

综合

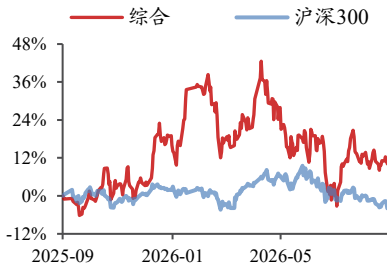
2026 年 09 月 27 日

## 海外科技：AI Agent 打开 CPU 空间，海外 X86 服务器供应链受益

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

《市场更加重视模型帕累托最优，模型厂&云厂&主权基金资金弹药充足——港股行业深度报告》-2026.9.24

《Oracle 云基建收入同比翻倍，Nebius 再度提价，算力厂商 CDS 走阔——行业周报》-2026.9.20

《中资模型在图文生成赛道集群式领先，子不语在 TikTok 渠道实现高增——行业周报》-2026.9.20

初敏（分析师）

chumin@kysec.cn

证书编号：S0790522080008

叶彬慧（分析师）

yangzhe@kysec.cn

证书编号：S0790524100001

#### ● AI 芯片：AI Agent 打开 CPU 空间，3.2T 重塑光模块电芯片价值

**Muse 验证消费级常驻 Agent，CPU 打开增量空间：**Meta Muse 上线后迅速登顶美国 App Store，验证消费级常驻 Agent 需求。其独立云端 VM 需持续调用 CPU 运行浏览器、代码、工具及多 Agent 任务。我们认为，若消费级常驻 Agent 由单 Agent 向多 Agent 扩展，AIDC CPU 需求有望超预期，AMD EPYC 等服务器 CPU 生态受益。3.2T 驱动光模块光电协同升级：Marvell 在 ECOC 展示 2nm 400G/lane PAM4 方案，推动光互联由 1.6T 向 3.2T 演进。我们认为，升级逻辑正由光器件扩展至 DSP、TIA、Driver 等 EIC 全链条；先进制程与更高模拟带宽有望抬升电芯片价值量，并进一步利好 Tower SiGe 业务。

#### ● 服务器：Agent 时代分布式推理迎产业验证，海外 X86 服务器供应链受益

9 月 24 日边缘算力服务商 Akamai 与 Anthropic 达成扩展合作，签署 7 年期 116 亿美元基础合同，协议最高可扩展至 200 亿美元，重点承接 Anthropic 的 CPU 类工作负载，覆盖 Agent 编排、网关调度、数据预处理与轻量推理等业务。为履行订单，Akamai 规划三年合计 55 亿美元资本开支（2026 年 17 亿、2027 年 31 亿、2028 年 7 亿），投向 X86 服务器整机、内存及网络设备，并与联想、Jabil 签署硬件总协议锁定供应链产能。Akamai 采用 GPU 集群+全球分布式边缘节点两级架构，CPU 算力池以 AMDEPYC 为主体，依托跨 130 个国家的 4400 个边缘节点实现低时延推理，通过语义缓存、智能调度优化 GPU 资源消耗，验证 Agent 应用带动 CPU 算力的新增需求。本次大额长约打开北美 CSP 推理硬件采购空间，联想作为核心 X86 服务器供应商直接受益，海外 AI 基础设施业务增长弹性持续凸显。

#### ● 投资建议：

**AI 芯片：**Muse 等常驻型 Agent 有望提升服务器 CPU 配置规模，X86 CPU 且有独立 72 卡 AI 服务器设计能力的公司有望提升份额，受益标的：AMD；随 100G/通道逐步向 400G/通道发展，电芯片代工厂的价值有望提升，受益标的：Tower 半导体。**整机：AI 大模型推理和 Agent 渗透推动算力需扩散，推动 X86 AI 和通用服务器份额企稳回升，企业端布局较深且具备保供能力的 OEM 品牌有望受益，受益标的：联想集团、DELL。AI 设备：**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，受益标的：ASMPT。

● **风险提示：**产能及供应链风险，监管政策变动，宏观经济增长放缓，地缘政治风险。

## 目 录

|                                                           |    |
|-----------------------------------------------------------|----|
| 1、 AI 硬件：Agent 驱动 CPU 增量，3.2T 推动电芯片升级 .....               | 3  |
| 1.1、 Meta Muse 走红，常驻 Agent 打开 CPU 增量需求想象空间 .....          | 3  |
| 1.2、 400G/通道迈向 2 纳米，3.2T 驱动光模块电芯片价值升级 .....               | 4  |
| 2、 服务器：Akamai 与 Anthropic 达成百亿美元通用服务器订单，联想成为核心硬件供应商 ..... | 5  |
| 3、 港股周度更新 .....                                           | 6  |
| 4、 模型信息更新 .....                                           | 8  |
| 5、 投资建议 .....                                             | 11 |
| 6、 风险提示 .....                                             | 11 |

## 图表目录

|                                                              |    |
|--------------------------------------------------------------|----|
| 图 1： Akamai 全球边缘网络基础设施可降低延迟成本 .....                          | 6  |
| 图 2： 本周恒生指数（2026.09.21-2026.09.25）在全球主要市场中涨跌幅居后位（单位：%） ..... | 6  |
| 图 3： 本周（2026.09.21-2026.09.25）恒生科技跌幅较大（%） .....              | 7  |
| 图 4： 本周（2026.09.21-2026.09.25）恒生汽车主题、恒生科技跌幅较大（%） .....       | 7  |
| 图 5： 2026 年 8 月以来港股通当日买入成交净额整体为正（亿元） .....                   | 7  |
| 图 6： 恒生沪港通 AH 溢价指数持续回弹 .....                                 | 8  |
| 图 7： OpenRouter 的 token 调用量持续提升 .....                        | 10 |
| 图 8： 国产 AI 在 ZenMux 的 Token 市场份额占比超过 60% .....               | 10 |
| 表 1： Akamai 预计投入 55 亿美元以保障 Anthropic 订单交付 .....              | 5  |
| 表 2： 本周（2026.09.21-2026.09.25）腾讯控股、阿里巴巴-W、智谱南向流入资金居前 .....   | 7  |
| 表 3： 9 月以来新上线模型与待发布模型 .....                                  | 8  |

## 1、AI 硬件：Agent 驱动 CPU 增量，3.2T 推动电芯片升级

### 1.1、Meta Muse 走红，常驻 Agent 打开 CPU 增量需求想象空间

Meta 于 9 月 8 日正式发布个人 AI Agent Muse，上线两天后即升至美国 App Store 免费榜第 2 位，并于 9 月 18 日登顶美国 iPhone 免费应用榜，超过 ChatGPT；Apptopia 数据显示，Muse 上线前 12 天在美国、加拿大 iOS 端累计下载约 180 万次，高于 ChatGPT 移动端上线同期的 130 万次。我们认为，Muse 受到高度关注的核心是其首次较大规模验证了常驻型 Agent 的消费级落地形态。

Meta 官方披露，Muse 运行于 Muse 安全虚拟机—每名用户拥有一台持久化、独立的云端虚拟机，配备自己的浏览器、存储、CPU 和内存；即使用户关闭 App，Muse 仍可继续在后台完成长期任务，并在需要用户批准或任务状态变化时主动返回。Meta 技术团队进一步披露，该虚拟机拥有足够的 CPU 和内存资源，可同时执行代码编译、浏览器操作、并发小 Agent 以及 Cron 定时任务，并明确将其描述为“an agent in your own cloud computer”（一个在您云端电脑的 Agent）。因此，Muse 与其他 Agent 最大的区别之一，是运行时间从“关机即结束”变成了“后台长期执行任务”。

此前 Clowdbot 等产品已经能够实现“常驻”，但其部署方式更偏向开发者和技术爱好者：用户通常需要自行准备节点等运行环境，安装命令行和 Gateway daemon，并配置模型 API/OAuth、通讯渠道及后台服务，因此相较于传统消费级 App，其自主部署和维护门槛明显更高；而 Muse 是消费级云端产品，大幅降低了用户使用门槛，Meta 官方明确指出，Muse 专为服务全球数十亿用户而设计，因此无需学习即可上手；任何人都可以直接使用，无需任何技术背景。

另，Muse 主打消费级应用场景，根据 Meta 官方对 Muse 的描述，消费级 Agent 的场景大致包含 5 类：1. 个人效率；2. 购物。3. 旅行服务；4. 个人财务管理；5. 长期目标管理。我们认为，这一举措拓宽了 Agent 使用范围的想象空间，从“商务端”到“消费端”。

凭借着 Meta 的平台如 Facebook、Instagram、WhatsApp 可以快速向海量用户分发。Sensor Tower 数据显示，Muse 上线后 Meta 迅速通过旗下产品推广 Muse；截至 9 月 25 日，Sensor Tower 估计 Muse 下载量已经超过 340 万。我们认为，这或意味着该消费级 Agent 产品能较快下沉到较多的用户端，带来较强的影响力。

我们认为，Muse 可能成为未来 AI 数据中心 CPU 需求被低估的重要增量来源。大模型核心推理仍主要由 GPU/ASIC 完成，但 Agent 运行所需的操作系统和工具调用均需要 CPU 承载；当 Agent 从一次性请求演变为持续虚拟机环境，并进一步同时运行多个定时任务后，单个活跃用户对应的 CPU 占用时长和服务器侧并发需求均可能显著高于传统聊天机器人模式。因此，未来 AI 基础设施或将随着更多消费级常驻 Agent 的出现而带动更多 CPU 的需求。

Meta 与 AMD 在 2026 年已进一步扩大合作，Meta 不仅部署了数百万颗 AMD EPYC CPU，还将成为第六代 EPYC Venice 的头部客户之一，并计划在首批 Helios 系统中采用 MI450 GPU+Venice CPU；双方正在针对 Meta 的 AI 训练、推理及新 AI 体验共同优化 CPU、GPU、网络和系统。我们认为，Muse 更重要的意义在于验证了“每个用户对应一个持续运行的云端 Agent 环境”的潜在新计算范式；若未来 OpenAI、Google、Meta 等持续推出类似的消费级常驻 Agent，并由单 Agent 进一步走向多 Agent，

AIDC 中的 CPU 有望由 AI 服务器的辅助计算单元逐步成为 Agent 基础设施的重要资源池，AMD、Intel 及 Arm 服务器 CPU 生态均有望受益，其中 AMD 凭借 EPYC 份额持续提升以及与 Meta 等头部 CSP 在 AI 基础设施上的深度合作，具备较高的产业弹性。

**受益标的：**AMD、英特尔等。

## 1.2、400G/通道迈向 2 纳米，3.2T 驱动光模块电芯片价值升级

本周 ECOC 2026 召开，Marvell 展示了行业首个基于 2 纳米工艺的 400G/通道光学 PAM4 方案。Marvell 表示，该技术面向 AI 数据中心从 1.6T 向 3.2T 及以上速率演进：当前主流 1.6T 光模块主要采用  $8 \times 200\text{G}$  架构，而随着单通道速率进一步翻倍至 400G，下一代 3.2T 模块可以延续 8 通道架构实现  $8 \times 400\text{G}$ ，在不显著增加光通道数量的情况下实现模块总带宽再次翻倍。Marvell 同时展示 2 纳米 1.6T ZR、Coherent-lite 以及 102.4T CPO 等方案，反映 3.2T 相关技术储备正在从实验室逐步进入产业验证阶段。

Marvell 光 DSP 已经形成清晰的制程演进路径：2023 年 Nova 采用 5 纳米、200G/通道，2024 年 Ara 进一步升级至 3 纳米、200G/通道，而本次 400G/通道 PAM4 进一步进入 2 纳米。Marvell 指出，先进节点主要用于提升效率、带宽密度并降低每 bit 功耗，这说明随着单通道速率由 100G 向 400G 提升，DSP 需要处理更高波特率、更复杂的均衡/FEC/SerDes，同时满足光模块严格的功耗约束，使数字电芯片必须持续向更先进制程迁移。此前 3 纳米 Ara 已经将 200G SerDes、DSP 及激光器调制器驱动进一步集成，并搭配 200G TIA（跨阻放大器），较上一代方案将 1.6T 模块功耗降低超过 20%，本质上反映光模块并非只有光器件升级，DSP、Driver、TIA 等 EIC 也需要同步迭代。

我们认为，3.2T 时代应从“光芯片价值提升”进一步扩展至“整套光模块 EIC 价值提升”的投资逻辑。一方面，DSP 从 5 纳米逐步向 2 纳米迁移，单颗芯片的设计复杂度、SerDes 性能及集成功能持续增加，我们预计将对 DSP 单颗价值量形成支撑；另一方面，400G/通道并不是简单将 200G/通道芯片超频，而要求 DSP/SerDes、TIA、Modulator Driver 等整条电信号链同步支持更高模拟带宽、更低噪声和更严格的信号完整性要求。例如 Marvell 目前的 1.6T Ara 已经不是单独销售 DSP 的简单架构，而是 DSP 内部集成 Laser Driver、外部配套专用 200G TIA，表明高速光模块越来越趋向“光电共同优化”。因此，从 400G→800G→1.6T→3.2T 的升级中，市场不应只关注 EML、硅光、CW 激光器等光器件的 ASP 提升，DSP、TIA、Driver 等 EIC 同样存在显著的规格升级和价值量提升机会；未来 400G/通道放量后，光模块价值链有望由此前偏重光器件的“光升级”，进一步演化为“光+电共同升级”。

当前高速光互联 EIC 的主要玩家包括 Broadcom、Marvell、Semtech 及 MACOM，其中 Tower 半导体已公开披露与 Broadcom、Semtech、MACOM 在高速光通信电芯片领域存在长期合作。Tower 半导体的 SiGe BiCMOS 平台主要面向 TIA、Laser/Modulator Driver、CDR 等高速模拟/混合信号芯片，随 AI 光互联由 800G 向 1.6T、3.2T 持续升级，单通道速率由 100G/lane 向 200G、400G/lane 提升，对 TIA/Driver 的带宽、线性度、噪声、功耗及均衡能力提出更高要求。Tower 半导体目前主力量产 SiGe 平台已面向 1.6T 及以上速率，管理层亦指出 AI 数据中心正在推动 SiGe TIA 和 Driver 需求快速增长。我们认为，若未来高速 EIC 的设计及制造难度随单通道速率

进一步提升，具备高性能 SiGe BiCMOS 工艺、量产经验及头部客户基础的 Tower 半导体有望进一步提升在光通信模拟 EIC 代工环节的价值量及市场份额，硅锗业务有望成为除硅光之外另一条受益于 AI 光互联升级的重要增长主线。

受益标的：Tower 半导体、格罗方德等。

## 2、服务器：Akamai 与 Anthropic 达成百亿美元通用服务器订单，联想成为核心硬件供应商

**事件：**9 月 24 日头部边缘算力服务商 Akamai 宣布与 Anthropic 进一步扩展合作关系，双方签署新一份为期七年、价值 116 亿美元的基础合同（最高扩展空间达 200 亿美元/7 年）。这项多年期合作将利用 Akamai Cloud 的分布式人工智能基础设施和软件，满足 Anthropic 日益增长的 CPU 工作负载需求。根据公司电话会，为履行协议，Akamai 拟配套规划 55 亿美元总资本开支，分三年逐步投入（2026 年约 17 亿、2027 年约 31 亿、2028 年约 7 亿），资金主力投向 X86 服务器整机、内存芯片及配套网络设备，已同步与联想、Jabil 签订硬件总协议以锁定核心服务器及内存供应链产能。

**表1：Akamai 预计投入 55 亿美元以保障 Anthropic 订单交付**

| 时间      | CapEx    | 用途           |
|---------|----------|--------------|
| 2026Q4  | 约 17 亿美元 | 预购内存等关键供应链部件 |
| 2027 全年 | 约 31 亿美元 | 服务器等主要设备部署   |
| 2028 全年 | 约 7 亿美元  | 完成全部硬件部署     |

数据来源：Akamai 公司公告、开源证券研究所

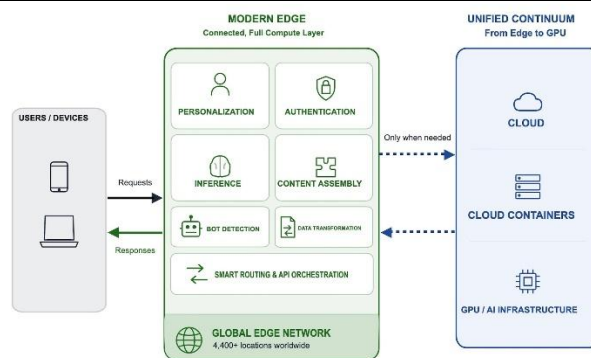
**Akamai 分布式算力网络锚定 AI 推理低成本和低时延需求，进一步验证 Agent 渗透驱动 CPU 部署趋势。**Akamai Inference Cloud 是业界首个全球规模的 NVIDIA AI Grid 部署方案，它能够智能地在其边缘、区域和核心网络中路由 AI 工作负载，从而平衡延迟、成本和性能，采用 GPU 云集群+分布式边缘节点的两级算力架构，支持跨 130 个国家、700 个城市、4400 个边缘节点的分布式推理：

**（1）GPU 侧：**已部署数千张 NVIDIA RTX PRO 6000 Blackwell 服务器级 GPU，建成全球规模领先的 Blackwell 架构 AI 集群，配套高带宽无损以太网与并行存储系统，适配大模型训练与高并发推理。

**（2）CPU 侧：**拥有超大规模 x86 CPU 算力池（AMD EPYC 为主体），也是本次与 Anthropic 合作订单的核心供给资源，可承载通用计算与 AI 推理混合负载，例如 Claude 大模型的网关、编排、agent、数据处理、安全及可能的轻量推理等。

随 AI 推理成为主流工作负载，企业侧加速垂类 AI Agent 部署，新型原生 AI 应用需要可预测的延迟和更高的成本效益，基于 CPU 的 Akamai 分布式基础设施布局可解决以上痛点：（1）利用语义缓存和智能编排，将请求定向到大小合适的资源，并为真正需要的工作负载预留优质的 GPU 计算层级，显著降低推理成本。（2）通过边缘网络为物理 AI 和自主代理提供快速响应，同时 GPU 集群以集中式规模与分布式边缘计算形成互补。

图1: Akamai 全球边缘网络基础设施可降低延迟成本



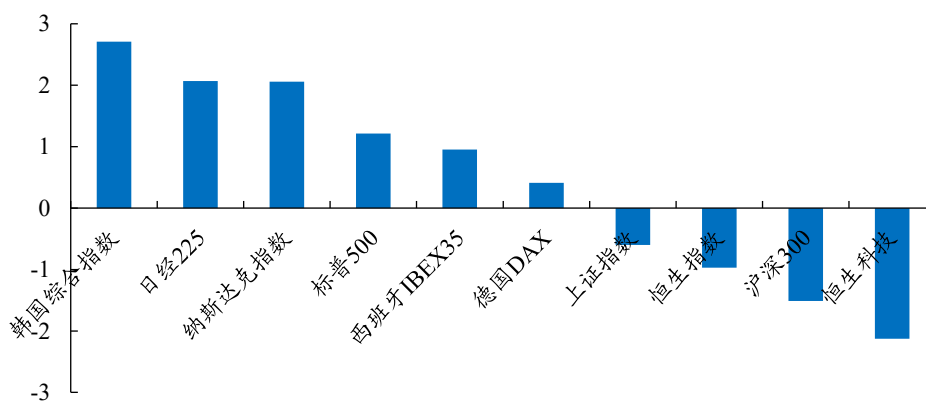
资料来源: Akamai 公司官网

联想为 X86 服务器核心供应商, 北美 CSP 推理订单有望进一步上行。联想作为全球 X86 服务器核心厂商, 同步与 Akamai 签下 7 年硬件总协议, 分担 AMDEPYC 架构 X86 服务器的交付任务, 直接分享本轮百亿算力订单的硬件采购红利; 同时依托本次合作进一步突破海外头部 AI 算力客户, 海外 AI 基础层设施业务的增长弹性持续凸显。

### 3、港股周度更新

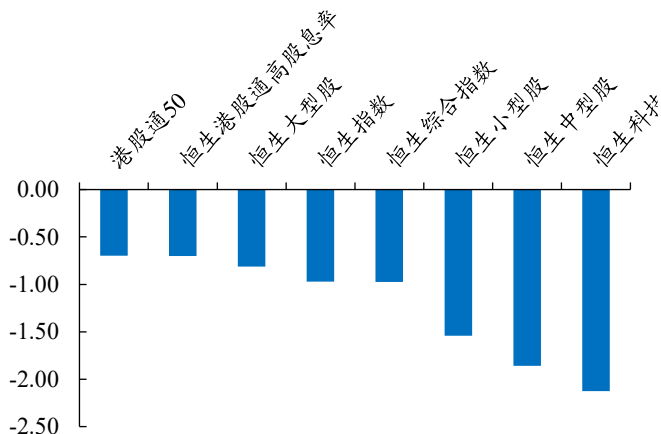
本周恒生指数下跌 0.97%, 在全球主要市场中涨跌幅靠后, 从成分上看, 恒生科技、恒生中型股等跌幅居前, 恒生科技本周下跌 2.13%。

图2: 本周恒生指数 (2026.09.21-2026.09.25) 在全球主要市场中涨跌幅居后位 (单位: %)



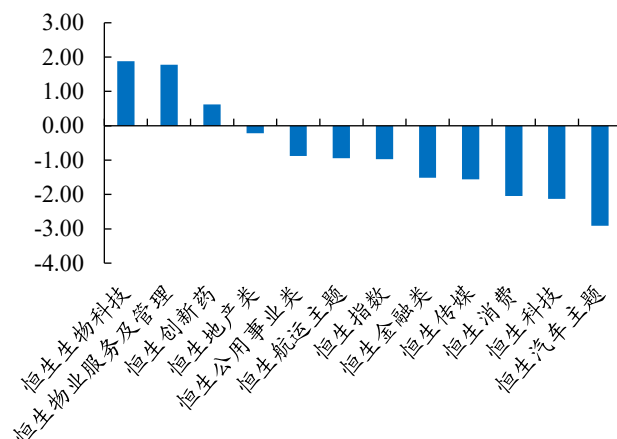
数据来源: Wind、开源证券研究所

图3: 本周 (2026.09.21-2026.09.25) 恒生科技跌幅较大 (%)



数据来源: Wind、开源证券研究所

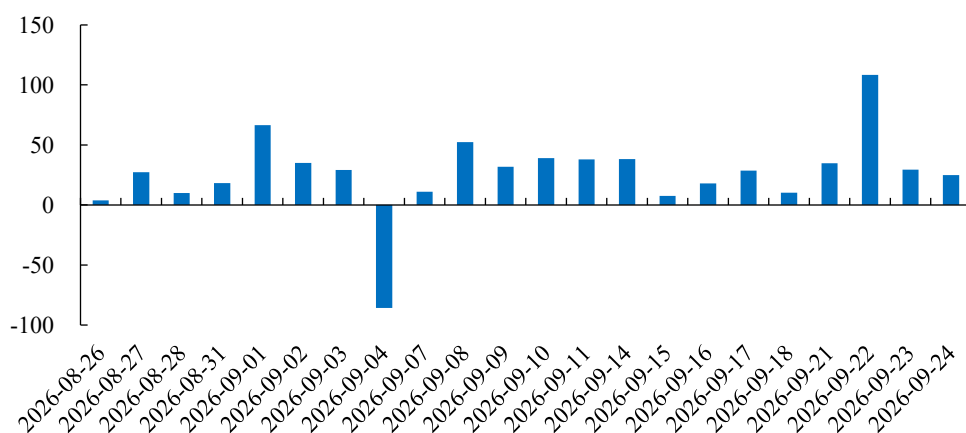
图4: 本周 (2026.09.21-2026.09.25) 恒生汽车主题、恒生科技跌幅较大 (%)



数据来源: Wind、开源证券研究所

本周港股通成交净额为流入 197.63 亿元, 环比增加, 从头部活跃个股交易层面来看, 腾讯 (+17.0 亿港元)、阿里巴巴-W (+12.4 亿港元)、智谱 (+9.3 亿港元) 资金净流入居前。

图5: 2026 年 8 月以来港股通当日买入成交净额整体为正 (亿元)



数据来源: Wind、开源证券研究所

表2: 本周 (2026.09.21-2026.09.25) 腾讯控股、阿里巴巴-W、智谱南向流入资金居前

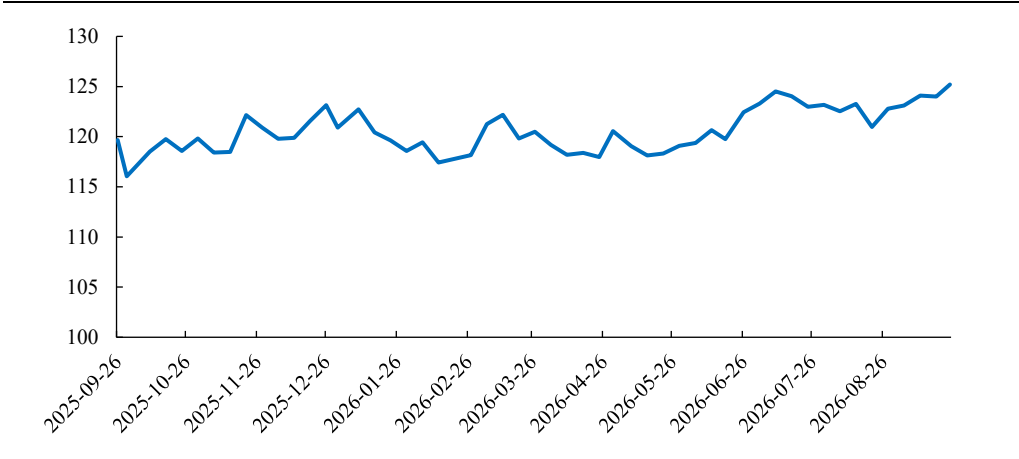
| 股票代码    | 股票名称         | 成交净买入 (亿港元) | 港股通持股数占总股本比例 |
|---------|--------------|-------------|--------------|
| 0700.HK | 腾讯控股         | 17.0        | 11.9%        |
| 9988.HK | 阿里巴巴-W       | 12.4        | 10.0%        |
| 2513.HK | 智谱           | 9.3         | 14.2%        |
| 6869.HK | 长飞光纤光缆       | 6.1         | 50.1%        |
| 3033.HK | 南方东英恒生科技 ETF | 5.6         | 0.0%         |
| 1347.HK | 华虹宏力         | 5.3         | 16.2%        |
| 3308.HK | 中际旭创         | 3.9         | 9.5%         |

| 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 |
|---------|-----------|------------|--------------|
| 0100.HK | MINIMAX-W | 2.8        | 10.3%        |
| 3986.HK | 兆易创新      | 2.1        | 52.3%        |
| 2476.HK | 胜宏科技      | 0.9        | 27.1%        |
| 1810.HK | 小米集团-W    | 0.6        | 19.6%        |
| 0883.HK | 中国海洋石油    | -0.4       | 22.3%        |
| 3696.HK | 英矽智能      | -0.5       | 27.2%        |
| 1548.HK | 金斯瑞生物科技   | -1.1       | 23.2%        |
| 9926.HK | 康方生物      | -2.7       | 47.0%        |
| 0148.HK | 建滔集团      | -3.5       | 24.5%        |
| 2099.HK | 中国黄金国际    | -5.2       | 18.5%        |
| 0981.HK | 中芯国际      | -11.5      | 28.6%        |
| 1888.HK | 建滔积层板     | -16.2      | 12.0%        |

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6：恒生沪港通 AH 溢价指数持续回弹



数据来源：Wind、开源证券研究所

4、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（2026.09.21-2026.09.25）新上线模型包括 MiMo-V2.6-Pro、MiMo-V2.6-Flash、Grok 4.7、GPT-6 Sol、GPT-6 Luna 与 Claude Opus 5.5；待正式发布模型有 Gemini 4 Pro 与 Claude Fable 5.2。

表3：9 月以来新上线模型与待发布模型

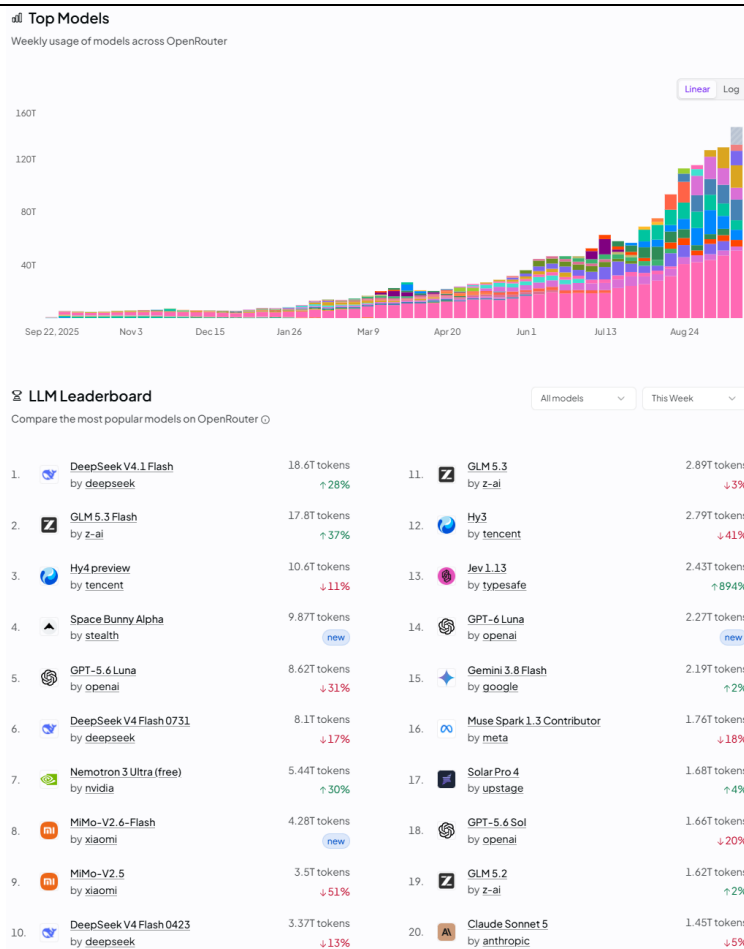
| 已发布模型             | 发布时间       | 厂商        | 输入价格（/M） | 缓存命中输入价格（/M） | 输出价格（/M） | 上下文窗口 |
|-------------------|------------|-----------|----------|--------------|----------|-------|
| Claude Fable 5.1  | 2026-09-01 | Anthropic | \$10.00  | \$0.25       | \$50.00  | 1M    |
| Claude Mythos 5.1 | 2026-09-01 | Anthropic | \$10.00  | \$0.25       | \$50.00  | 1M    |
| Gemini 3.8 Flash  | 2026-09-02 | Google    | \$0.75   | \$0.08       | \$3.75   | 1M    |

| 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 |
|---------------------|------------|-----------|-------------------------------|---------------|-----------|-------|
| 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 |
| DeepSeek-V4.1-Flash | 2026-09-10 | DeepSeek  | ¥2.00 (高峰)                    | ¥0.04         | ¥8.00     | 1M    |
|                     |            |           | ¥1.00 (空闲)                    | ¥0.02         | ¥4.00     | 1M    |
| 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    |
| 待正式发布模型             |            |           | 发布时间                          |               |           |       |
| Gemini 4 Pro        |            |           | Google                        |               |           |       |
| Claude Fable 5.2    |            |           | Anthropic                     |               |           |       |

资料来源：各模型厂商官网、IT之家、新智元、开源证券研究所

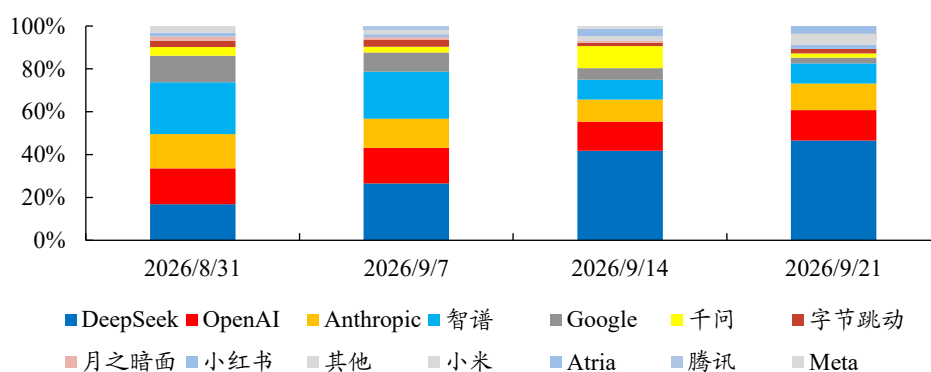
本周（2026.09.21-2026.09.25）OpenRouter 的 token 调用量持续提升，其中国产模型 DeepSeek V4.1 Flash、GLM 5.3 Flash 与 Hy4 preview 位居前三。截止 9 月 21 日，在 ZenMux 统计的 Token 市场份额中，国产 AI 占比超过 60%，DeepSeek 市场份额保持第一，占比为 46.6%。依托开源生态及出色的性价比优势，国产模型持续扩大市场影响力。

图7: OpenRouter 的 token 调用量持续提升



资料来源: OpenRouter 官网

图8: 国产 AI 在 ZenMux 的 Token 市场份额占比超过 60%



数据来源: ZenMux、开源证券研究所

## 5、投资建议

**AI 芯片：**Muse 的诞生为 Agent 提供了一个新的发展思路—常驻型 Agent，对未来的 24 小时不间断工作 Agent 有了更强的指引，也对 AIDC 的 CPU 数量提出了新的要求；电芯片方面，随着 100G/通道逐步向 400G/通道发展，电芯片的各种性能指标将会有新的要求，或进一步推进电芯片价值量提升，从而拉动电芯片代工厂的价值提升。建议关注电芯片核心代工厂 Tower 半导体；CPU 老牌企业且有独立 72 卡 AI 服务器设计能力的公司有望受益 Agent 渗透，受益标的如 AMD。

**AI 整机：**中型云、政企私有推理、新云客户更偏好品牌 OEM 的整机交付、维保与合规能力，推理负载占比上行长期利好品牌厂商，建议关注联想集团，受益标的如 DELL。

**AI 设备：**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，受益标的：ASMP

## 6、风险提示

**地缘政治影响下的原材料供应风险：**用于磷化铟激光器的高纯度铟的主要供应来自中国，如国际局势紧张，或导致关键原材料出口受限，从而影响激光器出货。

**量产良率爬坡风险：**AI 服务器属于复杂系统集成，涉及多个芯片和器件之间的调试和平衡，如出现良率问题，可能导致相关服务器出货延迟。

**业内竞争加剧风险：**目前原生 AI 芯片企业和云计算巨头同时研发 AI 芯片，如出货量大于实际需求，则可能导致竞争加剧而产品价格出现波动的风险。

## 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