

大模型迭代提升性能 AI 定制芯片出货预期提升

——电子行业研究周报



申港证券
SHENGANG SECURITIES

投资摘要:

每周一谈: 大模型迭代提升性能 AI 定制芯片出货预期提升

DeepSeek V3.2 强化 Agent 能力, 大模型性能进一步提升。根据 DeepSeek 公众号, 公司近期同时发布 DeepSeek-V3.2 和 DeepSeek-V3.2-Special 两个正式版模型。在公开的推理类 Benchmark 测试中, DeepSeek-V3.2 达到了 GPT-5 的水平, 仅略低于 Gemini-3.0-Pro。DeepSeek-V3.2 模型目标为平衡推理能力与输出长度, 可同时支持思考模式与非思考模式的工具调用。模型提出了一种大规模 Agent 训练数据合成方法, 构造了大量“难解答, 易验证”的强化学习任务 (1800+ 环境, 85,000+ 复杂指令), V3.2 并没有针对这些测试集的工具进行特殊训练, V3.2 在真实应用场景中能够展现出较强的泛化性。根据量子位, DeepSeek-V3.2 最大的架构创新是引入了 DSA 高效稀疏注意力机制, DSA 让模型在长上下文任务中显著加速推理, 且无明显性能损失。模型总训练 FLOPs 较少, 在世界知识广度方面距离领先的闭源模型仍有距离, 另外, 该两个模型需要生成更长的轨迹才能达到 Gemini-3.0-Pro 的输出质量, 显示其 Token 效率仍有提升空间。目前国产大模型积极对标顶级模型, 模型效率和性能差距有望逐步缩小, 另外, 近期阿里巴巴发布 AI 眼镜、豆包发布 AI 手机助手, 端侧 AI 应用有望加速突破。建议关注端侧 AI 相关的消费电子产业链, 以及国产算力产业链的芯片、存储、服务器、先进制程产能释放等环节的景气变化。算力芯片领域建议关注寒武纪-U、海光信息、芯原股份、华为昇腾产业链。

亚马逊推出 Trainium 3, AI 定制芯片出货预期提升。根据钛媒体 AGI, 亚马逊推出新一代自研 AI 芯片 Trainium 3, 同时推出集成 144 颗 Trainium3 的 UltraServer 服务器, 新服务器的计算能力是上一代产品的 4.4 倍, 内存带宽是上一代产品的 4 倍, 能够将能效提升 40%, 同时大幅提高了产品部署能力上限, 通过服务器之间的互联互通, 最多能部署多达百万颗 Trainium 3 芯片, 将此前的天花板提高了 10 倍。公司称与同等 GPU 系统相比, Trainium 在相关场景中可以将训练和推理成本降低 50%, 在下一代芯片 Trainium 4 中使用英伟达 NVLink Fusion 技术。根据芯智讯援引 DIGITIMES 数据, 2027 年数据中心 ASIC 有望突破 1000 万颗, 达到接近同期的 GPU 的 1177 万颗的销量。

根据量子位援引消息称, 谷歌开始向 Meta 与大型金融机构等, 推介在其自有数据中心部署 TPU 的方案。Meta 正在与谷歌探讨使用 TPU 训练新 AI 模型, 计划在 2027 年斥资数十亿美元在其数据中心使用 TPU, 并于明年从谷歌云租用谷歌芯片。谷歌云部门部分高管预测, 这类业务可为谷歌带来数十亿美元的年收入。建议关注谷歌产业链光模块、PCB 等龙头企业, 以及有望受益定制 AI 芯片行业崛起的芯原股份、翱捷科技-U、灿芯股份等。

投资策略: 建议关注消费电子尤其是 AI 端侧产品年底备货出货情况, 关注阿里巴巴夸克 AI 眼镜的出货量和供应链企业的受益机会。建议关注谷歌 TPU 部署及对数据中心算力芯片市场的拓展, 关注其产业链中有望受益的光模块、PCB 等龙头企业, 以及有望受益定制 AI 芯片行业崛起的芯原股份、翱捷科技、灿芯股份等。

风险提示: 贸易摩擦加剧, 需求复苏不及预期, 产能扩张不及预期, 竞争加剧

评级

增持 (维持)

2025 年 12 月 08 日

王伟

分析师

SAC 执业证书编号: S1660524100001

行业基本资料

股票家数	471
行业平均市盈率	64.9
市场平均市盈率	14.1

行业表现走势图



资料来源: 申港证券研究所

相关报告

- 1、《电子行业研究周报: 阿里 AI 产品加速推进 谷歌 TPU 部署受关注》2025-12-04
- 2、《电子行业研究周报: 中芯国际毛利率及产能指标修复 存储资本支出或对位元出货支撑有限》2025-11-19
- 3、《电子行业研究周报: 北美 CSP 资本开支旺盛 英伟达 GTC 大会展望超预期》2025-11-06

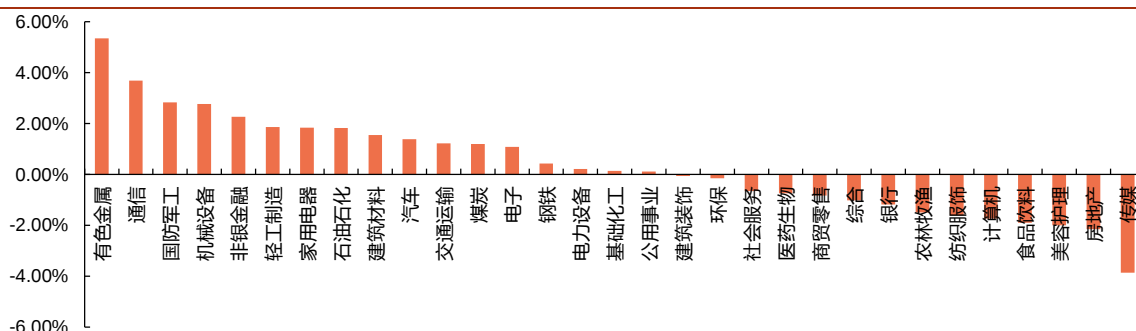
1. 市场回顾

上周（12.1-12.5）申万电子行业指数上涨 1.09%，在申万 31 个行业中排名第 13，跑输沪深 300 指数 0.19%。

本月（12.1-12.5）申万电子行业指数上涨 1.09%，在申万 31 个行业中排名第 13，跑输沪深 300 指数 0.19%。

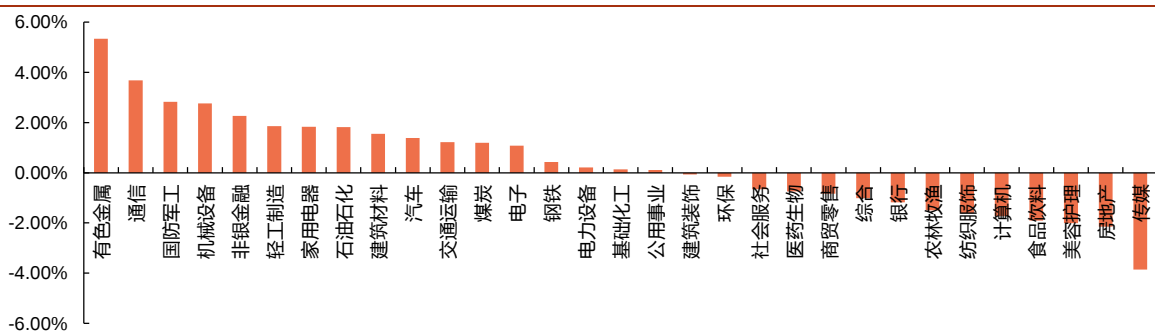
年初至今（1.1-12.5）申万电子行业指数上涨 42.15%，在申万 31 个行业中排名第 3，跑赢沪深 300 指数 25.65%。

图1：申万一级行业上周涨跌幅



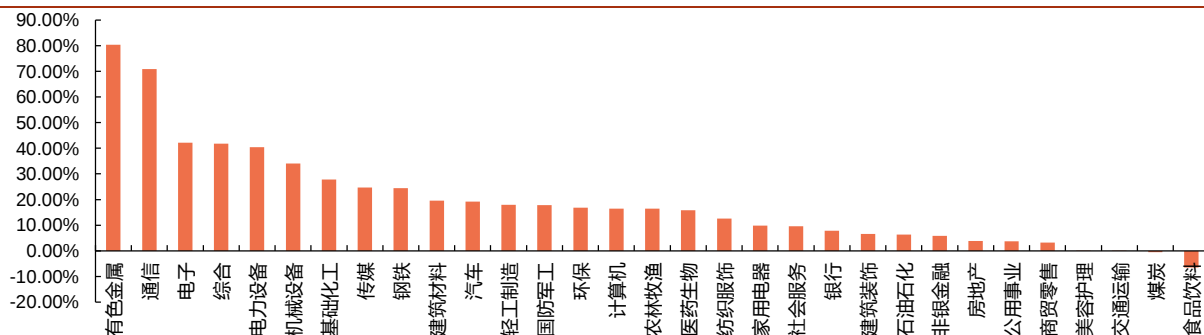
资料来源：wind，申港证券研究所

图2：申万一级行业本月以来涨跌幅



资料来源：wind，申港证券研究所

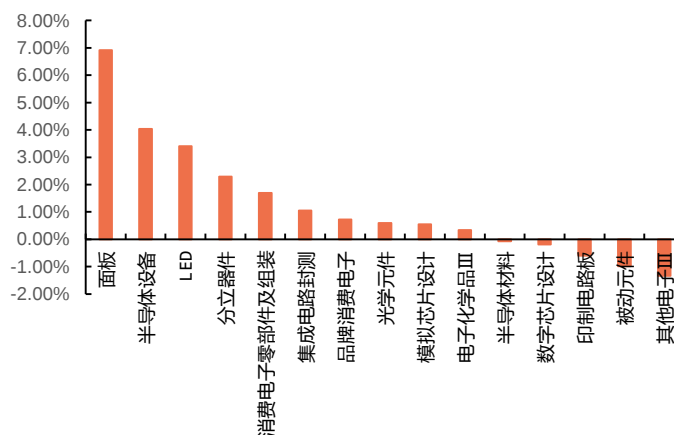
图3：申万一级行业年初至今涨跌幅



资料来源：wind，申港证券研究所

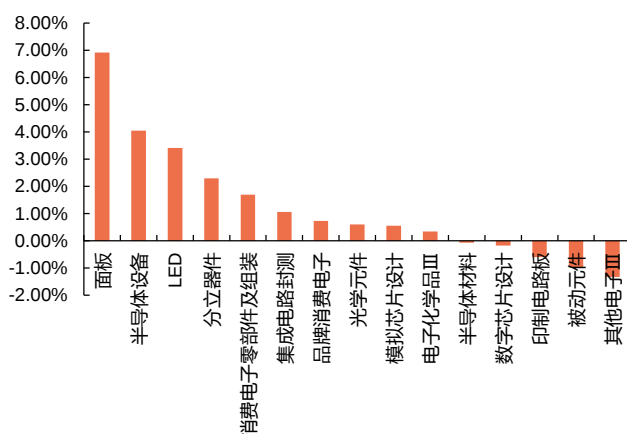
上周(12.1-12.5)申万电子行业三级子行业中面板、半导体设备、LED、分立器件、消费电子零部件及组装指数涨跌表现相对靠前,跑赢沪深300指数5.64%、2.76%、2.13%、1.01%、0.41%。

图4: 电子子行业上周涨跌幅



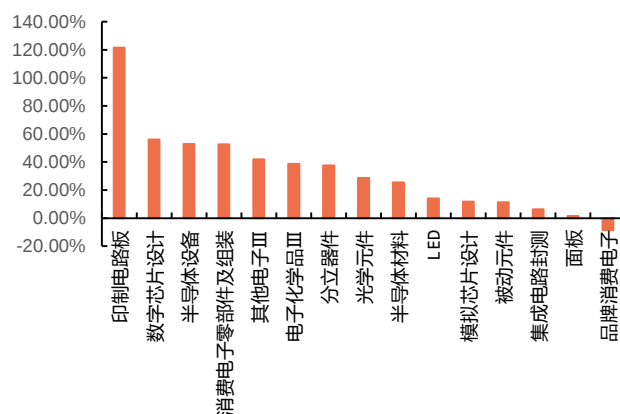
资料来源: wind, 申港证券研究所

图5: 电子子行业本月涨跌幅



资料来源: wind, 申港证券研究所

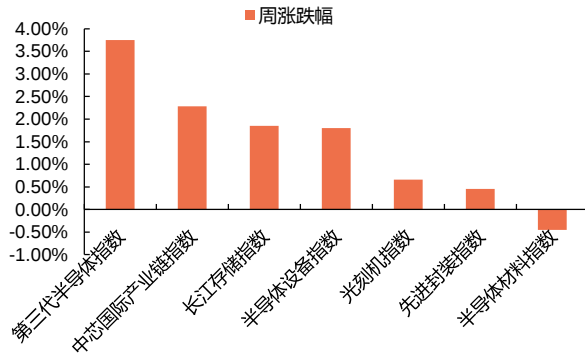
图6: 电子子行业年初至今涨跌幅



资料来源: wind, 申港证券研究所

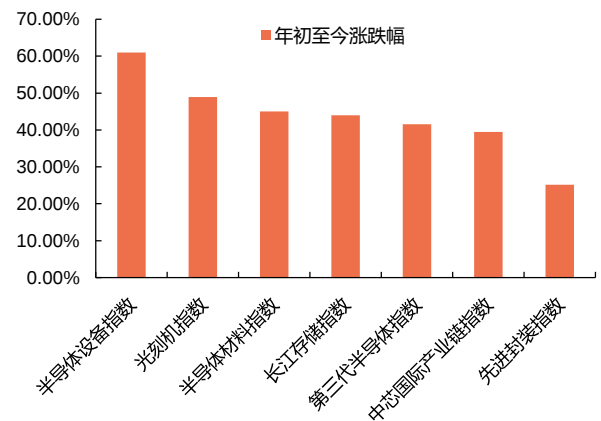
上周(12.1-12.5)万得半导体概念指数中第三代半导体指数、中芯国际产业链指数、长江存储指数、半导体设备指数、光刻机指数、先进封装指数、半导体材料指数涨跌幅分别为3.75%、2.28%、1.85%、1.81%、0.66%、0.45%、-0.45%。

图7：万得半导体概念指数上周涨跌幅



资料来源：wind，申港证券研究所

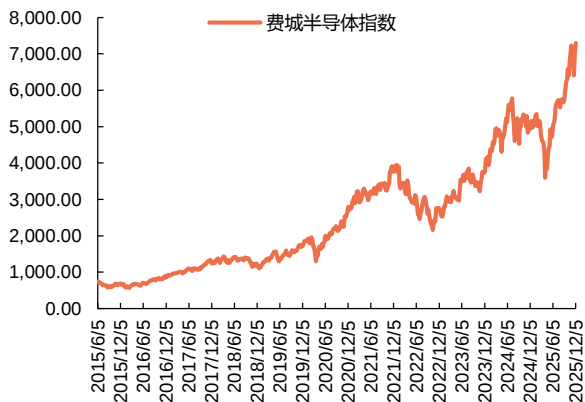
图8：万得半导体概念指数年初至今涨跌幅



资料来源：wind，申港证券研究所

截至12月5日，费城半导体指数收于7294.84点、周上涨3.84%。台湾半导体指数收于864.40点、周上涨1.34%。

图9：费城半导体指数



资料来源：wind，申港证券研究所

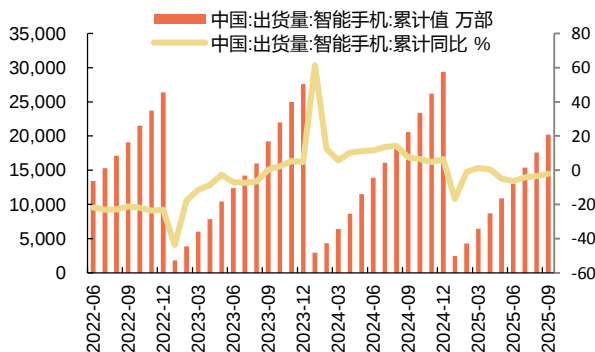
图10：台湾半导体指数



资料来源：wind，申港证券研究所

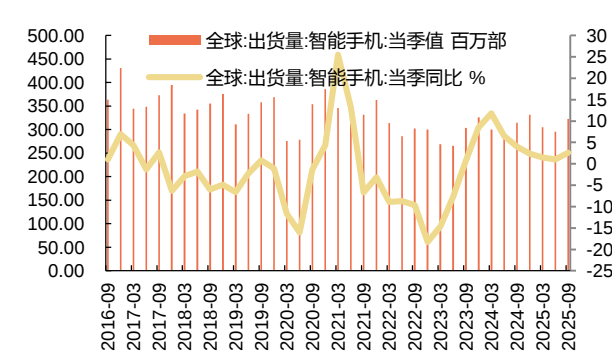
2. 行业数据跟踪

图11：中国智能手机出货量及累计同比



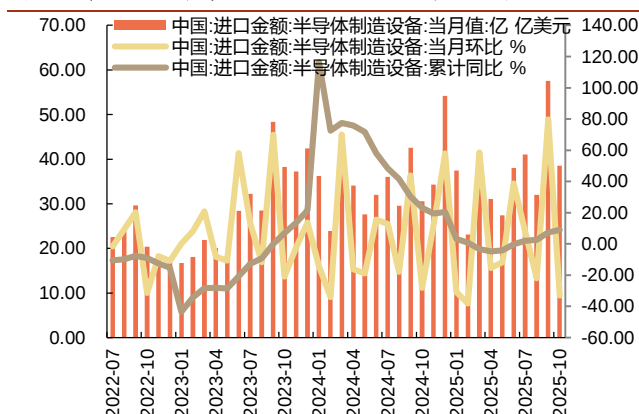
资料来源：wind，申港证券研究所

图12：全球智能手机出货量及当季同比



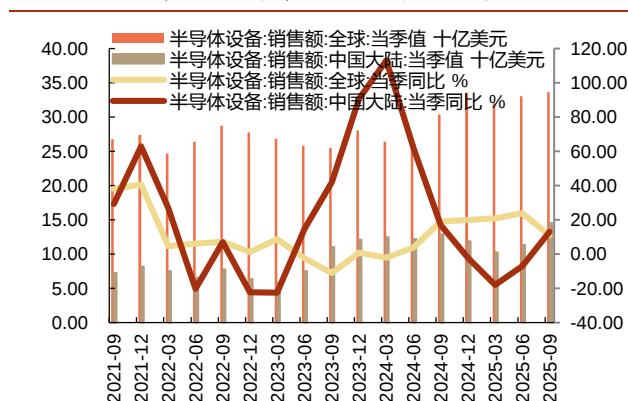
资料来源：wind，申港证券研究所

图13: 中国进口半导体制造设备金额及同环比情况



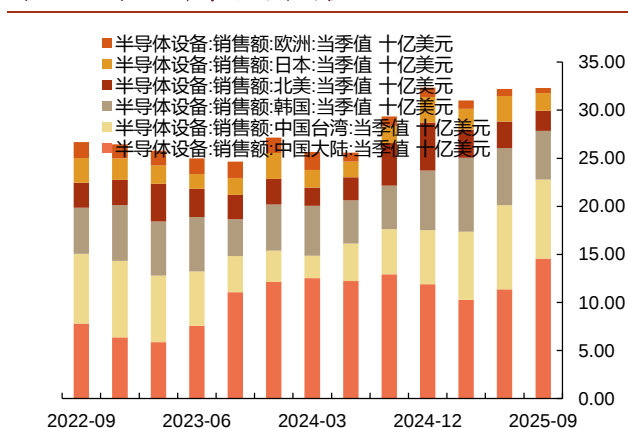
资料来源: wind, 申港证券研究所

图14: 全球及中国大陆半导体设备销售额及同比



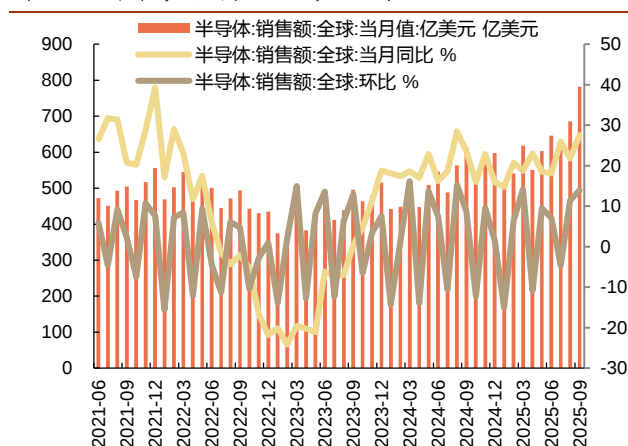
资料来源: ifind, 申港证券研究所

图15: 主要地区半导体设备销售额



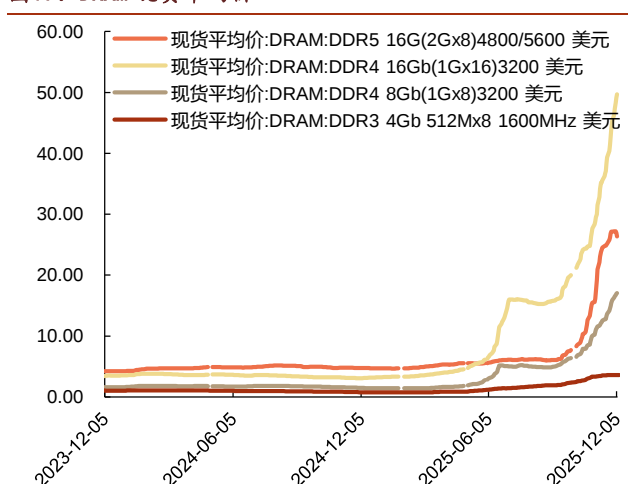
资料来源: ifind, 申港证券研究所

图16: 全球半导体销售额及同环比情况



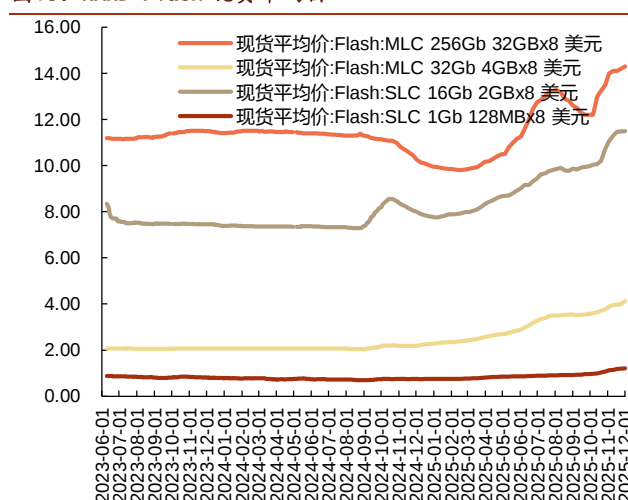
资料来源: ifind, 申港证券研究所

图17: DRAM 现货平均价



资料来源: ifind, 申港证券研究所

图18: NAND Flash 现货平均价



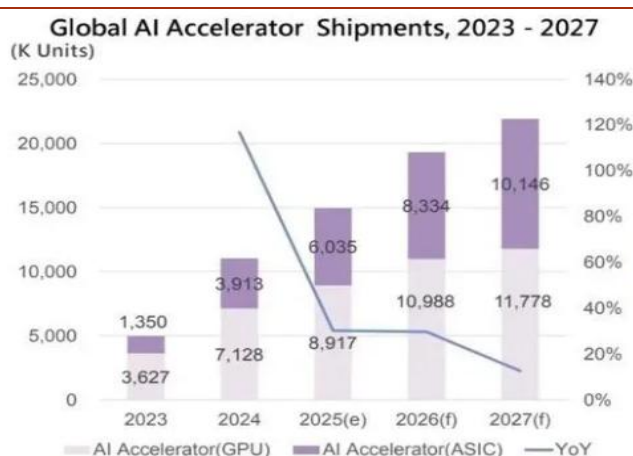
资料来源: ifind, 申港证券研究所

3. 每周一谈：大模型迭代提升性能 AI 定制芯片出货预期提升

DeepSeek V3.2 强化 Agent 能力，大模型性能进一步提升。根据 DeepSeek 公众号，公司近期同时发布 DeepSeek-V3.2 和 DeepSeek-V3.2-Special 两个正式版模型。在公开的推理类 Benchmark 测试中，DeepSeek-V3.2 达到了 GPT-5 的水平，仅略低于 Gemini-3.0-Pro。DeepSeek-V3.2 模型目标为平衡推理能力与输出长度，可同时支持思考模式与非思考模式的工具调用。模型提出了一种大规模 Agent 训练数据合成方法，构造了大量“难解答，易验证”的强化学习任务（1800+ 环境，85,000+ 复杂指令），V3.2 并没有针对这些测试集的工具进行特殊训练，V3.2 在真实应用场景中能够展现出较强的泛化性。根据量子位，DeepSeek-V3.2 最大的架构创新是引入了 DSA 高效稀疏注意力机制，DSA 让模型在长上下文任务中显著加速推理，且无明显性能损失。模型总训练 FLOPs 较少，在世界知识广度方面距离领先的闭源模型仍有距离，另外，这两个模型需要生成更长的轨迹才能达到 Gemini-3.0-Pro 的输出质量，显示其 Token 效率仍有提升空间。目前国产大模型积极对标顶级模型，模型效率和性能差距有望逐步缩小，另外，近期阿里巴巴发布 AI 眼镜、豆包发布 AI 手机助手，端侧 AI 应用有望加速突破。建议关注端侧 AI 相关的消费电子产业链，以及国产算力产业链的芯片、存储、服务器、先进制程产能释放等环节的景气变化。算力芯片领域建议关注寒武纪-U、海光信息、芯原股份、华为昇腾产业链。

亚马逊推出 Trainium 3，AI 定制芯片出货预期提升。根据钛媒体 AGI，亚马逊推出新一代自研 AI 芯片 Trainium 3，同时推出集成 144 颗 Trainium3 的 UltraServer 服务器，新服务器的计算能力是上一代产品的 4.4 倍，内存带宽是上一代产品的 4 倍，能够将能效提升 40%，同时大幅提高了产品部署能力上限，通过服务器之间的互联互通，最多能部署多达百万颗 Trainium 3 芯片，将此前的天花板提高了 10 倍。公司称与同等 GPU 系统相比，Trainium 在相关场景中可以将训练和推理成本降低 50%，在下一代芯片 Trainium 4 中使用英伟达 NVLink Fusion 技术。根据芯智讯援引 DIGITIMES 数据，2027 年数据中心 ASIC 有望突破 1000 万颗，达到接近同期的 GPU 的 1177 万颗的销量。

图19：2027 年数据中心 ASIC 芯片出货量有望超过 1000 万颗



资料来源：芯智讯，DIGITIMES，申港证券研究所

根据量子位援引消息称，谷歌开始向 Meta 与大型金融机构等，推介在其自有数据中心部署 TPU 的方案。Meta 正在与谷歌探讨使用 TPU 训练新 AI 模型，计划在 2027 年斥资数十亿美元在其数据中心使用 TPU，并于明年从谷歌云租用谷歌芯片。谷歌云部门部分高管预测，这类业务可为谷歌带来数十亿美元的年收入。建议关注谷歌产业链光模块、PCB 等龙头企业，以及有望受益定制 AI 芯片行业崛起的芯原股份、翱捷科技-U、灿芯股份等。

4. 重要公告

江波龙：12 月 3 日发布向特定对象发行 A 股股票预案。公司拟向特定对象发行股票拟募集资金总额不超过 37 亿元（含本数），扣除发行费用后的募集资金净额将用于面向 AI 领域的高端存储器研发及产业化项目、半导体存储主控芯片系列研发项目、半导体存储高端封测建设项目、补充流动资金等项目。

拓荆科技：12 月 5 日发布向特定对象发行 A 股股票募集说明书（申报稿）。本次向特定对象发行股票募集资金总额不超过人民币 46 亿元（含本数），扣除发行费用后的募集资金净额拟投入高端半导体设备产业化基地建设项目、前沿技术研发中心建设项目、补充流动资金。

佰维存储：12 月 2 日发布关于择机出售公司持有的部分股票资产的公告。公司拟根据股票市场行情以集中竞价交易的方式择机出售所持联芸科技（杭州）股份有限公司部分或全部 A 股股票，合计不超过 3744681 股（占其总股本的比例不超过 1%）

5. 行业动态

美光科技退出消费内存，聚焦 AI 存储

当地时间 12 月 3 日，美国存储原厂美光科技（Micron Technology）做出了一项重要决策：退出面向消费者的 Crucial 内存业务，将全部产能与资源聚焦于高性能 AI 芯片所需的高带宽内存（HBM）和数据中心存储产品。（来源：电子工程专辑）

2025Q3 全球半导体设备市场：中国大陆位居第一，占比高达 43%

12 月 3 日，国际半导体产业协会（SEMI）最新公布的统计数据显示，2025 年三季度全球半导体制造设备销售额同比增长 11%至 336.6 亿美元，为连续第 6 个季度呈现增长，增幅连续第 5 个季度达 2 位数百分比（10%以上）水准，环比销售额连续第 5 个季度高于 300 亿美元，超越 2024 年四季度的 335.6 亿美元，创有数据可供比较的 2005 年以来历史新高纪录。

从区域表现来看，三季度中国大陆市场销售额同比增长 13%至 145.6 亿美元，连续第 10 个季度成为全球最大半导体设备市场，占整体销售额比重为 43%，这也是近 4 个季度来首度超越 40%；中国台湾市场销售额暴涨 75%至 82.1 亿美元，增幅居所有市场之首，且连续第两个季度超越韩国，成为全球第 2 大半导体设备市场；韩国市场销售额同比 12%至 50.7 亿美元，位居第三；北美市场销售额增暴跌 52%至 21.1

亿美元，排名第四；日本市场销售额同比增加 5%至 18.3 亿美元，排名第五；欧洲市场销售也同比大跌 50%至 5.2 亿美元，排名第六；其他市场销售额同比增长 34%至 13.6 亿美元。（来源：芯智讯）

英伟达 20 亿美元入股 Synopsys，并达成战略合作

当地时间 12 月 1 日，英伟达（NVIDIA）与 EDA 大厂新思科技（Synopsys）公司宣布建立一项扩展的战略合作伙伴关系，旨在革新跨行业的设计与工程。同时，英伟达以每股 414.79 美元的价格投资了 20 亿美元的 Synopsys 普通股。

英伟达创始人兼首席执行官黄仁勋表示：“CUDA GPU 加速计算正在革新设计——实现前所未有的速度和规模仿真，从原子到晶体管，从芯片到完整系统，打造计算机内部功能齐全的数字孪生。”“我们与 Synopsys 的合作利用 NVIDIA 加速计算和人工智能的力量，重新构想工程与设计——赋能工程师发明将塑造我们未来的非凡产品。”（来源：芯智讯）

2026 年全球半导体营收将增长 26.3%，逼近 1 万亿美元大关

12 月 2 日消息，根据世界半导体贸易统计组织（WSTS）最新发布的全球半导体市场预测报告显示，2025 年全球半导体营收将同比增长 22.5%至 7720 亿美元，2026 年将再度增长 26.3%至 9750 亿美元，逼近 1 万亿美元大关。

WSTS 表示，2025 年全球半导体营收的增长主要受益于人工智能应用及数据中心基础设施的强劲需求，推动了对于逻辑芯片和存储芯片需求的增长。

其中，2025 年逻辑芯片营收有望同比增长 37.1%，是增幅最大的产品类别；其次是存储芯片，营收同比增长 27.8%；传感器营收将同比增长 10.4%；微处理器营收同比增长 7.9%；模拟芯片营收同比增长 7.5%；光电子元件营收同比增长 3.7%；受汽车领域需求疲软影响，分立元件营收同比下滑 0.4%。

展望 2026 年，WSTS 预期，全球半导体营收有望再增长 26.3%，达到 9,750 亿美元，存储芯片和逻辑芯片仍是主要成长动能，将分别成长 39.4%及 32.1%。美洲和亚太地区是同比增长最强劲的地区，分别增长 34.4%及 24.9%。（来源：芯智讯）

谷歌 TPU 若对外销售，2027 年有望拿下全球 20%的 AI 芯片市场

12 月 5 日消息，近期谷歌（Google）推出的基于其自研 TPU 芯片的 Gemini 3 获得了众多用户的追捧。而在今年 10 月底，谷歌宣布将向 Anthropic PBC 供应数百亿美元的 TPU 芯片，随后有报道称 Meta 计划斥资数十亿美元获取谷歌 TPU 的使用权，这也掀起了云端服务供应商（CSP）对导入谷歌 TPU 或自家研发的 AI 芯片的信心，虽然目前仍无法撼动英伟达（NVIDIA）在 AI 芯片市场的霸主地位，但可以在一定程度上降低对英伟达的依赖，并降低采购成本。

根据 DA Davidson 的技术研究主管 Gil Luria 的说法，若 CPS 公司希望摆脱对英伟达的依赖，谷歌 TPU 将是一个不错的选择。他预测，如果谷歌认真考虑对外销售 TPU，未来几年内可能会占据 20%的 AI 芯片市场，这将使该业务总价值将达到约 9,000 亿美元。（来源：芯智讯）

黑芝麻智能拟 4-5.5 亿元收购亿智电子控股权

近日，黑芝麻智能国际控股有限公司（2533.HK）发布内幕消息公告，更新此前披露的可能收购事项。根据公告，公司拟通过收购股权及向目标公司注资的方式，收购珠海亿智电子科技有限公司的控股权，预期总代价约介乎人民币 4 亿元至人民币 5.5 亿元。亿智电子则以边缘侧和端侧 AI 芯片为核心业务，围绕 AI 机器视觉算法和 SoC 芯片设计提供系统级解决方案，产品覆盖智能车载、智能硬件及智慧安防等应用场景。其自研的多系列视觉 AI 芯片已实现规模出货，并在部分车规级项目上完成流片和定点导入。

黑芝麻智能在车规级算力平台、亿智电子在端侧视觉 AI 与边缘应用方面的技术方向具有互补性。若交易落地，黑芝麻智能有望在保持汽车主业优势的基础上，进一步向安防、消费级智能硬件等领域延伸，形成“车规+端侧”多场景覆盖的产品与方案组合。黑芝麻智能表示，目前公司及其他相关方正积极推进可能收购事项及相关工作，预计就可能收购事项签署最终交易协议将于 2026 年第一季度内完成。（来源：集微网）

亚马逊最新 AI 芯片 Trainium4 兼容英伟达 GPU

亚马逊云科技（AWS）在 AWS re:Invent 全球大会上，正式推出第三代 AI 训练芯片 Trainium3，并同步披露下一代 Trainium4 研发进展。这款基于台积电 3nm 工艺打造的芯片，帮助企业更快地训练更大的人工智能模型，旨在冲击英伟达、谷歌等头部玩家的 AI 芯片市场格局。据亚马逊云科技（AWS）披露，Trn3 UltraServers 将多达 144 个 Trainium3 芯片集成到单个系统中，计算性能比 Trainium2 UltraServers 提升高达 4.4 倍。此次发布的一大亮点则是亚马逊对 AI 战略的重新定位。下一代 Trainium4 将首次集成英伟达 NVLink Fusion 技术，实现与 GPU 的协同工作。

亚马逊表示，“为了实现更强大的扩展性能，Trainium4 专为支持 NVIDIA NVLink Fusion 高速芯片互连技术而设计。这种集成将使 Trainium4、Graviton 和 Elastic Fabric Adapter (EFA) 能够在通用的 MGX 机架中无缝协作，为您提供经济高效的机架级 AI 基础架构，同时支持 GPU 和 Trainium 服务器。最终打造一个灵活的高性能平台，专为高要求的 AI 模型训练和推理工作负载而优化。”（来源：电子工程专辑）

追求利润最大化，三星拟将部分 HBM3E 产能转产通用 DRAM

据韩媒报道，三星电子正考虑大幅削减其 10nm 制程的第四代（1a）DRAM 芯片 HBM3E 的产能。鉴于其明年将“利润最大化”作为整体目标，三星计划将生产重心转移到利润率高于 HBM3E 的通用 DRAM 芯片上。

据业内人士透露，三星内部正在讨论将 30-40% 的 1a DRAM 产能转换为 10nm 级第五代（1b）DRAM。如果将成熟工艺线（例如 1z）的转换投资也算进去，三星将确保每月额外获得 8 万片（基于晶圆）的 1b DRAM 产能。

报道称，三星电子 12 层 HBM3E 产品的营业利润率预计在 30% 左右，良率也被认为已显著提升。然而，由于 DDR5 等通用 DRAM 的营业利润率预计将超过 60%，因此扩大

HBM3E 产能的必要性不大。

知情人士称，三星电子明年将“利润最大化”作为整体目标，期待实现更高的营业利润，目前的目标是保持 DRAM 产量的高速增长，并计划将大部分产能分配给利润率较高的通用 DRAM。（来源：闪存市场）

6. 风险提示

贸易摩擦加剧，需求复苏不及预期，产能扩张不及预期，竞争加剧

分析师承诺

负责本研究报告全部或部分内容的每一位证券分析师，在此申明，本报告的观点、逻辑和论据均为分析师本人独立研究成果，引用的相关信息和文字均已注明出处，不受任何第三方的影响和授意。本报告依据公开的信息来源，力求清晰、准确地反映分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与、未来也将不会与本报告中的具体推荐或观点直接或间接相关。

风险提示

本证券研究报告所载的信息、观点、结论等内容仅供投资者决策参考。在任何情况下，本公司证券研究报告均不构成对任何机构和个人的投资建议，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。市场有风险，投资者在决定投资前，务必要审慎。投资者应自主作出投资决策，自行承担投资风险。

免责声明

申港证券股份有限公司（简称“本公司”）是具有合法证券投资咨询业务资格的机构。

本报告所载资料的来源被认为是可靠的，但本公司不保证其准确性和完整性，也不保证所包含的信息和建议不会发生任何变更。

申港证券研究所已力求报告内容的客观、公正，但报告中的观点、结论和建议仅供参考，不构成所述证券的买卖出价或征价，不构成其他投资标的的要约和邀请，亦不构成任何合约或承诺的基础。投资者不应单纯依靠本报告而取代自身独立判断，应自主作出投资决策并自行承担投资风险，投资者据此做出的任何投资决策与本公司和作者无关。本公司并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相关的其他任何损失承担任何责任。

本报告所载资料、意见及推测仅反映申港证券研究所于发布本报告当日的判断，本报告所指证券或投资标的的价格、价值及投资收入可能会产生波动，在不同时期，申港证券研究所可能会对相关的分析意见及推测做出更改。本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。

本报告仅面向申港证券客户中的专业投资者及风险承受能力为 C3、C4、C5 的普通投资者，本公司不会因接收人收到本报告而视其为当然客户。本报告版权归本公司所有，未经事先许可，任何机构和个人不得以任何形式翻版、复制和发布。如转载或引用，需注明出处为申港证券研究所，且不得对本报告进行有悖原意的引用、删节和修改，否则由此造成的一切不良后果及法律责任由私自翻版、复制、发布、转载和引用者承担。

投资评级说明

申港证券行业评级说明：增持、中性、减持

增持	报告日后的 6 个月内，相对强于市场基准指数收益率 5% 以上
中性	报告日后的 6 个月内，相对于市场基准指数收益率介于-5%~+5%之间
减持	报告日后的 6 个月内，相对弱于市场基准指数收益率 5% 以上

（基准指数说明：A 股市场基准为沪深 300 指数；香港市场基准为恒生指数；美国市场基准为标普 500 指数或纳斯达克指数。）

申港证券公司评级说明：买入、增持、中性、减持

买入	报告日后的 6 个月内，相对强于市场基准指数收益率 15% 以上
增持	报告日后的 6 个月内，相对强于市场基准指数收益率 5%~15%之间
中性	报告日后的 6 个月内，相对于市场基准指数收益率介于-5%~+5%之间
减持	报告日后的 6 个月内，相对弱于市场基准指数收益率 5% 以上

（基准指数说明：A 股市场基准为沪深 300 指数；香港市场基准为恒生指数；美国市场基准为标普 500 指数或纳斯达克指数。）