

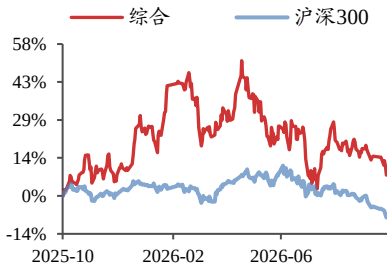
综合

2026 年 10 月 18 日

投资评级：看好（维持）

——行业周报

行业走势图



数据来源：聚源

相关研究报告

《海外科技：AMD 布局世界模型，模型能力升级与成本优化同步推进——行业周报》-2026.10.11

《海外科技：AI Agent 打开 CPU 空间，海外 X86 服务器供应链受益——行业周报》-2026.9.27

《港股消费周报：微软升级 Copilot 加速 AI 渗透，上美合作丽驻兰——行业周报》-2026.9.27

海外科技：AMD 推出 Helion 机架级平台，ASML 评估进一步扩产

初敏（分析师）

chumin@kysec.cn

证书编号：S0790522080008

叶彬慧（分析师）

yebinhui@kysec.cn

证书编号：S0790526070007

李秋莹（联系人）

liqiuying@kysec.cn

证书编号：S0790126060038

● AI 芯片：Agent 算力拓展，AI 产业链景气延续

AMD 推出 Helion 机架级平台：Agentic AI 有望推动 CPU 独立算力需求增长。Helion 提供纯 CPU 及 CPU/GPU 混合机架，面向 Agent 任务执行与模型推理。随着 Agent 并发规模扩大，CPU 有望从 GPU 配套资源升级为独立算力单元。**台积电 Q3 业绩强劲、Q4 收入指引超预期，验证 AI 算力产业链需求韧性。**Q3 营收同比增长 41.2%，Q4 指引中值超一致预期 5.9%。HPC 及 2nm 需求强劲，管理层维持 AI 长期增长信心，持续看好先进制程、GPU 及先进封装产业链。**Semtech 锁定 Tower SiGe 产能并上调业绩指引，验证 AI 高速互联需求旺盛。**Semtech 预计 FY2030 数据中心收入达 21 亿美元。我们认为，随着 800G 向 1.6T 及 3.2T 升级，高速模拟芯片有望量价齐升，Tower 的 SiGe 代工业务有望持续受益。

● AI 设备：ASML 评估进一步扩产，联想量产 Rubin 机架出货

ASML2026Q3 营收、净利均创历史新高，毛利率升至 57.4%；公司 2027-2028 年产能扩产 30%仍无法覆盖 AI 驱动的设备需求，头部客户订单已接洽至 2029-2030 年，光刻设备供给刚性持续强化。OCP 全球峰会落地下一代 AI 机架标准，英伟达开放 VeraRubinMGX1.2 规范、AMD 推出 Helion 开放机架，共同推进行业向机架级系统升级；联想率先实现 VeraRubinNVL72 系统量产出货，验证规模化交付能力。整体来看，AI 设备长周期景气确认，上游设备瓶颈凸显国产替代价值，下游架构升级利好具备整机柜系统能力的厂商。

● 投资建议：

AI 硬件：推荐标的：Tower 半导体，受益标的：Lumentum、AMD；**AI 设备：**受益标的：ASMPT、联想、Dell、HPE。

风险提示：供应链、毛利率与地缘政治风险、CPU 竞争格局变化风险、CPO 产业化不及预期风险。

目 录

1、 AI 硬件： Agent 算力拓展， AI 产业链景气延续	3
1.1、 AMD 推出 Helion， CPU 独立机架正式成为 Agent 产品方向	3
1.2、 台积电财报收入超预期， AI 先进制造没有出现系统性订单拐点	4
1.3、 Semtech 向 Tower 半导体锁定产能， 指引上调暗示行业需求旺盛	4
2、 AI 设备： ASML 评估进一步扩产， 联想量产 Rubin 机架出货	6
2.1、 ASML： 2026Q3 业绩创新高， 扩产仍难匹配 AI 需求	6
2.2、 OCP 峰会： 机架级开放标准落地加速， 联想量产 Rubin 出货	7
3、 港股周度更新	8
4、 模型与算力价格数据更新	10
4.1、 模型信息更新	10
4.2、 算力租赁价格更新	13
5、 投资建议	13
6、 风险提示	14

图表目录

图 1： 2026Q3 ASML 实现销售额 116.6 亿欧元/同比+55.2%	6
图 2： 2026Q3 ASML 归母净利润 49.6 亿欧元/同比+133.5%	7
图 3： 2026Q3 ASML 毛利率和归母净利率同比+5.8/+14.3pct	7
图 4： 联想公布 NVIDIA Vera Rubin 平台	8
图 5： 本周恒生指数（2026.10.12-2026.10.16）在全球主要市场中涨跌幅靠前（单位：%）	8
图 6： 本周（2026.10.12-2026.10.16）恒生中小型股跌幅较大（%）	9
图 7： 本周（2026.10.12-2026.10.16）恒生金融类、公共事业类涨幅居前（%）	9
图 8： 2026 年 9 月 17 日以来港股通当日买入成交净额整体为正（单位：亿元）	9
图 9： 恒生沪港通 AH 溢价指数或已触底	10
图 10： OpenRouter 的 token 调用量高位运行	12
图 11： 国产 AI 在 ZenMux 的 Token 市场份额占比超过 50%	12
图： B200 租赁价格创阶段性新高（单位：\$/小时）	13
图： H200 租赁价格高位运行（单位：\$/小时）	13
图： H100 租赁价格高位震荡（单位：\$/小时）	13
图： A100 租赁价格有所下降（单位：\$/小时）	13
表 1： AMD 新品-CPU 独立机架、CPU/GPU 融合机架相关配置	3
表 2： 2026Q3 收入超预期，营业利润率小幅超预期	4
表 3： 2026Q4 收入指引中值超预期，营业利润率略低于预期	4
表 4： Semtech 本次指引和前次指引对比	5
表 5： Semtech 对于 FY2030 年的指引	5
表 6： 本周（2026.10.12-2026.10.16）友邦保险、腾讯控股与康方生物流入资金居前	9
表 7： 9 月以来新上线模型一览	10

1、AI 硬件：Agent 算力拓展，AI 产业链景气延续

1.1、AMD 推出 Helion，CPU 独立机架正式成为 Agent 产品方向

10 月 12 日，AMD 在 OCP Global Summit 期间正式推出 AMD Helion Agentic AI Rack，标志着公司在以 GPU 为核心的 AI 训练及推理基础设施之外，进一步将产品布局延伸至 Agent 执行计算领域。Helion 基于开放计算项目（OCP）标准设计，支持客户根据不同 Agent 工作负载，灵活配置 CPU 核心密度、单核性能、GPU 加速能力及机架功耗，并提供风冷、液冷等多种部署方案。公司计划将 Helion 相关技术规格提交 OCP，预计 2026 年 11 月起由生态合作伙伴推出相关机架配置。

从具体产品配置来看，Helion 主要面向超大规模数据中心和企业级 Agent 应用两类客户。第一类为 CPU-only Rack，搭载第六代 AMD EPYC 9006 系列处理器，最高支持 32,768 个 Zen 6 CPU 核心，主要面向超大规模数据中心，为海量 Agent 任务提供独立执行算力，并与外部 GPU 推理集群协同工作。第二类为 CPU/GPU All-in-one Rack，将 EPYC CPU、Instinct MI350P GPU 及通用计算节点集成至同一机架，使 Agent 任务执行、模型推理、数据库及存储服务能够在统一基础设施内完成。AMD 表示，两类配置可以根据客户对 Agent 数量、响应延迟及算力效率的要求进行灵活调整。

表1：AMD 新品-CPU 独立机架、CPU/GPU 融合机架相关配置

配置	核心硬件	主要应用
CPU-only Rack	第六代 EPYC CPU，最高 32,768 个 CPU 核心	超大数据中心 Agent 任务
CPU/GPU All-in-one Rack	EPYC CPU + Instinct MI350P GPU	企业 Agent 综合计算

资料来源：AMD Helion 官网、开源证券研究所

我们认为，Helion 的核心产业意义，在于首次将 Agent 执行计算作为独立的机架级工作负载进行产品设计，有望推动 AI 数据中心 CPU 需求从 GPU 配套向独立算力资源拓展。传统生成式 AI 以大模型训练和推理为主要计算任务，GPU 承担矩阵运算及模型推理等核心负载，CPU 更多负责数据预处理、任务调度及 GPU 资源管理。然而，随着 AI 应用从 Chatbot 向 Agentic AI 演进，一次用户请求可能涉及任务规划、信息检索、代码执行、工具调用及结果验证等多个环节。其中，模型推理仍主要依赖 GPU，而 Agent 运行环境、代码执行沙箱、数据库访问及外部工具调用等任务则大量依赖 CPU。随着 Agent 并发数量及任务执行复杂度提高，CPU 工作负载有望持续增长，并逐渐形成独立于 GPU 推理集群的算力需求。

1.2、台积电财报收入超预期，AI 先进制造没有出现系统性订单拐点

10 月 15 日，台积电公布 2026Q3 业绩。Q3 业绩强劲，收入、利润再创新高。Q3 营收 467.2 亿美元，同比+41.2%、环比+16.2%；归母净利润 7,630.6 亿新台币，同比+68.7%、环比+8.0%。毛利率 66.4%，环比下降 1.3pct，主要受 2nm 初期量产爬坡影响；营业利润率 58.7%，高于此前指引上限 58%。公司预计 2026Q4 营收 523–543 亿美元，中值 533 亿美元，环比增长约 14.1%，较彭博一致预期 503.5 亿美元高出 5.9%。毛利率指引 65%–67%，中值 66%，符合一致预期；营业利润率指引 56.5%–58.5%，中值 57.5%，低于一致预期 58%。本次超预期主要体现为需求强劲。

表2：2026Q3 收入超预期，营业利润率小幅超预期

指标	2026Q3 实际值	2026Q3 一致预期
单季度收入	467.2 亿美元	458.34 亿美元
毛利率	66.4%	66.5%
营业利润率	58.7%	58.1%
EPS	0.93 美元	0.89 美元

资料来源：台积电 2026Q3 法说会、彭博、开源证券研究所（1 美元兑 31.8 新台币）

表3：2026Q4 收入指引中值超预期，营业利润率略低于预期

指标	2026Q4 指引	2026Q4 一致预期
收入	523 亿—543 亿美元，中值 533 亿美元	503.5 亿美元
毛利率	65%–67%，中值 66%	66.1%
营业利润率	56.5%–58.5%，中值 57.5%	58.0%

资料来源：台积电 2026Q3 法说会、彭博、开源证券研究所（1 美元兑 31.8 新台币）

公司 2026Q3 HPC 业务收入环比增长 22%，收入占比提升至 68%；2nm 晶圆收入占比由 2026Q2 的 3%提升至 8%，3nm 占比 33%。7nm 及以下先进制程合计贡献 79%的晶圆收入，AI 加速器、CPU 及先进制程需求共同支撑公司高增长。

管理层维持 AI 需求乐观判断。根据法说会最新公开报道，CEO 魏哲家表示 AI 需求仍极为强劲，尚未看到 AI 长期趋势放缓迹象。CFO 黄仁昭表述收入的增速会高于 CAPEX 的增速是公司追求的目标。公司也已经向亚利桑那以外（德州）进军，公司维持 2026 年资本开支 600–640 亿美元指引。不过，考虑零部件涨价及宏观不确定性，管理层对 2027 年消费及价格敏感型终端市场仍保持审慎。

我们认为，本次台积电业绩验证了 AI 算力产业链基本面的韧性。在近期市场担忧 AI 资本开支持续性的背景下 Q4 收入指引显著超预期，有望改善市场对 AI 硬件需求的预期。我们继续看好 AI GPU、ASIC、CPU、服务器、先进制程及先进封装产业链，同时关注 CoWoS 产能扩张及 2nm 量产进展。

1.3、Semtech 向 Tower 半导体锁定产能，指引上调暗示行业需求旺盛

10 月 13 日，Tower 半导体与 Semtech 宣布扩大双方长达 20 年的合作关系，签

署多年期产能合作协议。根据公告, Semtech 将通过投资方式, 锁定 Tower 位于美国、以色列及日本的专用 SiGe (硅锗) 晶圆制造产能, 以支持 AI 数据中心光互联及高速铜互联业务的长期发展。双方合作覆盖跨阻放大器 (TIA)、激光器驱动芯片 (Laser Driver)、调制器驱动芯片 (Modulator Driver) 及线性均衡器 (Linear Equalizer) 等高速模拟芯片, 应用场景涵盖传统可插拔光模块、线性可插拔光模块 (LPO)、近封装光学 (NPO)、有源铜缆 (ACC) 等互联架构。与此同时, Tower 披露 Semtech 在 ECOC 2026 获奖的 224Gbps/lane TIA 系列产品采用了 Tower 的 SiGe 制造工艺, 表明双方合作已经延伸至 200G/lane 及以上的新一代高速互联产品。

10 月 15 日, Semtech 举办数据中心分析师日, 上调 FY2027Q3 业绩指引, 并公布 FY2030 长期财务目标。其中, FY2027Q3 收入指引中值由 4.10 亿美元上调至 4.175 亿美元, Non-GAAP 调整后 EPS 指引由 1.05 美元上调至 1.14 美元。长期来看, 公司提出 FY2030 总收入达到 32 亿美元、数据中心收入达到 21 亿美元的目标, 对应 FY2026—FY2030 总收入复合增速约 38%、数据中心收入复合增速约 75%。以 FY2030 目标测算, 数据中心收入预计占总收入约 65.6%, 有望成为 Semtech 最核心的增长来源。同时, 公司提出 FY2030 调整后毛利率达到 70%、调整后营业利润率达到 48%—50% 的目标, 体现管理层对高附加值高速模拟芯片增长及产品结构优化的积极预期。

表4: Semtech 本次指引和前次指引对比

财务指标	最新披露	此前披露
FY27Q3 收入指引	4.175 亿美元	4.10 亿美元
FY27Q3 调整后 EPS	1.14 美元	1.05 美元

资料来源: Semtech 8-K 文件、开源证券研究所

表5: Semtech 对于 FY2030 年的指引

财务指标	最新目标
FY2030 总收入目标	32 亿美元
FY2030 数据中心收入目标	21 亿美元
数据中心 FY26—FY30 收入 CAGR 目标	75%
FY2030 调整后毛利率目标	70%
FY2030 调整后营业利润率目标	48%至 50%

资料来源: Semtech 8-K 文件、开源证券研究所

我们认为, 随着 800G 向 1.6T 及 3.2T 升级, 高速模拟芯片有望迎来性能要求提升与单模块价值量增长的双重驱动。对于 Tower 半导体而言, 此次合作的意义不仅在于增加晶圆订单, 更在于强化 SiGe 工艺平台在高速光互联电芯片供应链中的战略地位。随着单通道速率持续提升, 高速模拟电芯片对晶圆制造工艺及设计工艺协同优化的依赖有望增强, 具备成熟 SiGe 工艺积累的特色晶圆厂有望形成更高的客户黏性与技术壁垒。

推荐标的: Tower 半导体。

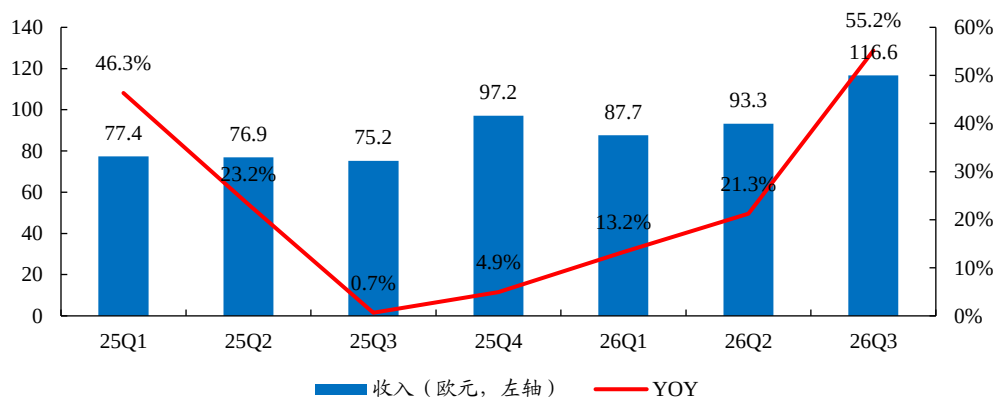
2、AI 设备：ASML 评估进一步扩产，联想量产 Rubin 机架出货

ASML：2026Q3 业绩创新高，扩产仍难匹配 AI 需求

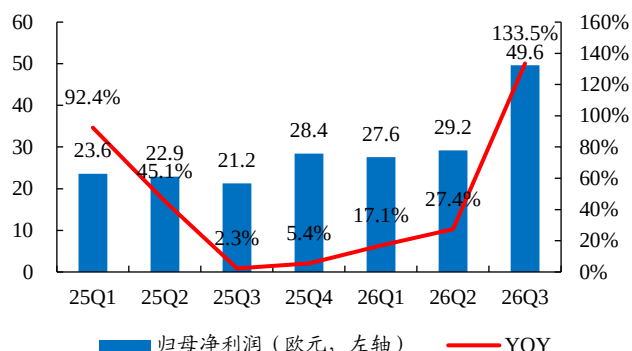
ASML 2026Q3 业绩创历史新高。2026Q3 公司实现销售额 116.6 亿欧元/同比增长 55.2%，环比增长 25.0%，略高于指引区间（110-120 亿欧元）中枢，创历史单季新高；其中系统设备净销售额 84 亿欧元，EUV 系统销售额 41 亿欧元（含 1 台 HighNA EUV 交付）。2026Q3 实现净利润 49.6 亿欧元，同比大幅增长 133.5%，环比增长 70%；毛利率达到 57.4%/同比+5.8pct，环比提升 3.5pct，高于指引上沿，一是高附加值的 EUV 设备及装机升级业务占比提升，二是产能爬坡带来的规模效应持续释放，主要来高附加值的 EUV 设备及装机升级业务占比提升和产能爬坡带来的规模效应持续释放。**展望全年，公司维持约 56% 的毛利率指引，Q4 单季预计约 57%，盈利能力仍处于上行通道。**

供给刚性持续强化，扩产仍难匹配 AI 需求。管理层在电话会中明确表示，客户基于对 AI 产业投资规模与持续周期的判断，持续上调设备采购预期，需求已明确延伸至 2028 年以后，部分头部客户已正式启动 2029-2030 年的设备订单接洽，订单能见度进一步拉长。公司此前已公布 2027-2028 年 Low-NA EUV 与浸没式 DUV 产能各提升 30% 的扩产计划，但目前头部晶圆厂远期产能锁定计划使得公司规划产能仍无法充分满足客户需求，目前正在评估进一步扩大 2028 年产能的方案。

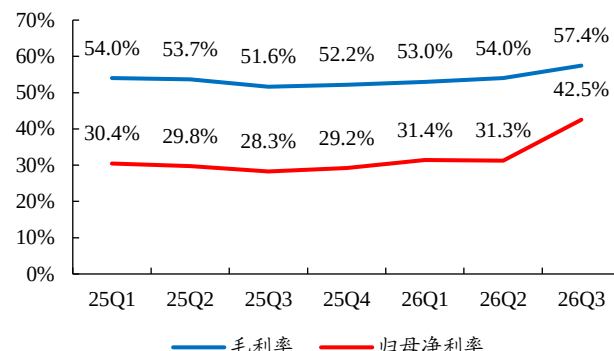
图1：2026Q3 ASML 实现销售额 116.6 亿欧元/同比+55.2%



数据来源：Bloomberg、开源证券研究所

图2：2026Q3 ASML 归母净利润 49.6 亿欧元/同比 +133.5%


数据来源：Bloomberg、开源证券研究所

图3：2026Q3 ASML 毛利率和归母净利率同比 +5.8/+14.3pct


数据来源：Bloomberg、开源证券研究所

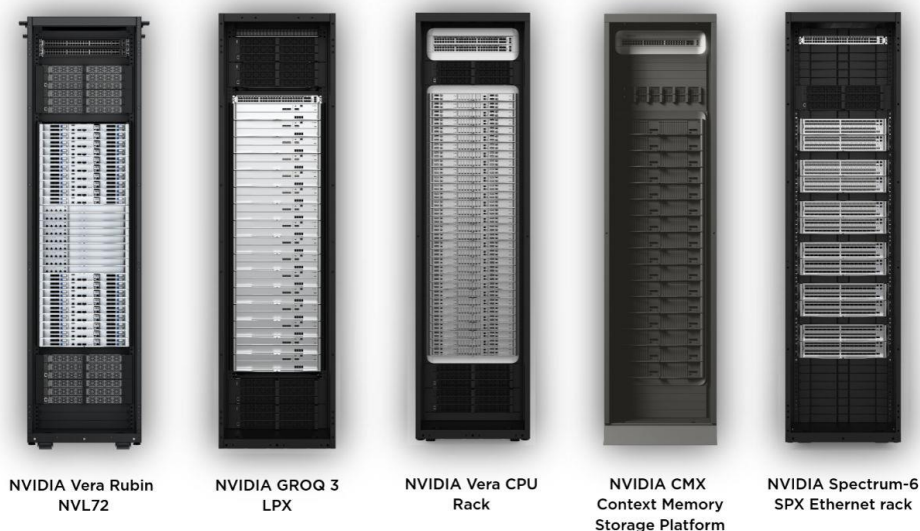
2.2、OCP 峰会：机架级开放标准落地加速，联想量产 Rubin 出货

新一代 AI 算力基础设施已完成标准定义，进入规模化落地阶段。（1）英伟达本次向 OCP 开放 VeraRubin 平台核心的 MGX1.2 机架规范，明确 250kW 级单柜功率上限，定义盲插液冷接口、5000A 液冷母排等核心标准，实现计算、互连、供电、冷却的机柜级一体化；完整 Vera Rubin 体系覆盖 NVL72 训练机架、推理加速系统、CPU 机架、存储平台及网络系统五大品类，其中 NVL72 集成 72 颗 RubinGPU 与 36 颗 VeraCPU，依托 NVLink6 高速互连，每兆瓦吞吐量最高提升 30 倍，每百万 Token 推理成本降至前代的 1/45，通过开放标准撬动全球生态规模化部署。

（2）AMD 同步推出 HelionAgenticAI 机架参考设计，原生兼容 OCP 规范，支持 CPU-only、CPU+GPU 混合灵活配置，纯 CPU 机架最高可集成 32768 个 Zen6 核心，专为智能体分布式工作负载优化；公司同时宣布将 Helion 规格贡献至 OCP 社区，以开放路线对标英伟达 MGX 生态，差异化聚焦 CPU 高密度与混合负载适配能力，预计于 2026 年 11 月推出。

终端厂商层面，联想在本次峰会上宣布已启动 VeraRubinNVL72 系统全球交付，为首批实现该平台量产出货的头部厂商，并将完整 VeraRubin 架构纳入其 AI 云超级工厂解决方案，覆盖计算、推理、存储、网络全栈交付；依托全球超 7 万台机架的年产能（含 1.1 万台液冷机架），联想的快速落地验证了开放架构下头部厂商的系统集成与量产交付能力，意味着下一代 250kW 级液冷 AI 服务器已从技术验证正式切入规模化部署周期，整机柜交付、液冷集成、全栈方案能力将成为 AI 基础设施厂商的核心竞争壁垒。

图4: 联想公布 NVIDIA Vera Rubin 平台

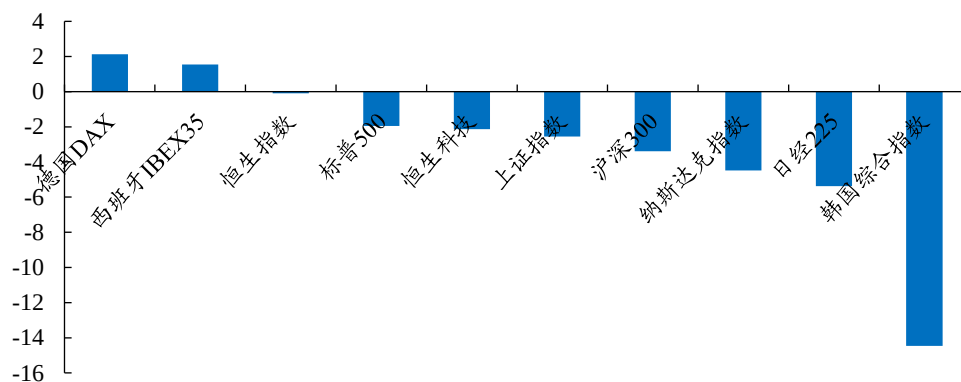


资料来源: 联想官网

3、港股周度更新

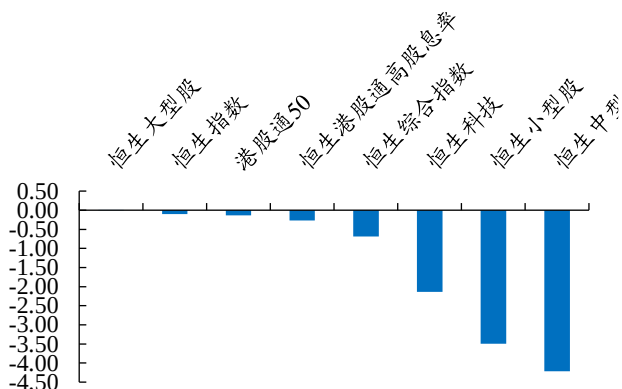
本周恒生指数下跌 0.10%，在全球主要市场中涨跌幅靠前，从成分上看，恒生中小型股等跌幅居前，恒生科技本周下跌 2.13%。

图5: 本周恒生指数（2026.10.12-2026.10.16）在全球主要市场中涨跌幅靠前（单位：%）



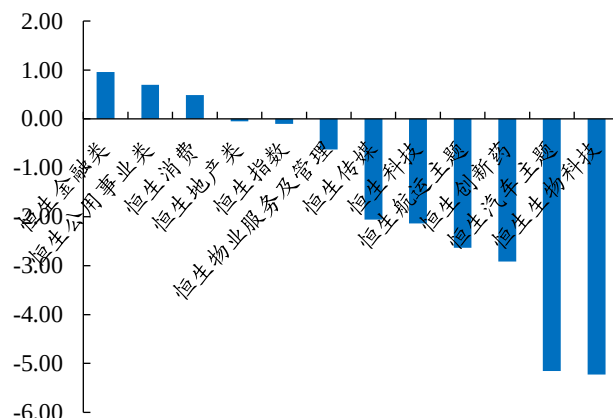
数据来源: Wind、开源证券研究所

图6: 本周（2026.10.12-2026.10.16）恒生中小型股跌幅较大（%）



数据来源: Wind、开源证券研究所

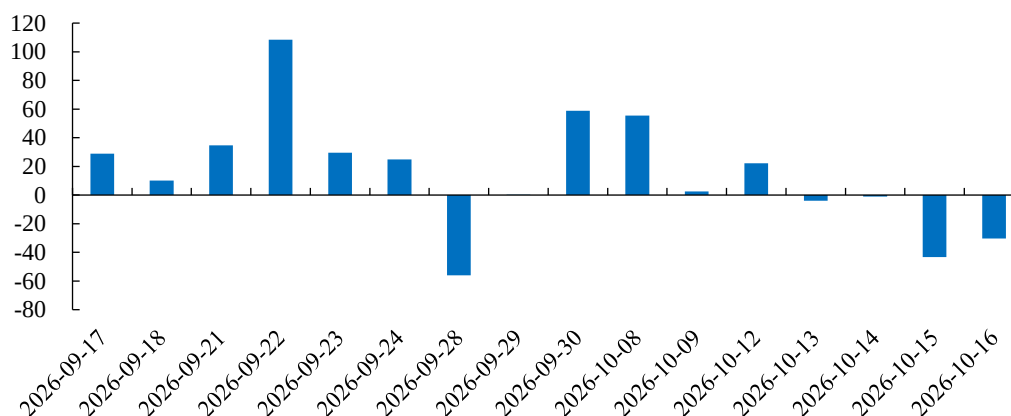
图7: 本周（2026.10.12-2026.10.16）恒生金融类、公共事业类涨幅居前（%）



数据来源: Wind、开源证券研究所

本周港股通成交净额为流出 56.52 亿元，环比下降，从头部活跃个股交易层面来看，友邦保险（+14.9 亿港元）、腾讯控股（+14.3 亿港元）、康方生物（+7.7 亿港元）资金净流入居前。

图8: 2026 年 9 月 17 日以来港股通当日买入成交净额整体为正（单位: 亿元）



数据来源: Wind、开源证券研究所

表6: 本周（2026.10.12-2026.10.16）友邦保险、腾讯控股与康方生物流入资金居前

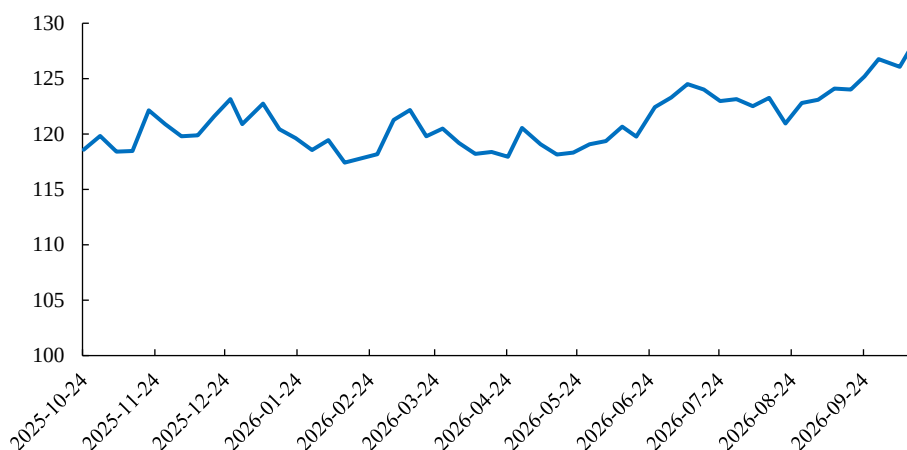
股票代码	股票名称	成交净买入（亿港元）	港股通持股数占总股本比例
1299.HK	友邦保险	14.9	4.7%
0700.HK	腾讯控股	14.3	12.0%
9926.HK	康方生物	7.7	46.8%
3690.HK	美团-W	2.8	18.7%
6166.HK	剑桥科技	0.7	59.4%
1378.HK	中国宏桥	0.7	11.7%
0100.HK	MINIMAX-W	-0.8	11.0%
2315.HK	百奥赛图-B	-0.8	41.5%

股票代码	股票名称	成交净买入（亿港元）	港股通持股数占总股本比例
0883.HK	中国海洋石油	-1.1	22.4%
0148.HK	建滔集团	-2.4	22.6%
2513.HK	智谱	-4.0	14.3%
0992.HK	联想集团	-4.2	7.6%
1548.HK	金斯瑞生物科技	-4.4	20.9%
1810.HK	小米集团-W	-7.6	19.7%
2269.HK	药明生物	-8.4	24.5%
1888.HK	建滔积层板	-10.7	11.9%
1347.HK	华虹宏力	-11.0	15.2%
6869.HK	长飞光纤光缆	-24.5	45.5%
0981.HK	中芯国际	-30.2	27.5%

数据来源：Wind、开源证券研究所

本周恒生沪港通 AH 溢价指数为 128.56，环比有所上升，AH 股溢价触底回弹，随着后续更多 A 股科技类标的赴港上市，或带动港股板块的“成长性溢价”持续向 A 股靠拢。

图9：恒生沪港通 AH 溢价指数或已触底



数据来源：Wind、开源证券研究所

4、模型与算力价格数据更新

4.1、模型信息更新

9 月以来全球大模型行业进入新品密集发布期，海内外头部厂商先后推出十余款新模型，产品定位与定价梯队分化明显。本周（2026.10.12-2026.10.16）新上线模型包括闭源模型 Muse Wonder Full、Muse Wonder 与开源模型 Muse Spark Open、DeepSeek-V4.1-Flash-1017。

表7：9 月以来新上线模型一览

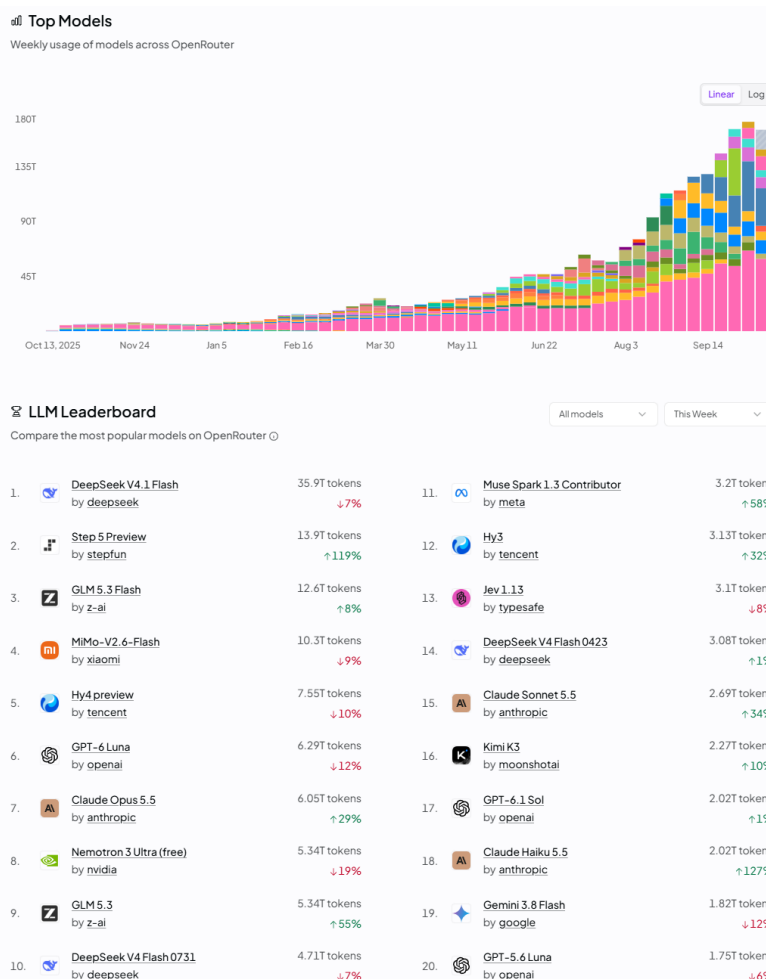
已发布模型	发布时间	厂商	输入价格 (/M)	缓存命中输入价格 (/M)	输出价格 (/M)	上下文窗口
-------	------	----	--------------	------------------	--------------	-------

已发布模型	发布时间	厂商	输入价格 (/M)	缓存命中输入价格 (/M)	输出价格 (/M)	上下文窗口
Claude Fable 5.1	2026-09-01	Anthropic	\$10.00	\$0.25	\$50.00	1M
Claude Mythos 5.1	2026-09-01	Anthropic	\$10.00	\$0.25	\$50.00	1M
Gemini 3.8 Flash	2026-09-02	Google	\$0.75	\$0.08	\$3.75	1M
Muse Spark 1.3	2026-09-02	Meta	\$1.25	\$0.15	\$4.25	1M
GPT-6 Astra	2026-09-03	OpenAI	\$10.00	\$1.00	\$50.00	1.05M
Kimi K2.8 Preview	2026-09-11	月之暗面	目前只在 Kimi Code 与 Kimi Work 上线			1M
GLM-5.3-FlashX	2026-09-18	智谱	¥2.00	¥0.57	¥7.00	1M
Step 5 Preview	2026-09-20	阶跃星辰	¥7.00	¥0.35	¥20.00	1M
MiMo-V2.6-Pro	2026/9/22	小米	¥3.00	¥0.03	¥6.00	1M
MiMo-V2.6-Flash	2026/9/22	小米	¥1.00	¥0.02	¥2.00	1M
Grok 4.7	2026/9/22	SpaceXAI	\$2.00	\$0.50	\$6.00	500K
GPT-6 Sol	2026/9/23	OpenAI	\$2.00	\$0.20	\$10.00	1.05M
GPT-6 Luna	2026/9/23	OpenAI	\$0.10	\$0.01	\$0.50	1.05M
Claude Opus 5.5	2026/9/23	Anthropic	\$4.00	\$0.20	\$20.00	1M
MiniMax M3.1 Flash Preview	2026/9/27	MiniMax	目前只在 MiniMax Code 上线			1M
Claude Sonnet 5.5	2026/9/28	Anthropic	\$2.00	\$0.20	\$10.00	1M
GPT-6.1 Sol	2026/9/29	OpenAI	\$2.00	\$0.10	\$10.00	1.05M
Gemini 4 Argon	2026/10/01	Google	\$2.00	\$0.10	\$10.00	1M
Beam	2026/10/05	Reflection AI	暂无			1M
Claude Haiku 5.5	2026/10/8	Anthropic	\$0.10	\$0.01	\$0.50	1M
Muse Wonder Full	2026-10-12	Meta	目前只在 Muse Code 订阅中开放			1M
Muse Wonder	2026-10-12	Meta	\$4.00	\$0.20	\$20.00	1M
Muse Spark Open	2026-10-15	Meta	\$1.25	\$0.15	\$4.25	1M
DeepSeek-V4.1-Flash-1017(高峰)	2026-10-17	DeepSeek	\$0.30	\$0.006	\$1.20	1M
DeepSeek-V4.1-Flash-1017(空闲)	2026-10-17	DeepSeek	\$0.15	\$0.003	\$0.60	1M

资料来源：各模型厂商官网、IT之家、Openrouter、开源证券研究所

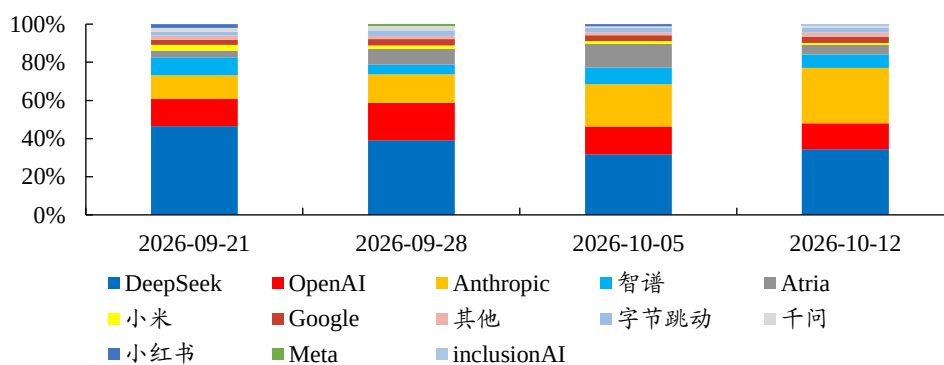
本周（2026.10.12-2026.10.16）OpenRouter 的 token 调用量持续提升，其中国产模型 DeepSeek V4.1 Flash、Step 5 Preview、GLM 5.3 Flash 与 MiMo-V2.6-Flash 位居前列。截至 10 月 12 日，在 ZenMux 统计的 Token 市场份额中，国产 AI 占比超过 50%，DeepSeek 市场份额保持第一，占比为 34.5%。依托开源生态及出色的性价比优势，国产模型持续扩大市场影响力。

图10: OpenRouter 的 token 调用量高位运行



资料来源: OpenRouter 官网

图11: 国产 AI 在 ZenMux 的 Token 市场份额占比超过 50%



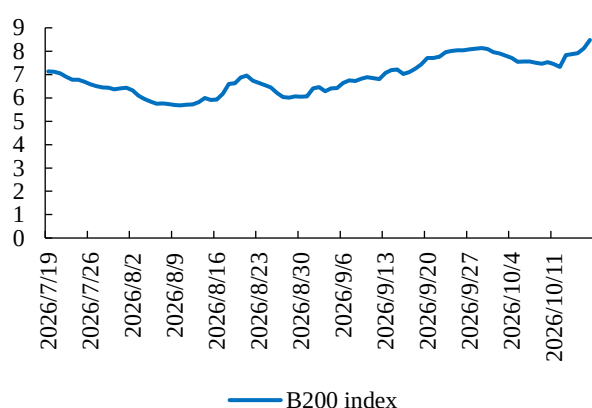
数据来源: ZenMux、开源证券研究所

4.2、算力租赁价格更新

GPU 租赁价格整体高位运行，不同性能 GPU 价格走势有所分化。根据 Ornn 统计的基于 neocloud 市场的租赁交易价格，GPU 租赁价格整体保持高位运行，不过不同性能 GPU 价格走势有所分化，高性能型号 B200 与 H200 租赁价格走势坚挺，创下阶段性新高；而 H100 与 A100 租赁价格走势则偏弱。

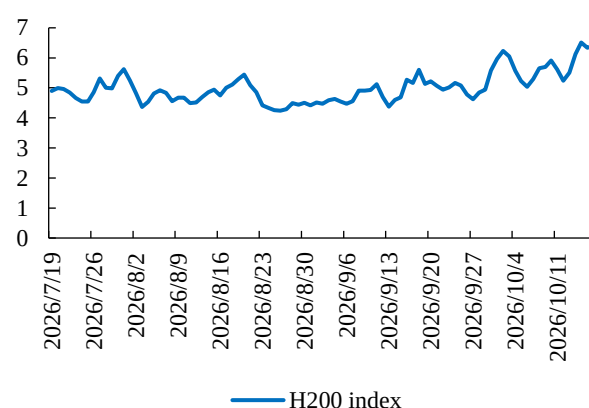
截至 10 月 17 日，B200 租赁价格为 8.48 \$/小时；H200 租赁价格为 6.39 \$/小时；H100 租赁价格为 2.78 \$/小时；A100 租赁价格为 0.97 \$/小时。

图：B200 租赁价格创阶段性新高（单位：\$/小时）



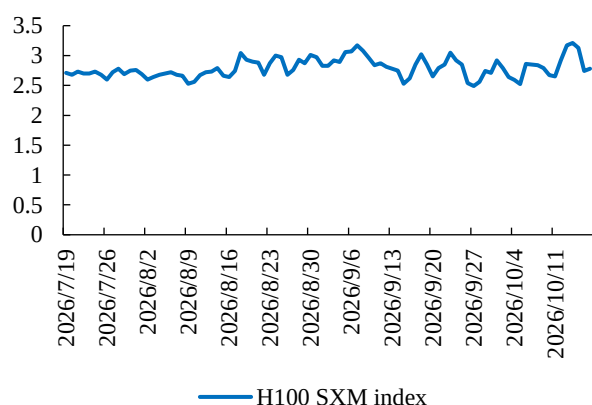
数据来源：Ornn、开源证券研究所

图：H200 租赁价格高位运行（单位：\$/小时）



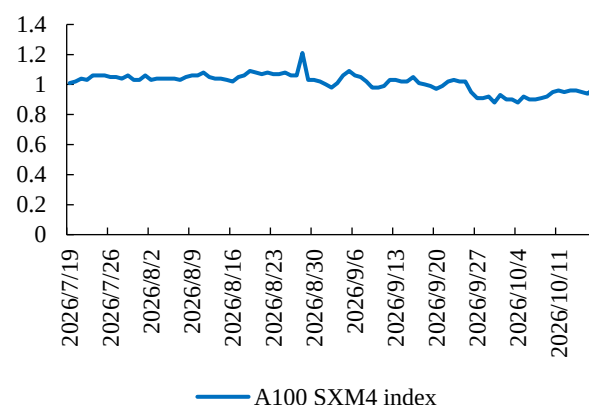
数据来源：Ornn、开源证券研究所

图：H100 租赁价格高位震荡（单位：\$/小时）



数据来源：Ornn、开源证券研究所

图：A100 租赁价格有所下降（单位：\$/小时）



数据来源：Ornn、开源证券研究所

5、投资建议

AI 硬件：我们认为本周 AMD 推出的独立 CPU 机架印证 AI 未来的发展方向会更偏向 Agent，而台积电的收入超预期隐含了 AI 的持续发展，而 Semtech 的收入上调说明了“互联”环节需求的强劲，我们继续推荐在“互联”方向的硅光/硅锗“卖铲者”Tower 半导体，其他受益标的包括 AMD、台积电。

AI 硬件：半导体景气向上游传导链条完整流畅，先进存储和逻辑扩产资本开支上修将加速半导体设备需求向上、弹性更大的设备厂商，受益标的如 ASMP、Besi、ASML 等。服务器 OEM 行业进入 AI 算力+传统 IT 换机双轮驱动的上行周期，需求从单一超大规模云厂扩散至新云/企业/主权算力，产品形态从单机硬件向整机柜液冷一体化方案升级；行业盈利模型重构，具备客户结构、全栈方案、供应链调度、高密度交付能力的头部 OEM 厂商业绩确定性更强，受益标的：联想、Dell、HPE。

6、风险提示

供应链、毛利率与地缘政治风险：内存、先进封装、电力和机房资源仍可能制约交付，原材料涨价和完整系统产品占比提升可能压低毛利率；对中国市场的出口许可、供应链区域化及关键原材料限制亦可能影响收入、成本和交付节奏。

CPU 竞争格局变化风险：Vera 放量既验证 CPU 需求，也可能加剧 Arm 与 x86 竞争；AMD 等传统 CPU 厂商的实际受益程度取决于产品竞争力和客户份额。

CPO 产业化不及预期风险：CPO 仍涉及光源可靠性、硅光良率、散热、封装、客户认证和可维护性，Spectrum-6 支持 CPO 不等同于 CPO 版本已形成大规模收入。

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已于2017年7月1日起正式实施。根据上述规定，开源证券评定此研报的风险等级为R4（中高风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为境内专业投资者及风险承受能力为C4、C5的普通投资者。若您并非境内专业投资者及风险承受能力为C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师承诺

本研究报告的署名人员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，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（outperform）	预计相对强于市场表现 5%～20%；
	中性（Neutral）	预计相对市场表现在 - 5%～ + 5%之间波动；
	减持（underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（underperform）	预计行业弱于整体市场表现。

备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数、港股基准指数为恒生指数、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）、美股基准指数为标普 500 或纳斯达克综合指数。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取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

分析、估值方法的局限性说明

本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（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。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。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。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。本报告中所指的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。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户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。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。在任何情况下，本公司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。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，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。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

开源证券研究所

上海

地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

深圳

地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

北京

地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

西安

地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn