



通信行业研究

买入（维持评级）

行业周报

证券研究报告

通信组

分析师：张真桢（执业 S1130524060002）

zhangzhenzhen@gjzq.com.cn

HBM 市场极度紧缺，字节获 296 亿美元银团贷款

通信周观点：

1) 智谱 2026 年上半年实现营收 9.54 亿元，同比增长 399.7%，其中开放平台及 API 业务收入同比飙升 2735.7%，占总收入比例跃升至 86.5%，8 月 ARR 已达 16 亿美元，商业化闭环加速形成。尤为值得注意的是，智谱首次在超大规模推理需求中全面使用国产芯片，10 万级国产芯片集群的规模化推进验证了国产算力的可用性。国产大模型公司业绩增速迅猛，商业化进程加速推进，国内 AI 模型及算力生态长期发展势头强劲，看好相关产业链持续受益。2) 字节跳动获 296 亿美元的银团贷款，为今年亚洲规模第二大的美元贷款，充裕的现金流储备为公司激进扩张提供了有力支撑。公司正考虑将今年资本支出提升至高达 700 亿美元，超过去年总额两倍，重点用于扩建数据中心集群及其他 AI 基础设施。头部 CSP 大厂融资能力持续加强，为后续大规模 AI 服务器芯片采购及 IDC 建设提供了良好的现金流基础，国内光模块、服务器及 IDC 产业景气度有望持续向上，产业链订单能见度进一步提升。3) 燧原科技科创板 IPO 发行价锁定 142.18 元/股，网上发行有效申购户数达 703.2045 万户，回拨后最终中签率仅为 0.02455%，显示市场对国产算力芯片资产的高度追捧。燧原科技与摩尔线程、沐曦股份、壁仞科技并称为“国产 GPU 四小龙”，随着又一家国产芯片公司登陆科创板，国产算力芯片板块持续扩容，产业链关注度与估值体系有望迎来系统性提升，国产 GPU 产业生态逐步走向成熟。4) 海外大模型竞赛持续升温，密集迭代节奏不减。OpenAI 正式发布 GPT-6 Astra，在代表高阶数学能力的 FrontierMath Tier 4 上得分达 97.6%，在强调抽象推理的 ARC-AGI-3 上得分从 7.8% 跃升至 99.9%，并在漏洞利用测试 ExploitBench 中取得满分 100%，展示出强大的通用智能水平和网络安全能力。此外，Anthropic 发布 Claude Fable 5.1 与 Mythos 5.1，在多项基准测试中超越前代及 GPT-5.6 Sol。海外大模型密集迭代，AI 能力边界持续扩展，下游应用场景不断拓宽，算力需求有望长期受益，AI 算力产业链高景气度具备持续性。5) HBM 供需缺口持续拉大，产业链景气度维持高位。三星已将 2031 年前约 70% 的 HBM 产能分配给长约订单，致使现货市场极度紧缺，HBM 现货价已涨至长约价的 4-5 倍。2026 年全球 HBM 市场规模预计增长 58% 至 546 亿美元，占 DRAM 市场近四成，产能缺口仍维持在 50% 至 60%。HBM 抢货情况相当激烈，供给紧缺格局短期难以逆转，存储芯片与先进封装产业链景气度有望维持高位，相关设备及材料供应商将持续受益于扩产周期。6) 国产存储迎来里程碑式突破。长鑫存储 LPDDR6 内存正式量产，实现全球首次落地商用，标志着中国存储企业首次在高端内存标准上打破海外厂商长期垄断产品先发的局面，具有重要的产业战略意义。目前长鑫存储已成长为全球第四大 DRAM 生产商，同时已开始小批量生产 HBM3E，计划 2027 年扩大规模，有望显著缓解国产 AI 处理器在 HBM 上的关键制约，补齐国产 AI 芯片供应链短板。随着国产芯片进入好用阶段，国产存储及算力产业链价值量有望进一步抬升，带动上游设备与材料环节需求放量。7) DeepSeek 计划在内蒙古建设吉瓦级大型数据中心，内部署至少 16 万颗华为昇腾 950DT 芯片，若建成将成为已知规模最大的华为 AI 芯片集群之一。这一大规模部署计划充分印证了国产芯片正从“可用”向“好用”阶段实质性迈进，利好国产算力模块，国产 AI 芯片设计、制造及配套服务环节将深度受益。

投资建议与估值

建议关注海外 AI Capex 持续超预期带动的光模块、光纤光缆与服务器板块；国内大模型持续迭代，国产芯片适配度进一步提升，建议关注国产算力链。

风险提示

AI 资本开支节奏不及预期；地缘政治与出口管制不确定性；存储价格大幅波动。



内容目录

一、细分行业观点.....	3
二、核心数据更新.....	5
三、本周行情.....	7
四、本周重要新闻.....	8
风险提示.....	11

图表目录

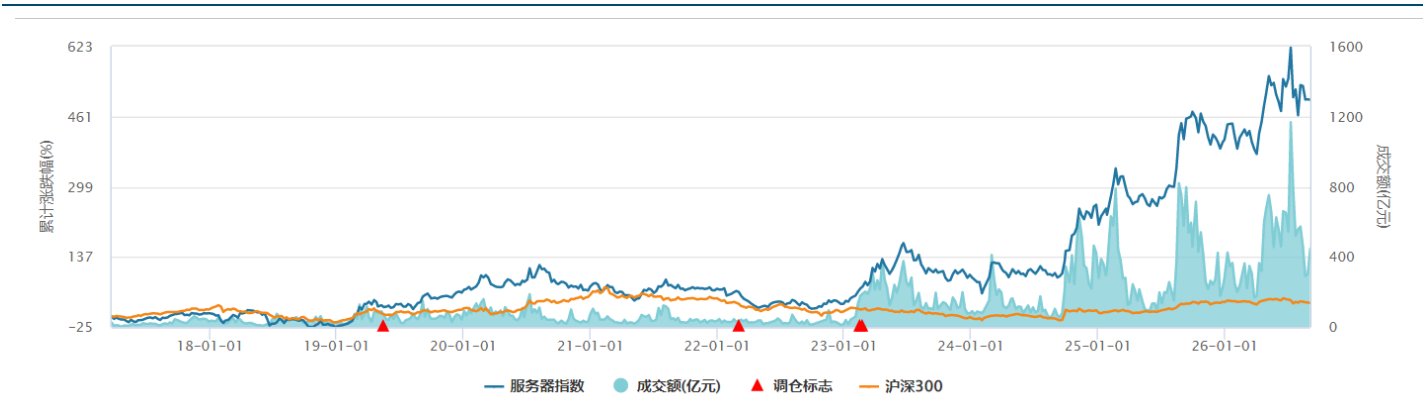
图表 1: 服务器指数 (8841058.WI) 走势	3
图表 2: 光模块 (CP0) 指数 (8841258.WI) 走势	3
图表 3: IDC 指数 (866052.WI) 走势	4
图表 4: 光纤指数 (884944.WI) 走势	4
图表 5: 本周通信板块景气度	4
图表 6: 千兆用户占比超三成	5
图表 7: 截至 6 月末 5G 用户占比约七成	5
图表 8: 上半年移动互联网累计流量同比增长 17.7%	5
图表 9: 6 月当月 DOU 达 23.57GB/户·月	5
图表 10: 电话通话量持续下滑	6
图表 11: 移动短信业务量大幅下降	6
图表 12: 千兆光纤宽带网络建设稳步推进	6
图表 13: 5G 网络建设持续深化	6
图表 14: 5 月光模块出口金额累计同比增加 18.6%	6
图表 15: 5 月光模块出口金额当月同比增加 50.7%	6
图表 16: 截至 6 月末蜂窝物联网终端用户数同比增长 5.76%	7
图表 17: 26Q1 全球蜂窝物联网模组出货量同比增长 16%	7
图表 18: 板块周涨跌幅排序 (%)	7
图表 19: 通信板块个股周涨跌幅 (剔除*ST 公司)	8



一、细分行业观点

服务器：本周服务器指数-0.08%，本月以来，服务器指数-4.45%。英伟达正式通知微软、谷歌、甲骨文等大客户，搭载其 AI 芯片的服务器价格将普遍上涨超 15%，涨价压力沿产业链向云厂商与终端传导，AI 服务器需求与整机价值量同步提升，服务器产业链景气度延续。本周 MiniMax 宣布 H3 Max 768P/480P 正式接入开放平台，5 秒内生成 768p 音视频。视频生成实时化将大幅拉动国产推理算力需求，国内 AI 算力产业链景气度持续向上。

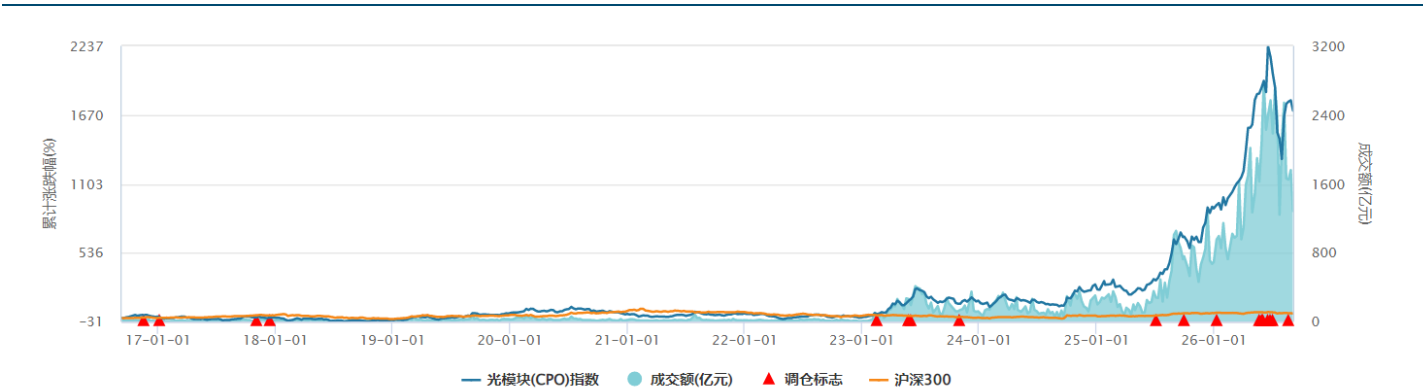
图表1：服务器指数 (8841058.WI) 走势



来源：wind，国金证券研究所

光模块：本周光模块 (CPO) 指数-4.48%，本月以来，光模块 (CPO) 指数-6.19%。中际旭创 26H1 营收同比增长 182.49%，归母净利润同比增长 241.7%；新易盛 26H1 营收同比增长 100.34%，归母净利润同比增长 90.98%。AI 数据中心需求持续拉动光模块及光器件放量，光模块产业链景气度持续验证。

图表2：光模块 (CPO) 指数 (8841258.WI) 走势

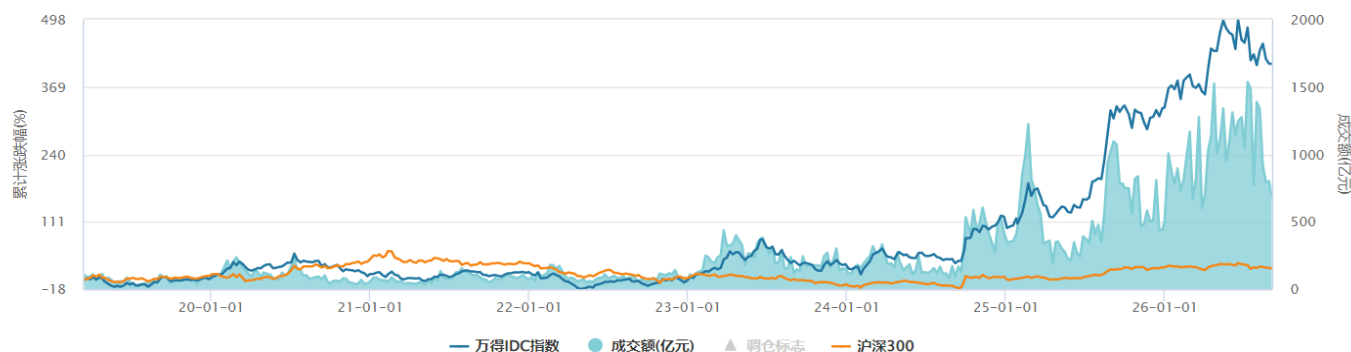


来源：wind，国金证券研究所

IDC：本周 IDC 指数-0.09%，本月以来，IDC 指数-3.36%。字节跳动已获得一笔 296 亿美元的银团贷款，这是今年亚洲规模第二大的美元贷款，国内 CSP 投入资金充足，国内 IDC 需求景气度持续受云厂商 Capex 增长拉动。



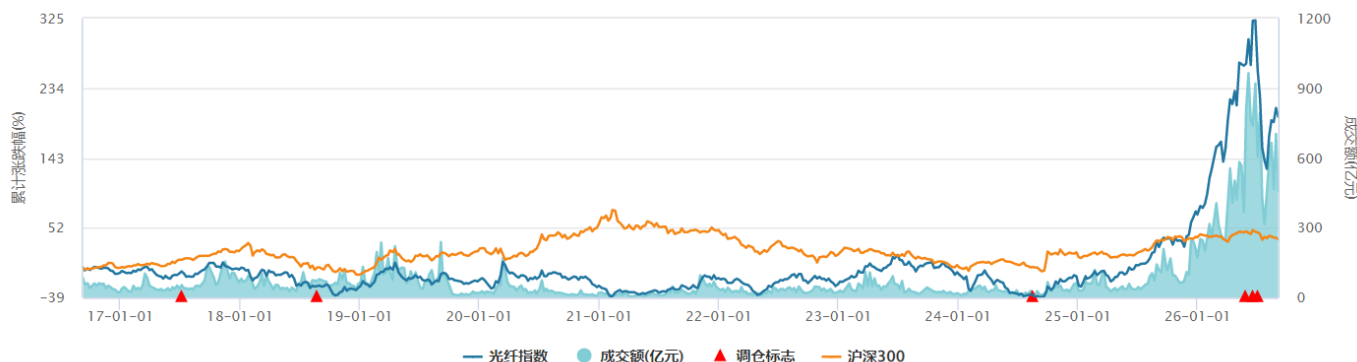
图表3: IDC 指数 (866052.WI) 走势



来源: wind, 国金证券研究所

光纤: 本周光纤指数-3.73%, 本月以来, 光纤指数-3.46%。光纤涨价周期有望持续, 同时多芯光纤、空芯光纤等新技术变革将进一步打开成长空间。

图表4: 光纤指数 (884944.WI) 走势



来源: wind, 国金证券研究所

图表5: 本周通信板块景气度

板块	景气度指标	本期景气度说明
运营商	稳健向上	5G 投资周期结束, 电信行业端承压, 但云与 IDC 业务放量接力成长, 整体景气度稳健向上。
光模块	稳健向上	中际旭创 2026 年上半年营收同比增长 182.49%, 归母净利润同比增长 241.7%; 新易盛 2026 年上半年营收同比增长 100.34%, 归母净利润同比增长 90.98%。AI 数据中心需求持续拉动光模块及光器件放量, 光模块产业链景气度持续验证。
服务器	稳健向上	英伟达正式通知搭载其 AI 芯片的服务器价格将普遍上涨超 15%, AI 服务器产业链景气度延续。
交换机	稳健向上	26H2 及 2027 年超节点成为国产芯片主力出货方向, 交换网络的层级和带宽要求持续提升, 交换机板块需求旺盛。
连接器	稳健向上	康宁计划将在美国的光连接制造能力扩大 10 倍、光纤生产能力扩大 50%以上, 看好光无源器件释放业绩。
IDC	加速向上	字节跳动已获得一笔 296 亿美元的银团贷款, 这是今年亚洲规模第二大的美元贷款, 国内 CSP 投入资金充足, 国内 IDC 需求景气度持续受云厂商 Capex 增长拉动。
物联网	加速向上	苹果加码推进端侧 AI, 依托自研芯片在 iPhone、Mac 等终端设备本地运行 AI 模型, 同时正考虑收购端侧 AI 初创公司 LiquidA。
液冷	高景气维持	《促进人工智能与能源双向赋能行动方案》明确新建万架级智算中心 100%液冷、28 年前完成全部风冷存量机房改造;20kW 以上机柜强制液冷, 液冷板块高景气维持。

来源: 红星新闻, 证券之星, 看点资讯, 今朝新闻, 东方财富, 新浪财经, 光纤在线, 今日头条, 国金证券研究所



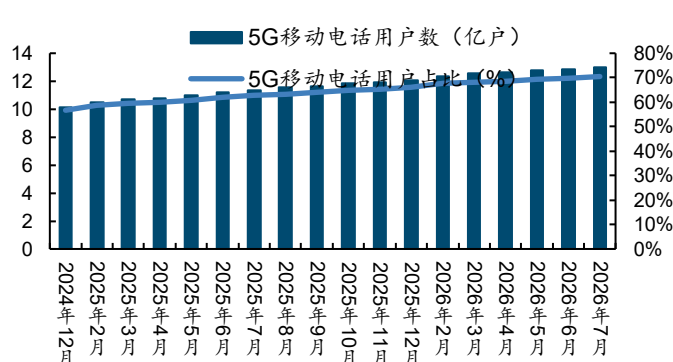
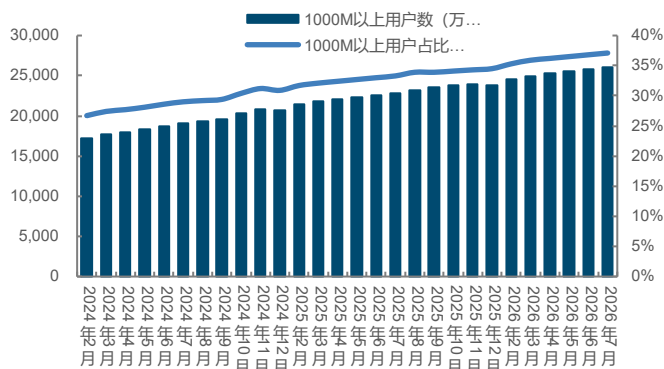
二、核心数据更新

运营商：运营商数据维持稳健增长

截至7月末，三家基础电信企业及中国广电的移动电话用户总数达 18.47 亿户，比上年末净增 2062 万户。其中，5G 移动电话用户达 13.03 亿户，比上年末净增 9875 万户，占移动电话用户的 70.5%。

图表6：千兆用户占比超三成

图表7：截至6月末5G用户占比约七成



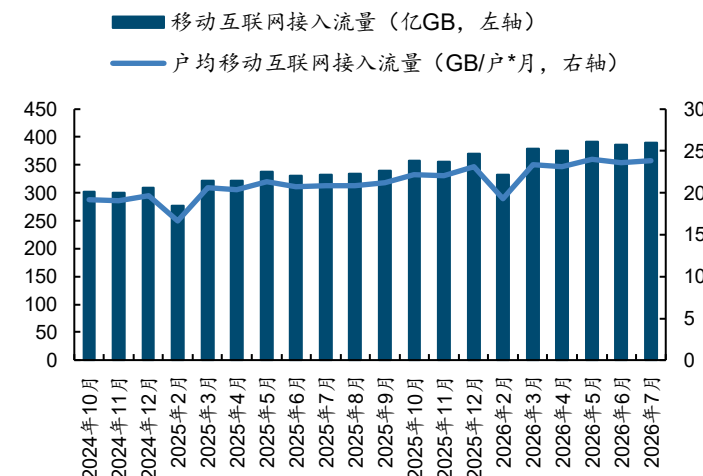
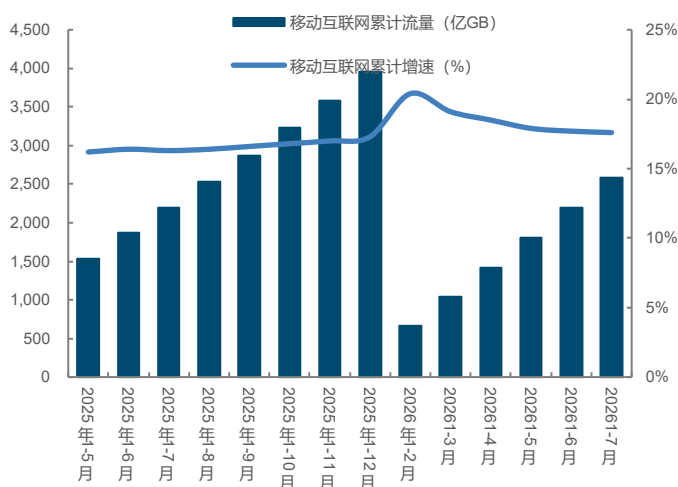
来源：工信部，国金证券研究所

来源：工信部，国金证券研究所

前7个月，移动互联网累计流量达 2587 亿 GB，同比增长 17.6%。前7个月，移动互联网用户数达 16.38 亿户，比上年末净增 2858 万户。7月当月移动互联网接入流量（DOU）达到 23.81GB/户·月，同比增长 13.9%，比上年底高 0.77GB/户·月。

图表8：上半年移动互联网累计流量同比增长 17.7%

图表9：6月当月 DOU 达 23.57GB/户·月



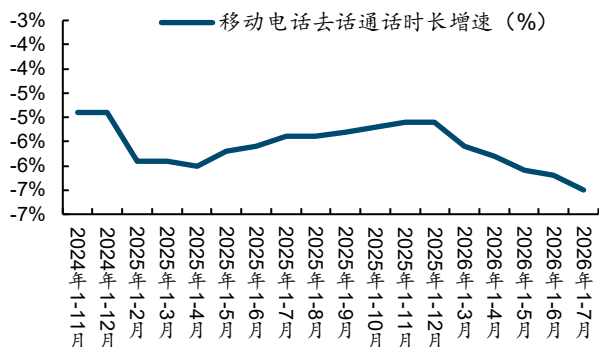
来源：工信部，国金证券研究所

来源：工信部，国金证券研究所

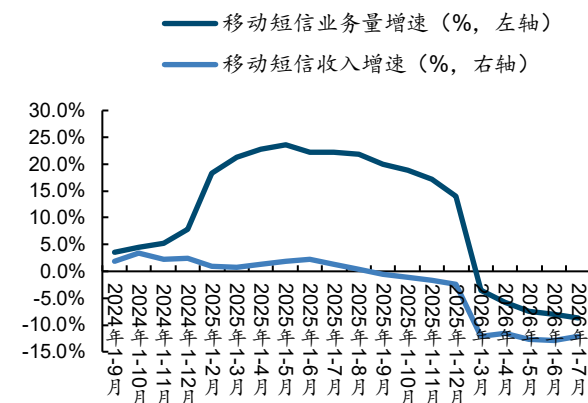
前7个月，移动电话去话通话时长完成 1.1 万亿分钟，同比下降 6.5%；固定电话主叫通话时长完成 327.6 亿分钟，同比下降 18.6%。前7个月，全国移动短信业务量同比下降 8.7%；移动短信业务收入同比下降 12.2%。



图表10：电话通话量持续下滑



图表11：移动短信业务量大幅下降

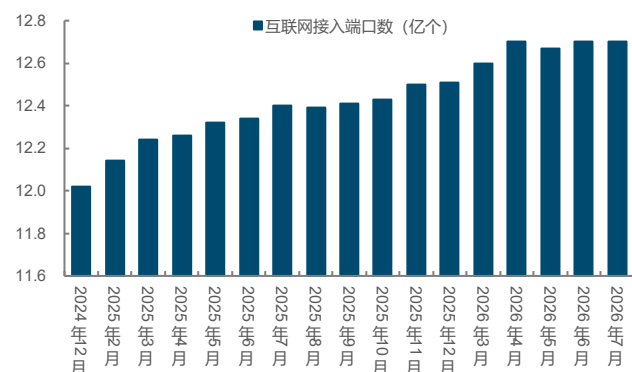


来源：工信部，国金证券研究所

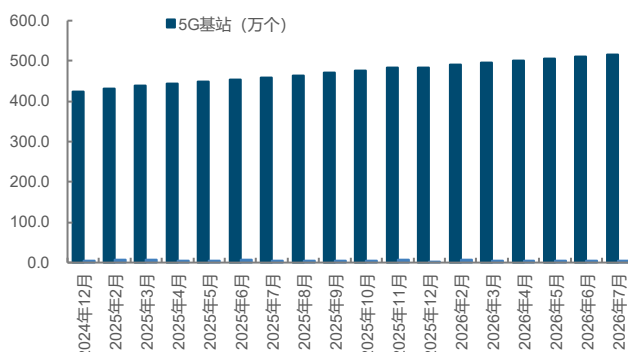
来源：工信部，国金证券研究所

截至7月末，全国互联网宽带接入端口数量达12.7亿个，比上年末净增2152万个。其中，光纤接入（FTTH/O）端口达到12.3亿个，比上年末净增1935万个，占互联网宽带接入端口的96.7%。截至7月末，具备千兆网络服务能力的10G PON端口数达3304万个，比上年末净增141.7万个。

图表12：千兆光纤宽带网络建设稳步推进



图表13：5G网络建设持续深化

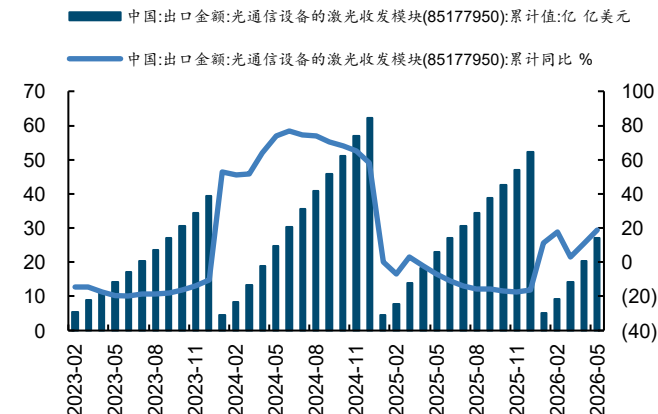


来源：工信部，国金证券研究所

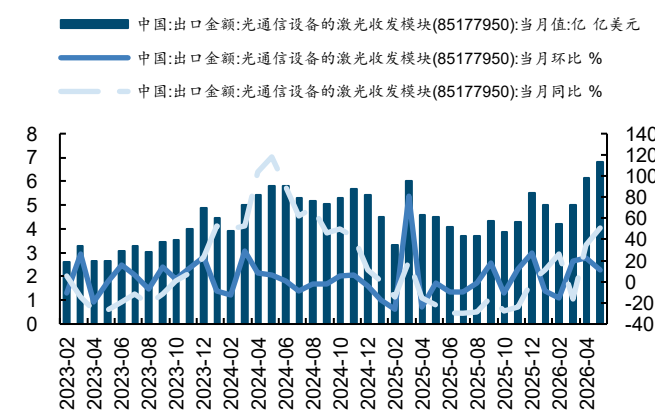
来源：工信部，国金证券研究所

光模块数据：2026年5月我国光模块出口数据同比+50.67%；1-5月累计同比+18.6%。

图表14：5月光模块出口金额累计同比增加18.6%



图表15：5月光模块出口金额当月同比增加50.7%



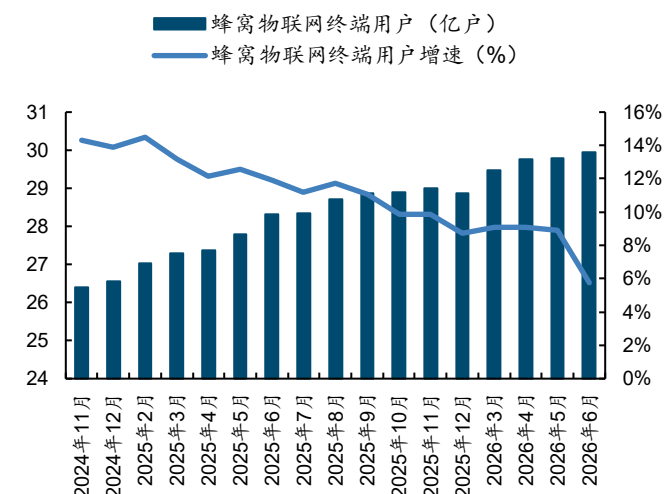
来源：wind，国金证券研究所

来源：wind，国金证券研究所

物联网数据：截至6月末，三家基础电信企业发展移动物联网终端用户29.94亿户，比上年末净增1.06亿户。互联网电视（IPTV、OTT）用户数达4.09亿户，比上年末净增99万户。

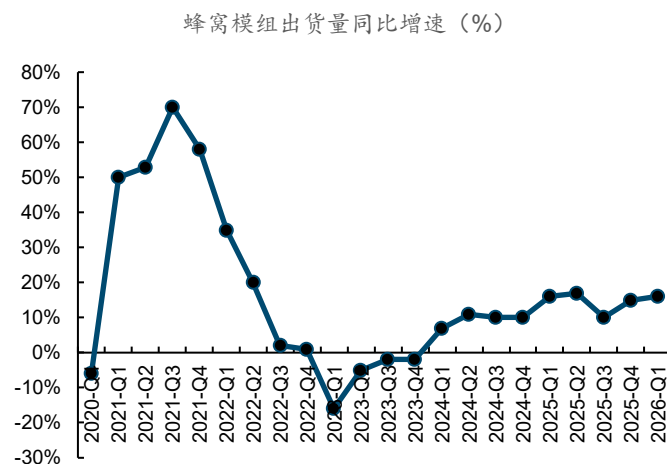


图表16：截至6月末蜂窝物联网终端用户数同比增长5.76%



来源：工信部，国金证券研究所

图表17：26Q1 全球蜂窝物联网模组出货量同比增长16%

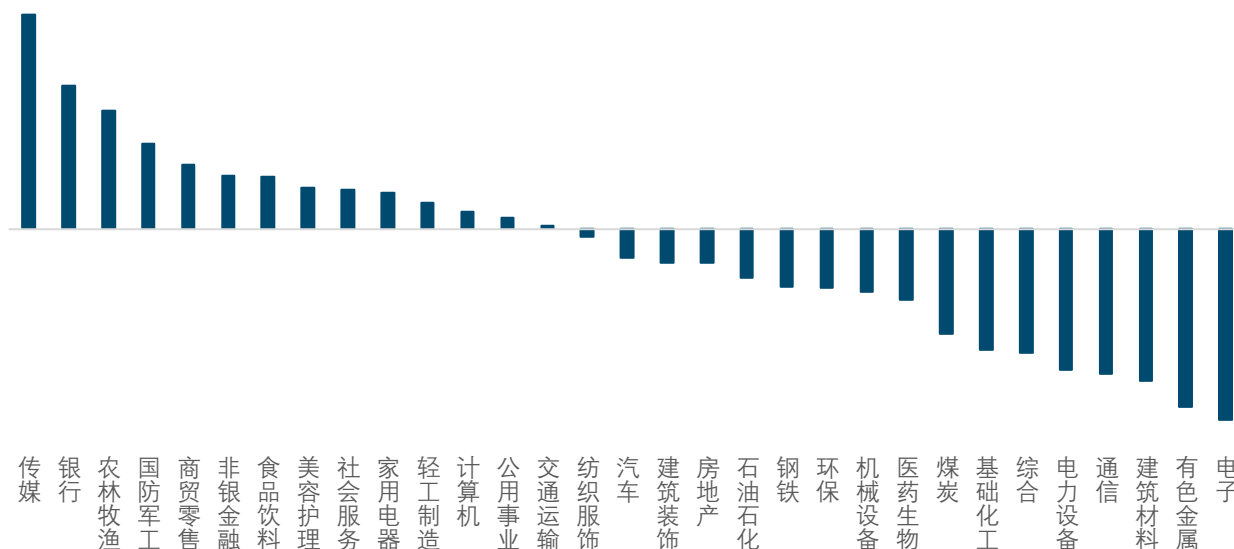


来源：国金证券数字未来实验室，RFID 世界网，C114 通信网，国金证券研究所

三、本周行情

回顾本周行情（8月31日-9月4日），参考申万一级行业划分，通信板块涨跌幅为-4.06%，排名全行业第二十八。

图表18：板块周涨跌幅排序 (%)

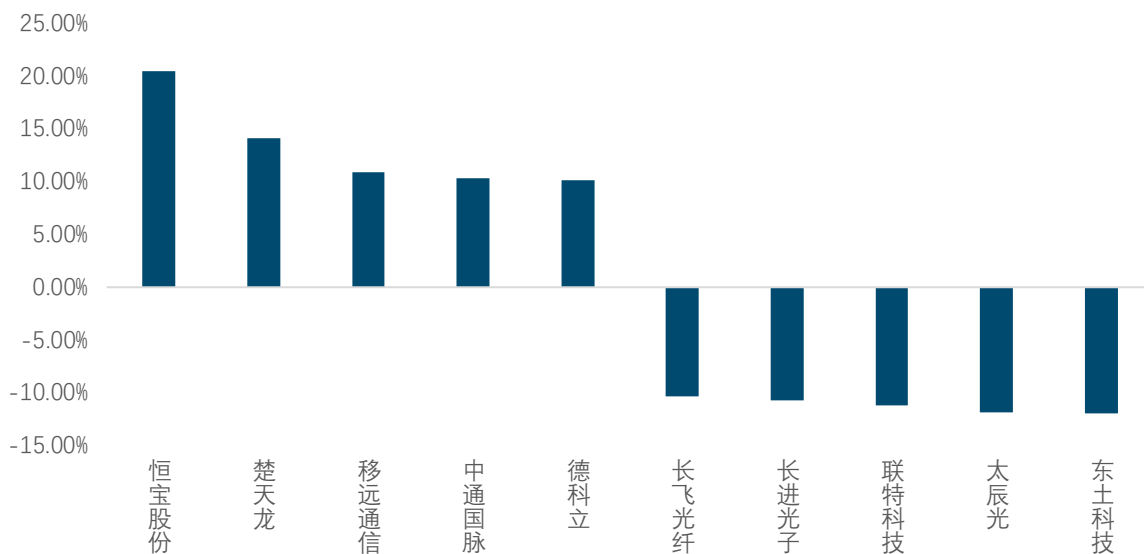


来源：wind，国金证券研究所

从个股情况来看，本周恒宝股份、楚天龙、移远通信、中通国脉、德科立为通信（申万）涨幅前五大公司，涨跌幅分别为+20.50%、+14.16%、+10.94%、+10.33%、+10.10%。通信（申万）跌幅前五为东土科技、太辰光、联特科技、长进光子、长飞光纤，涨跌幅分别为-11.99%、-11.89%、-11.17%、-10.75%、-10.31%。



图表19：通信板块个股周涨跌幅（剔除*ST公司）



来源：wind，国金证券研究所

四、本周重要新闻

4.1 行业新闻

三星将 2031 年前约 70% 的 HBM 产能分配给长约订单，HBM 抢货激烈

8 月 31 日，三星存储器业务部门已将 2031 年前约 70% 的 HBM 产能分配给长约订单，主要签约客户包括英伟达、微软及谷歌等大型科技公司，但连这些大型科技公司都无法取得合同约定的全部供货量，HBM 抢货情况相当激烈。据内存市场研究公司 MegaGrid Supply 数据，一块 36GB 的 HBM3E 产品现货售价为 2100 美元，约为长约价格的四到五倍；16 层 HBM4 现货价格更高达 3500 美元。供应压力正从 HBM 蔓延至一般 DRAM，随着向英伟达等大型科技客户供应的 HBM4 持续增加，HBM 生产吸收大量 DRAM 产能，进一步推升 DRAM 价格，且 HBM4 仍处于量产初期，生产良率低于上一代 HBM3E，制造过程需要耗用更多 DRAM，也使整体 DRAM 供应短缺进一步恶化。三星电子和 SK 海力士均认为存储器短缺短期内无法解决，正从多方面寻求产能扩张。据财联社报道，三星电子第三季度营收预计达 206.64 万亿韩元，营业利润将达到 116.38 万亿韩元，远高于第二季度的 171.50 万亿韩元和 89.49 万亿韩元；SK 海力士第三季度营收预计达 101.76 万亿韩元，比上一季度增长 28.2%；营业利润为 79.16 万亿韩元，增长 30.7%。HBM 供给紧缺格局延续，存储芯片与先进封装产业链景气度有望维持高位。

SK 海力士考虑英特尔代工 HBM4E 基底芯片

据财联社 8 月 31 日报道，韩媒爆料 SK 海力士正考虑将 HBM4E 的基底芯片交由英特尔代工，但该合作目前仅处于评估阶段，尚未敲定最终协议。当前，SK 海力士 HBM4 基底芯片全部由台积电以 12nm 级工艺生产，台积电基础芯片成本比 SK 海力士自产核心芯片高出 3 至 4 倍。今年 6 月，SK 海力士已向头部客户送出 12 层 HBM4E 样品，英伟达签下 2790 亿美元多年内存采购大单。由于 HBM 芯片大多签订长期供货协议，成本上涨很难直接转嫁给下游客户。此前，SK 海力士 HBM4 曾出现交付延迟，被三星抢占了部分市场份额。对于 SK 海力士而言，双供应商模式将是关键的供应链避险策略，避免单一工厂出现产能不足和延期交付风险。尽管该合作短期不会对台积电营收造成实质影响，但长远来看，若英特尔能实现合格良率，三星和美光大概率也会跟进评估同样方案，台积电在先进封装领域的议价能力可能被逐步削弱。HBM 产业链分工进一步细化，先进封装与散热环节价值量持续提升。

韩国 8 月半导体出口 466.5 亿美元同比增 209%，今年第五次刷新月度纪录

据韩国产业通商资源部 9 月 1 日发布的“8 月进出口动向”，8 月半导体出口额达 466.5 亿美元，同比增长 209%，创月度历史新高，这是今年第五次刷新纪录——2 月以 251 亿美元创当时月度最高，3 月达 328 亿美元，5 月达 372 亿美元，6 月达 448 亿美元，8 月再创新高，出口额已连续三个月突破 400 亿美元。8 月整体出口额 982.5 亿美元，同比增长 68.7%，为月度出口业绩中第三高（历史最高为 6 月 1020 亿美元，第二为 7 月 990 亿美元）；考虑工作日因素，日均出口额 44.7 亿美元，同比增长 72.5%。分类别看，计算机出口 62.4 亿美元增长 419.5%，创月度历史最高，得益于企业级 SSD 核心部件 NAND 闪存价格持续上涨；无线通信设备出口因 Galaxy S26 系列销售增长而增加 21.2%，连续 10 个月保持增长。按地区看，对华出口 241 亿美元增长 119.3%，对美出口 165 亿美元增长 89.3%，对东盟出口 190.9 亿美元增长 75.4%，创历年 8 月最高。进口方面，8 月进口额 635.1 亿美元增长 22.5%，贸易顺差 347.5 亿美元，连续三



个月保持在 300 亿美元以上。产业部长官金正宽表示，在半导体出口持续走强的同时，半导体以外品类出口也录得 20% 高增长率，出口基础正在扩大，但美国关税政策、欧盟钢铁免关税配额、中东局势等贸易环境不确定性较高，需密切审视。全球 AI 基础设施投资持续扩大，半导体产业链景气度延续，国内光模块、光器件等供应链环节有望受益。

特朗普政府考虑对进口半导体加征新一轮关税

据彭博社 9 月 2 日报道，美国商务部长霍华德·卢特尼克表示，特朗普政府正考虑对进口半导体加征新一轮关税，以吸引更多制造业回流美国。卢特尼克对 CNBC 表示，关税措施很可能与对投资美国本土制造业的公司提供的关税减免配套推出，效仿制药行业做法——“如果你在这里建厂，你就不用交税；但如果你不在这里建厂，那就准备好交钱才能进入全球最大的市场”。卢特尼克表示，所有公司都知道这些政策要来了。目前已有约 1.2 万亿美元投资承诺用于在美国建设半导体产能，台积电等公司正在投资美国本土制造设施，卢特尼克透露台湾下周还将宣布新一轮 200 亿至 300 亿美元投资承诺。全球半导体贸易政策不确定性上升，国产芯片自主可控逻辑有望强化，国产算力产业链价值量有望重估。

大模型 API 调用价格持续下降，AI 推理成本进入“1 美元时代”

据 21 世纪经济报道 9 月 4 日报道，大语言模型 Token 支出指数首次跌破每百万 Token 1 美元关口，报 0.97 美元，创历史最低纪录。该指数以彭博代码 SDLLMTK 发布，按使用量加权计算全球主流大模型的实际成交价格，被市场视为衡量 AI 推理价格的核心基准。这一价格已不到今年夏季早些时候高点的一半，仅过去 7 天就下跌了 8.6%。本轮价格下行中，中国模型厂商的策略调整加速了行业价格体系的整体下移。DeepSeek、月之暗面等公司普遍采用混合专家架构压低算力消耗；今年 5 月 DeepSeek 将 V4-Pro 的 API 价格永久下调至原价的四分之一；中国头部模型几乎全部开放权重，开发者可免费私有化部署。9 月 1 日阿里云宣布对 Qwen3-VL-Rerank 模型进行降价：文本输入从 0.7 元/百万 tokens 下调至 0.5 元/百万 tokens，多模态输入从 1.8 元/百万 tokens 大幅下调至 0.5 元/百万 tokens。9 月 3 日，阿里云更新旗舰模型 Qwen3.8-Max，新版整体性能进一步提升，已在千问 AI 平台提供 API 服务。面对中国模型的价格策略，OpenAI 在 7 月底下调了两款 GPT-5.6 价格，其他模型厂商也相继推出动态定价。9 月 2 日，月之暗面宣布 Kimi API 已原生支持 OpenAI Responses API 格式与 Anthropic Messages API 格式，Codex 和 Claude Code 用户无需转换即可直连 Kimi K3 等多款模型。Syz Group 首席投资官表示，Token 价格通缩将压缩营收，但运算资源的承诺成本却维持固定。大模型从“拼参数”进入“拼价格”阶段，AI 使用成本持续走低，推理算力需求有望加速释放。

4.2 公司新闻

MiniMax:

8 月 31 日，MiniMax 宣布 H3 Max 768P、H3 Max 480P 正式接入开放平台和 MiniMax Design。H3 Max 生成一段长度 5 秒、768p 的音视频耗时不到 3 秒，已跨过实时直播所需的速度门槛。海外开发者已围绕 H3 Max 搭建 Twitch 直播、24 小时“AI 电视台”，以及邀请观众实时改写剧情的创作实验。fal 工程师将 H3 Max 接进直播流，做成“跨维度电视”，在 X 上的观看量已近 350 万次。知名开发者 Pieter Levels 也上线了 24 小时 AI 直播网站，观众在聊天室输入内容即可实时改变下一段视频画面。业绩方面，8 月 26 日 MiniMax 发布上市后首份中期业绩，上半年总收入 1.17 亿美元，同比增长 283.1%，已超过 2025 年全年收入；其中开放平台及其他 AI 企业服务收入同比增长 703.1% 至 7390 万美元，占比升至 63.4%，成为第一大收入来源；净亏损 3.58 亿美元，同比收窄 11%。公司管理层在业绩会上披露，7 月平台 Token 消耗量已达 1 月的 20 倍，8 月 ARR 超过 8 亿美元。H3 于 7 月 31 日发布并开放权重，三个多星期下载量超 2400 万，产生超 300 个公开衍生模型。算力方面，MiniMax 创始人、CEO 闫俊杰在业绩会上表示，公司正在推进 M3 和 H3 的国产芯片适配，大规模国产算力集群将很快上线，并逐步承担真实生产流量。截至 6 月底，公司持有现金约 13.23 亿美元。国产大模型厂商商业化进展提速，叠加国产芯片在算力侧的应用拓展，AI 算力产业链景气度有望延续。

智谱:

8 月 31 日，智谱发布 2026 年半年报。上半年实现营收 9.54 亿元，同比增长 399.7%；经调整净亏损 19.64 亿元，同比收窄 12.1%。业务结构发生根本性变化：开放平台及 API 业务收入约 8.25 亿元，同比增长 2735.7%，占总收入比例由去年同期的 15.2% 跃升至 86.5%；同期企业级通用大模型（本地化部署）收入由 1.48 亿元降至 6704 万元。商业化重心正从一次性本地化部署全面转向按调用量计费的云端 MaaS 业务。用户和调用量方面，MaaS 平台付费日活用户数较年初增长 603%，前十大用户日均调用量较年初增长 98 倍，平台 Token 调用量较年初增长超过 40 倍，API 平均售价较年初提升约 101%。模型能力方面，GLM-5.3 与 GLM-5.2 在架构和总参数保持一致的情况下，通过扩大后训练规模，端到端编码完成率提升超过 50%。智谱首次在超大规模需求中全面使用国产芯片，完成 10 万级国产芯片规模化推进验证，推动性能提升 3 倍。MiniMax 8 月 ARR 已突破 8 亿美元，远超杰富瑞此前 5.8 亿美元的预测。国产大模型商业化进程加速，训练与推理已具备全链路国产算力支撑，看好国内 AI 模型长期发展。

Anthropic:

9 月 1 日，Anthropic 正式发布新一代旗舰大模型 Claude Fable 5.1 与 Claude Mythos 5.1，公司称其为“全球最先进的编程与知识工作模型”。两款模型共享同一套底层模型，Fable 5.1 面向所有用户与企业全面开放，Mythos 5.1 则仅向经过审核的美国网络安全与生命科学机构开放。据 Anthropic 官方介绍，Mythos 5.1 的核心定位是协助美国政府及私营机构应对国家级网络安全威胁、开发生物防御及疫苗研发方案，不对外公开。Claude Fable 5.1 上下文窗口保持 500K token，支持图像、表格、图表等多模态输入，在多语言数学、推理、代码生成等基准测试中均达到该规模模型最优水平。定价方面，缓存读取费用从每百万 token 0.45 美元下调至 0.1125 美元，降幅达 75%，大幅降低了长上下文场景下的使用成本。据《华尔街日报》8 月 31 日报道，Anthropic 已与英伟达支持的云服务商 Lambda 签署了



价值 350 亿美元的云计算协议，英伟达本身持有得克萨斯州数据中心的租约，该数据中心由比特币矿商兼数据中心开发商 Hut 8 开发。Anthropic 本月早些时候还与英伟达支持的 Nscale 签署了 450 亿美元协议，租用西弗吉尼亚州设施资源。海外 AI 模型竞争持续升温，模型能力跃升与成本下降有望加速 AI 应用渗透，云服务与数据中心产业链景气度延续。

DeepSeek:

据彭博社 9 月 4 日报道，DeepSeek 计划在内蒙古建设的大型数据中心内部署至少 16 万颗华为昇腾 950DT 芯片，该数据中心规模达吉瓦级，若建成将成为已知规模最大的华为 AI 芯片集群之一。DeepSeek 计划使用昇腾 950DT 芯片运行其模型，安装时间表取决于华为生产能力。受高端内存等组件短缺影响，华为今年 950DT 芯片产量被限制在数十万片较低水平，满足 DeepSeek 订单可能需要一年以上。DeepSeek 已请求北京方面协助说服华为加快分配。若按计划建成，中国可能已拥有多个同等规模集群。

长鑫存储:

据 The Information 8 月 31 日报道，长鑫存储已开始小批量生产先进高带宽内存 HBM3E，产品目前仅落后于三星、SK 海力士和美光即将大规模量产的内存一代，计划 2027 年扩大生产规模。数家中国芯片设计公司正在测试该内存与自家处理器的兼容性，包括阿里巴巴平头哥和寒武纪，若顺利最早明年可在产品中使用长鑫存储的内存。此前数周长鑫存储在上海上市募资 86 亿美元，首日股价飙升 472%，上半年营收 1503 亿元（约合 223 亿美元），同比增长近十倍，净利润 776 亿元（约合 115 亿美元）。长鑫存储成立于 2016 年，已成长为全球第四大 DRAM 生产商，其 HBM3E 可与英伟达 Blackwell 及谷歌 TPU 等领先 AI 处理器兼容。HBM 通过堆叠存储芯片并置于处理器附近，使数据快速传输，对训练和运行大模型至关重要。华盛顿于 2024 年限制对华出口先进 HBM，长鑫存储的突破可能缓解中国自主研发 AI 处理器面临的关键制约。不过据知情人士透露，其在 HBM 技术上仍落后领先厂商三至五年，且仍在良率问题上挣扎。先进 HBM 比 DRAM 等传统内存更难制造，需将多个 DRAM 芯片进行堆叠和封装，每一步都增加失败可能性。

长存控股:

9 月 3 日，据上交所官网披露，长江存储控股股份有限公司科创板 IPO 审核状态变更为“已问询”。公司于 8 月 21 日获受理，保荐机构为中信证券、中信建投，拟募资 330 亿元，是目前科创板 IPO 拟募资规模最大的项目。募集资金中 208 亿元用于长江存储量产线技术升级项目，122 亿元用于研发相关募投项目。长存控股是国内领先的 NAND 闪存芯片制造商，产品广泛应用于智能手机、PC、服务器、数据中心等领域。国产存储芯片产业链关注度有望持续提升，上游设备、材料等配套环节有望受益。

燧原科技:

9 月 2 日，燧原科技披露首次公开发行股票并在科创板上市网上发行申购情况及中签率。本次发行股份数量为 4303.5173 万股，发行价 142.18 元/股；网上发行有效申购户数 703.2045 万户，有效申购股数 420.66 亿股。因网上发行初步有效申购倍数超 100 倍，启动回拨机制，将 344.3 万股从网下回拨至网上。回拨后，网上最终发行 1032.85 万股，最终中签率 0.02455315%。中一签需缴款 7.11 万元。燧原科技与摩尔线程、沐曦股份、壁仞科技并称为“国产 GPU 四小龙”，截至目前，摩尔线程、沐曦股份已实现 A 股上市，壁仞科技已在港股上市。伴随燧原科技上市，“国产 GPU 四小龙”将齐聚资本市场。

Crusoe:

据彭博社 9 月 4 日报道，数据中心初创企业 Crusoe 已与交易公司简街资本签署了一份为期五年的云服务合同，价值约 130 亿美元，将通过 Crusoe 云平台提供 GPU 集群及其他 AI 训练推理所需基础设施。Crusoe 此前已与 Meta、Oracle 等公司签约供应 AI 算力。Crusoe 成立于 2018 年，最初专注于从能源田捕获过剩天然气为加密货币挖矿供电，后转向 AI 基础设施领域，因开发大型数据中心项目而闻名，包括 OpenAI 旗舰站点及 Oracle 在得克萨斯州的 Stargate 项目。3 月微软同意租赁 Crusoe Stargate 项目扩建部分。Crusoe 在最新一轮融资中筹集超 30 亿美元，估值约 300 亿美元，彭博社 4 月曾报道简街资本在洽谈共同领投 Crusoe 竞争对手 Fluidstack 的一轮融资。AI 算力需求从科技巨头向金融机构等传统行业延伸，算力基础设施需求持续多元化。

Thinking Machines Lab:

据 The Information 9 月 3 日报道，由前 OpenAI 首席技术官 Mira Murati 领导的 AI 开发商 Thinking Machines Lab，正洽谈以至少 400 亿美元投前估值筹集至少 10 亿美元资金。现有投资者 Accel 正洽谈领投，英伟达也在考虑参与。拟议估值低于该公司去年年底寻求的估值（当时讨论以超 500 亿美元估值融资 40 亿至 50 亿美元）。该公司于 7 月发布了首个模型 Inkling，每年产生数亿美元收入。Thinking Machines Lab 成立不到两年即达到此估值水平，AI 初创企业融资热度不减，头部 AI 公司估值持续攀升。

4.3 海内外大厂重点跟踪

英伟达:

1) 9 月 3 日，英伟达宣布已同意以 129.3 亿美元收购全球最大开源 AI 平台 Hugging Face。英伟达承诺将维持 Hugging Face 作为开放平台运营，开发者可以自由选择所需的模型、框架、云服务及推理服务提供商，在 Hugging Face 上构



建或部署应用时无需使用英伟达的计算资源。Hugging Face 平台年化经常性收入约 1.5 亿美元。英伟达进一步掌控开源生态入口，有望带动开源大模型部署与 token 使用量进一步增长。

2) 8 月 31 日，英伟达与联发科宣布深化长期合作伙伴关系，共同构建从边缘到云端的 AI 计算平台。根据协议，英伟达已斥资 35 亿美元认购联发科发行的可转换债券。联发科将正式代工生产英伟达 NVLink Fusion 芯片小芯片、交换芯片以及定制内存，这一硬件可以把非英伟达的 GPU、CPU 以及其他专用芯片互相连接起来。联发科将采用全新 NVIDIA NVLink Fusion 平台，助力客户将定制 XPU 带入 NVIDIA NVLink 连接的机架级 AI 工厂。两家公司结合英伟达加速计算、AI、图形及软件平台优势与联发科在定制硅片、高性能计算、高能效 SoC 设计、先进封装、互连和连接技术方面的领先地位。

OpenAI:

1) 9 月 3 日，OpenAI 正式发布 GPT-6 Astra，称其为“目前全球最智能、且对齐程度最高的模型”。Astra 基于 OpenAI 有史以来最大规模的训练运行，在得克萨斯州 Stargate 站点使用了超过 10 万块 GPU，是首个被认定为达到“关键”网络安全阈值的模型，可在没有逐步人工指导的情况下跨防护良好的系统发现并利用此前未知的漏洞。在代表高阶数学能力的 FrontierMath Tier 4 上，Astra 得分达到 97.6%；在 ARC-AGI-3 上得分从 GPT-5.6 Sol 的 7.8% 跃升至 99.9%；在 ExploitBench 中得分达到 100%。

2) 9 月 2 日，OpenAI CEO 奥特曼在播客节目中确认，公司将自研人形机器人本体，同时并行开发数据中心特种机器人，原世界模拟项目升级为 OpenAI Robotics，短期研发辅助基建工人的机器人，长期目标实现人手一台个人机器人。AI 应用场景从软件向硬件端持续扩展，下游应用爆发将为算力需求提供更长期支撑。

3) 据 CNBC 报道，OpenAI 广告业务已实现 10 亿美元年化收入运行率，该业务虽仅成立约 200 天，但已在其企业服务、消费者订阅及 API 等现有收入引擎基础上发展起来。ChatGPT 广告现已在 40 多个国家上线，面向 Go 订阅用户及免费用户展示，免费用户占其每周 10 亿活跃用户的绝大多数，广告有明确标识且不会影响 ChatGPT 回答，广告商无法访问用户私人对话。公司正为预计规模庞大的 IPO 做准备，面临向投资者证明其估值合理性的压力。

博通:

9 月 2 日，博通公布 2026 财年第三季度财报，营收 295.91 亿美元，同比增长 86%，环比增长 18%，创单季营收历史新高；GAAP 净利润 87.62 亿美元，同比增长 92%；调整后每股收益 3.32 美元，超出市场预期。AI 半导体业务收入 167 亿美元，同比增长 221%，环比增长 54%，占公司总营收比重超 56%，成为整体业绩的核心驱动力。据博通财报电话会披露，谷歌签订长期协议未来几年每年将交付数百亿美元 TPU；Anthropic 部署规模从 2026 年 1GW 增至 2027 年 5GW、2028 年再增 10GW，预计 2027 年成为博通最大 XPU 客户；OpenAI 的 Jalapeno 芯片正在部署，下一代 XPU 已接近流片，预计 2028 年成为第二大 XPU 客户。公司预计第四季度营收约 348 亿美元，首次给出未来三年 AI 收入指引：2026 财年 AI 收入 580 亿美元，2027 财年翻倍至约 1150 亿美元，2028 财年达 2300 亿美元。博通为 OpenAI 开发的 Jalapeno 芯片性能优于英伟达 Grace Blackwell，成本仅为 GPU 的一半。定制芯片与光互联双轮驱动，AI 算力需求向定制化芯片与网络互联环节传导，国内光模块、光器件产业链有望持续受益。

谷歌:

9 月 3 日，谷歌发布 Gemini 3.8 Flash 与专攻网络安全的 Gemini 3.8 Flash Cyber 两款模型。距离 3.7 Flash 发布仅三周，是谷歌六周内推出的第三款 Flash 系列模型。Gemini 3.8 Flash 被定位为“迄今最聪明的 Flash 模型”，重点强化长周期软件工程、自主 Agent 任务和复杂多步骤推理。性能方面，Gemini 3.8 Flash 在 DeepSWE v1.1 测试中得分 73.7%，接近 Claude Opus 5 的 74.0%；Terminal-Bench 2.1 得分 89.4%；HLE-Verified 得分 54.9%。定价维持输入 \$0.75/百万 Token、输出 \$3.75/百万 Token，但复杂任务输出 Token 增加约 30%，实际成本上升约 40%。安全版 Cyber 在 CyberGym 漏洞发现基准得分 86.2%，CWE-Bench 修复测试达 47.2%。Chrome 安全团队测试显示，其正确补丁数量是竞品的 2.6 倍；谷歌云团队借助其在不到 2 小时内发现关键漏洞。该模型仅通过 Fairwind 计划向受信防御者开放。

字节跳动:

据彭博社报道，字节跳动已获得一笔 296 亿美元银团贷款，为今年亚洲规模第二大的美元贷款。公司最初寻求 200 亿美元融资额度，在获得银行强劲承诺后扩大规模，该交易尚未签署，银行仍在签署确认配额过程中。所得资金将主要用于一般企业用途及 AI 基础设施建设。公司正考虑将今年资本支出提升至高达 700 亿美元，超过去年总额两倍，用于扩建数据中心及其他 AI 基础设施。据彭博社报道，四家美国超大规模云服务商——亚马逊、Alphabet、微软、Meta——也计划今年进行高达 7250 亿美元资本支出投资，主要聚焦 AI 数据中心设备。头部 CSP 大厂融资能力加强，为后续 AI 服务器芯片投资提供较好现金流基础，国内相关产业景气度持续向上。

风险提示

1、AI 商业价值不及预期的风险：目前 AI 市场应用仍处于初级阶段，盈利模式仍需探索，市场尚未成熟。若商业模式无法持续发展新客户，需求大幅减弱，或市场接受度偏低，可能对营业收入造成较大负面影响，损害相关公司的盈利能力及产品或服务的商业价值。

2、技术发展速度不及预期的风险：目前 AI 模型的使用仍受限于诸多因素，在特定领域无法达到预期的提高生产力效果。该领域目前仍面临较大的技术挑战，包括模型训练效果不稳定、算法不成熟等问题。若技术落地不及预期，可能



影响 AI 的应用领域和运行效率，造成较大的投资损失。

3、供应链集中度过高的风险：AI 行业基础设施建设目前高度依赖某几家核心供应商，极易受到相关供应商供应短缺的影响。此外在训练方面，AI 技术依赖于大量优质数据的输入。不可靠、低质量的数据来源会一定程度上影响 AI 模型训练的性能，同时提高训练过程中的不可控成本。

4、行业监管加剧的风险：目前生成式 AI 工具仍存在法律、伦理、安全风险。AI 生成内容的产权问题仍存在较大争议。各国可能针对 AI 的使用及 AI 生成内容进行更严格的监管及抵制，影响投资预期，并阻碍 AI 技术在产业上进一步落地。公司面临法律诉讼和声誉受损等负面影响风险。

5、市场竞争加剧的风险：在如今巨头科技公司加大 AI 投入，大量创业公司涌入竞争的大环境下，技术的迅速迭代及新算法的涌现可能使得公司技术迅速落后竞争对手，影响相关公司的市场份额和投资回报的稳定性。



行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。



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。

任何形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

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。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

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：北京市东城区建国内大街 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