施，低空飞行需要网联化、数字化、智能化的低空信息基础设施，提供通信、导航、感知、监管等必要信息服务保障，而频谱则是低空信息基础设施发展的基石。

他指出，从 3G 到 5G，我国移动通信产业实现跨越式发展，科学前瞻频谱规

划，有力保障我国 TDD 创新战略成功实施。利用已授权频谱建设低空信息基础设施，可提供更高质、高效的网络保障，激发低空经济规模化发展活力。相比传统点对点通信的自建链路，4G/5G 低空网在安全性、组网效率、带宽容量和感知能力上优势显著。

在安全性方面，4G/5G 网络利用授权频谱，具备安全认证和实时跟踪能力，而点对点通信工作于非授权频段，易受干扰，缺乏安全管控能力。

在组网效率方面，4G/5G 网络可规模组网，实现广域连续覆盖，且建设运营经验成熟，而自点对点通信无法规模组网，超视距飞行受限，专网建设运维成本高。

在带宽容量方面，4G/5G 网络提供大容量、高速率（百兆级）、低时延的通信能力，支持高清视频等复杂业务，而点对点通信容量有限，速率仅约 1-5Mbps。

在感知能力方面，4G/5G 网络支持通感一体技术，资源利用率高，满足城市低慢小目标监测需求，而点对点通信不支持感知能力，需额外建设感知网。

为了更好发挥频谱优势，中国移动积极响应国家战略，构建了全球首个“四驱两翼”低空智联能力体系，通过终端、网络、平台和应用四轮驱动、AI 与安全双翼护航，全面赋能低空经济高质量发展和高水平安全。

面向低空三维立体覆盖需求，中国移动提出充分复用地面网、按需新建低空网、灵活利用卫星网策略，结合 700MHz/2.6GHz/4.9GHz 多频协同优势，构建全球规模最大、质量最优、能力最强的低空通信网。（C114）

扎克伯格放大招：Meta 确认将在 AI 领域豪掷数千亿美元

据路透社今晚报道，马克·扎克伯格周一表示，Meta Platforms 将投入数千亿美元用于计算能力，以打造“超级智能”，进一步加码他对这一前沿技术的争夺战。此前，他已通过高薪争夺顶尖人工智能工程师展开人才大战。

这项声明发布之际，科技巨头们正竞相进行高调收购，并开出数百万美元年薪的薪酬方案，只为吸引人工智能领域的顶尖人才，争夺下一代 AI 技术的主导权。扎克伯格在社交平台 Threads 上发帖称：“我们有来自业务的资金来实现这一目标。”

在 Llama 4 模型遭遇挫折、多名核心员工离职之后，Meta 近期宣布成立新部门“Meta 超级智能实验室”，以整合公司内部的人工智能资源。

该实验室将由前 Scale AI 首席执行官亚历山大·王（Alexandr Wang）与前 GitHub 首席执行官纳特·弗里德曼（Nat Friedman）共同领导。此前，Meta 向 Scale AI 注资 143 亿美元（IT 之家注：现汇率约合 1025.87 亿元人民币），并加速招募顶尖 AI 人才的步伐。（IT 之家）

3. 技术前沿

中国移动人形机器人采购项目：智元、宇树中标

7月14日消息 今年6月初，中国移动旗下中移（杭州）信息技术有限公司（即杭州研发中心）启动了2025-2027年人形双足机器人代工服务采购项目，总预算高达1.2405亿元（含税）。

项目分为2个标包：

标包1为全尺寸双足人形机器人，预算金额7800万元（含税）；

标包2为小尺寸人形机器人、算力背包及五指灵巧手，预算金额4605万元（含税）。

近期公布的结果显示，智元机器人中标标包1，宇树科技中标标包2。

值得一提的是，宇树科技作为当前最火热的人形机器人公司，已得到中国移动旗下基金的投资。（C114）

亚马逊 AWS 展示 Project Rainier 分布式集群，有望成为最强 AI 模型训练计算机

亚马逊 AWS 在发布于6月24日的一篇文章中介绍了其大型分布式集群 Project Rainier，该项目有望成为世界上最强大的 AI 模型训练计算机。

Project Rainier 分布在美国境内的多个不同数据中心中，基于 AWS 旗下 Annapurna Labs 芯片部门开发的 AI 芯片 Trainium2：每个服务器配备 16 颗 Trainium2，每 4 个服务器合为一个 UltraServer，数以万计的 UltraServer 互联构成 Project Rainier 这一 UltraCluster 超级集群。

在互联方面，蓝色电缆的 NeuronLinks 负责 Tn2 UltraServer 内部高速互联，而单一数据中心内部和跨数据中心互联则交由黄色电缆的 Elastic Fabric Adapter (EFA) 技术负责。

Anthropic 将使用 Project Rainier 构建和部署旗舰模型 Claude 的未来版本，Annapurna Labs 产品和客户工程总监 Gadi Hutt 表示，Project Rainier 可提供五倍于 Anthropic 目前最大训练集群的算力。（IT之家）

SpaceX 将向 xAI 投资 20 亿美元：加强对抗 OpenAI

据媒体报道，SpaceX 已承诺向 xAI 投资 20 亿美元，作为其 50 亿美元股权融资的一部分。在马斯克旗下 xAI 加速追赶竞争对手 OpenAI 之际，此举进一步强化了其旗下公司的协同效应。

紧随 xAI 与 X 的合并完成，这笔投资使合并后公司估值跃升至 1130 亿美元。目前，Grok 聊天机器人已应用于星链（Starlink）客服业务，并计划未来集成至特斯拉（Tesla）的 Optimus 人形机器人。

不过，Grok 近期因其部分回复内容引发争议，被发现发布了多条不当言论。Grok 官方 X 账号于周六公开致歉，将问题归因于一条意外激活的“废弃代码路径”，导致其在长达 16 小时内易受平台极端言论“污染”，并强调这与底层语言模型无关。公司表示已移除相关代码并重构系统以防止复发，同时承诺将公开新的系统提示词。

尽管如此，马斯克仍盛赞 Grok 为“全球最聪明的 AI”。与此同时，xAI 正大力投入模型训练和基础设施建设。

此前（7 月 10 日），xAI 正式发布了定位为“全球最强大人工智能模型”的新一代大语言模型 Grok 4，同步推出其多智能体协作版本 Grok 4 Heavy。

马斯克宣称 Grok 4 的智能水平已“全面超越博士智力”，尤其在学术问题上能于所有学科达到甚至超越博士级别。他更乐观预测，Grok 4 有望在今年内助力人类实现“新技术发现”的突破，并可能在明年揭示新的物理现象。

据 xAI 介绍，Grok 4 的训练计算量较 Grok 2 实现了 100 倍的显著飞跃。特别是在推理和强化学习（RL）方面，其 RL 计算量是其他任何模型的 10 倍以上。这得益于 xAI 构建的、拥有 10 万块 H100 GPU 的“世界超级计算机”。（快科技）

4. 终端

中国联通上线了 eSIM 手机业务办理界面

7 月 14 日消息 近日，有网友发现，中国联通已在官方 App 及部分线上渠道悄然上线了 eSIM 手机业务的开通办理页面。

根据联通官方说明，在 eSIM 手机商用试验阶首次申请 eSIM 手机服务的用户，需携带本人有效身份证件及拟开通设备前往联通自有营业厅办理（合作营业厅暂不支持），或预约工作人员上门办理。非首次办理用户则可通过线上渠道重新提交证件信息等方式完成操作。

eSIM（嵌入式 SIM 卡）技术作为一种新型的 SIM 卡形式，具有无需插拔实体卡、支持远程配置多个运营商 Profile、节省设备内部空间等优势，广泛应用于智能手表、平板电脑以及高端智能手机等领域。

此前，中国联通已在可穿戴设备领域率先开展 eSIM 独立号码业务试点，并在部分城市实现了 eSIM 一号双终端业务的商用部署。此次面向手机用户的 eSIM 服务试水，意味着 eSIM 技术在国内的应用正逐步从“附属设备”走向“主终端”。（C114）

vivo 位列全球智能手机厂商市场份额第四

7 月 16 日 综合市场研究机构 Canalys（现并入 Omdia）和国际数据公司（IDC）数据，vivo 位列全球智能手机厂商市场份额第四。

其中，基于 Canalys（现并入 Omdia）公布的 2025 年第二季度全球智能手机市场情况，vivo 实现逆势增长，海外市场份额提升明显，以 9% 的全球市场份额位列第四，成为全球智能手机市场份额唯一正增长的中国手机品牌。

基于 IDC 公布的 2025 年第二季度全球智能手机市场出货份额信息，vivo 海外市场份额提升明显，全球市场份额同比增幅 4.8%，以 9.2% 的全球市场份额位列第四，成为全球智能手机市场出货量同比增幅最大的中国手机品牌。（C114）

苹果 AI 框架 MLX 正新增英伟达 CUDA 支持，助开发者降低硬件成本

AppleInsider 发现，苹果专为 Apple Silicon 设计的机器学习框架 MLX 正新增 CUDA 支持。

这意味着开发者可直接在 Apple Silicon Mac 上利用 MLX 开发应用，随后导出至 CUDA 环境运行。

在此之前，由于 MLX 深度集成于苹果平台的 Metal 框架，其无法在 macOS 之外运行，故导致开发者需购买额外硬件以完成部署和测试。

IT 之家注意到，该项目由 GitHub 开发者 @zcbenz 主导，其数月前开始构建 CUDA 相关支持代码，随后将项目拆分为多个模块，逐步并入苹果 MLX 主分支。

需明确说明的是，此支持仅限从适配苹果环境的 MLX 框架导出代码，使其可在 Nvidia 显卡及服务器硬件运行。这既不意味着 Mac Pro 或外接显卡坞可连接 Nvidia 显卡本地运行机器学习应用，也无法让基于 CUDA 编写的项目直接在 Apple Silicon 上运行。

对开发者而言，此项目优势在于成本控制——Nvidia 硬件配置成本极高，甚至数倍于顶配 Mac 售价。对于小规模开发者，组织可通过 Apple Silicon 设备进行开发测试，仅在量产阶段使用 Nvidia 硬件以节省开支。

此外，MLX 项目在 CUDA 系统运行时性能远超 Mac，结合 Nvidia 硬件在机器学习领域的广泛应用，开发者有望获得更强算力支持。（IT之家）。

三、本周关注：华为 Cloud Matrix 384

2025 年 4 月 10 日，华为公司推出 CloudMatrix 384 超节点，并宣布已在芜湖数据中心规模上线。5 月 27 日，华为在昇腾 AI 开发者峰会上表示，多个昇腾 384 超节点可以组成十万卡级的超节点集群，以支持更大规模的模型训练需求，帮助更多行业更快实现智能化和高效化。超节点打破了以 CPU 为中心的冯诺依曼架构，创新提出了对等计算架构，通信带宽提升了 15 倍，单跳通信时延也从 2 微秒做到 200 纳秒，降低了 10 倍。

2020 年，OpenAI 发布了论文《Scaling Laws for Neural Language Models》。与此前业界聚焦的算法创新不同，Scaling Laws 认为 AI 大模型性能与模型参数量、训练数据量以及计算资源量三者之间存在一定的关联关系，通过扩大规模，当超越临界点后，智能会随之涌现。

GPT3.5 的爆发带动全球大模型的蓬勃发展，主流模型参数规模迈向万亿量级。随着大模型参数规模的不断扩展、MoE 架构的广泛应用以及上下文长度的持续增加，对 AI 基础设施提出了前所未有的要求。GPU 集群规模不断增长。但通信开销和内存墙问题明显制约了硬件效率的充分发挥，分布式并行训练中的集合网络通信效率成为了当前大模型训练和推理的最大挑战。

受制于美国对中国的管制政策，英伟达的 GPU 难以获取，自主可控的芯片仍有一定的差距。任正非在采访中表示，“用数学补物理、非摩尔补摩尔，利用集群计算的原理，可以达到满足我们现在的需求。”华为在《Serving Large Language Models on Huawei CloudMatrix384》的论文中提出了超节点的基数细节，并从体系架构、资源调度和算子优化等多个方面展示了技术创新理念。和英伟达计算为主的架构体系不同，华为充分发挥了通信优势，通过 Scale Out 构造了超节点模式。

架构设计上，Cloud Matrix 将系统划分为三个功能子系统：预填充（Prefill）、解码（Decode）和缓存（Caching）。每个子系统独立运行，并通过显式的 KV 缓存传输接口进行通信。这种设计使得每个子系统能够根据负载需求弹性扩展，从而最大化资源利用率和端到端性能。其核心是超高带宽、低延迟的统一总线（UB）网络，通过全面点对点（Fully Peer-to-Peer）的高带宽互联，实现 CPU、NPU、DRAM、SSD、网卡等其他资源的高性能通信，使得 CloudMatrix 能够作为一个紧密耦合的大型逻辑节点运行，计算、内存和网络资源能够动态池化、统一访问以及独立扩展，支持同时运行 AI 和数据密集型应用。

CloudMatrix 融合了三个互补的网络 Plane：一是 UB Plane，以无阻塞的 All-to-All 拓扑直接连接所有 384 个 NPUs 和 192 个 CPUs，高效实现细粒度并行策略并快速的点对点访问池化内存；二是 RDMA Plane，实现云矩阵 384 超级节点与外部 RDMA 兼容系统之间的扩展通信；三是 VPC plane，通过高速网卡（擎天）将 CloudMatrix 连接到更广泛的数据中心网络。

CloudMatrix384 超节点设计为 16 个机架其中 12 个计算机架，总共托管 48 个 Ascend 910C 节点（总共 384 个 NPUs）及 4 个通信机架。其中，Ascend 910C NP 是华为 2024 年的旗舰 AI 加速器，与每个节点内的 L1 级 UB 交换机相连。二通信机架容纳二级（L2）UB 交换机，这些交换机互连超节点内的所有节点。L2 交换机被划分为 7 个独立的 Sub-plane，与 L1 级 UB 交换机相连。

算法策略上，华为也做了一系列优化。一是通过设计开发了 FusedDispatch 和 FusedCombin，降低了当前三次 All-to-All 通信延迟，解决传统 MoE 存在的通信开销大、动态形状导致效率低、串行依赖等问题；二是通过算子融合、原生 NZ 存储

及自适应分块等技术协同优化，显著提升了 MLA 的推理性能；三是基于微批次实现了一条非对称的解码流水线。该流水线由两个交错执行的流组成，各自负责解码过程中的不同阶段，根据计算需求配置了不同的 NPU 资源，实现了两个微批次之间的高效重叠执行，还能运行时条件动态调整两个流的资源分配；四是设计并实现了稳健分层的 INT8 量化方案。

软件和生态系统方面，华为开发的 CANN 作为中间软件层，实现了高级 AI 框架（如 PyTorch 和 TensorFlow）与 Ascend NPUs 的低级硬件接口之间的高效集成。CANN 架构分为驱动层（操作系统与昇腾 NPU 之间的底层接口），运行时层（管理应用程序的生命周期，编排模型计算，并提供模型和操作符的全面设备控制、内存管理和执行管理，通过 Ascend 计算语言 ACL API 访问）和库层（提供一系列高度优化的软件组件，以加速各种 AI 工作负载），通过基础设施软件部署到云端，带来了提供自主创新的整套解决方案。

论文中测评结果显示，CloudMatrix-Infe 在 Prefill 阶段实现了每个 NPU 6,688 tokens/s 的吞吐量，在 decode 阶段实现了每个 NPU 1,943 tokens/s 的吞吐量（在 <50 ms TPOT 下）。此外，Ascend 910C 上的 INT8 量化在 16 个不同的基准测试中保持了与官方 DeepSeek-R1 API 相当的模型准确性。并且节点间带宽退化小于 3%，节点间延迟增加不到 1 微秒，对 AI 任务的端到端性能影响可以忽略不计。

我们认为，CloudMatrix 的创新突破，为方兴未艾的 AI 大模型进一步应用提供了全新的解决方案。尽管其前期投入和运维费用更高，但未来可望进一步优化提升。不仅为自主可控带来底气，同时也提供了全新的生态链演进方向，并且对于光模块、AI 服务器、云服务等相关业务有较强的推动作用。

四、投资建议

通信行业传统业务稳健发展，创新应用积极布局。经过近期市场大涨，估值水平有所回升。通信企业在当前业绩增长的基础上，积极探索未来新的增长点。近期市场波动加大，普涨后调整，此后或将走势分化。建议保持谨慎乐观，关注盈利增长持续，网络价值提升的运营商；受益于流量增长和算力网络的光通信公司；以及技术创新持续投入，核心竞争力突出的优质企业。

五、风险提示

- 1、产业发展不及预期；
- 2、技术创新进展缓慢；
- 3、大国博弈升级。

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

彭斌，在此声明，本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。

本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿等。国新证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，以下简称本公司）已在知晓范围内按照相关法律规定履行披露义务。本公司的资产管理和证券自营部门以及其他投资业务部门可能独立做出与本报告中的意见和建议不一致的投资决策。本报告仅提供给本公司客户有偿使用。

本公司不会因接收人收到本报告而视其为客户。本公司会授权相关媒体刊登研究报告，但相关媒体客户并不视为本公司客户。本报告版权归本公司所有。未获得本公司书面授权，任何人不得对本报告进行任何形式的发布、复制、传播，不得以任何形式侵害该报告版权及所有相关权利。

本报告中的信息、建议等均仅供本公司客户参考之用，不构成所述证券买卖的出价或征价。本报告并未考虑到客户的具体投资目的、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，必要时可就研究报告相关问题咨询本公司的投资顾问。本公司市场研究部及其分析师认为本报告所载资料来源可靠，但本公司对这些信息的准确性和完整性均不作任何保证，也不承担任何投资者因使用本报告而产生的任何责任。本公司及其关联方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，敬请投资者注意可能存在的利益冲突及由此造成的对本报告客观性的影响。

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