

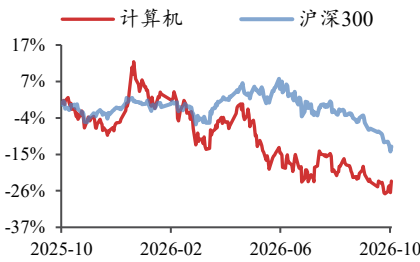
计算机

2026 年 10 月 25 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

《物理 AI 或成编码外最大落地场景，全球巨头重点发力——行业点评报告》
-2026.10.20

《服务器上游系列（一）：先进封装破局后摩尔时代，玻璃基板进入量产前夕——行业周报》-2026.10.18

《Anthropic 三季度营收增速放缓，关注 AI 应用投资机会——行业点评报告》
-2026.10.17

国产算力系列：单芯片性能与集群规模同步跃升，在手订单全面提速

张越（分析师）

zhangyue1@kysec.cn

证书编号：S0790524090003

叶进新（联系人）

yejinxin@kysec.cn

证书编号：S0790126060053

● 国产 GPU 性能迭代与订单放量共振，商业化进程全面提速

GPU 产业链呈“上游卖铲人—中游设计者—下游算力消耗方”三级传导结构，四类厂商梯队成型：独立专业厂商走向持续交付，华为昇腾 384 超节点已部署 750 余套、综合成熟度保持领先，百度昆仑芯、阿里平头哥依托云业务规模应用，景嘉微拓展行业计算场景，海光与寒武纪进入规模商用前列。IDC 数据显示，2025 年中国 AI 加速卡出货中国产厂商份额达 41%，2026 年上半年英伟达份额降至 34%、华为昇腾以 28% 紧随其后。产品端，国产算力进入年度迭代节奏，性能沿低精度算力与高带宽显存两条主线跃升，MTT S5000 单卡 Prefill 吞吐超 4,000 tokens/s，部分主流产品亦在大模型推理场景中展现出竞争力。财务端，2025 年摩尔线程、沐曦股份、壁仞科技、寒武纪收入分别同比增长 243.84%、121.27%、207.19%、453.41%，2026 年上半年多家公司收入超上年全年；腾讯、阿里、字节资本开支集体上修，训推分线格局下国产卡成为重要选择。

● AI 产业动态密集更新，国产推理生态与海外算力投入并进

本周 AI 产业围绕模型能力、推理生态与算力基建持续迭代。模型端，蚂蚁数科发布 LingDT-Decision-1.0，token 消耗较传统大模型降低超 90%、响应速度提升约 10 倍；科大讯飞发布基于全国产算力训练的星火 X2.6，端侧模型下载量超 256 万次；OpenAI 前沿模型及 Codex 接入谷歌云，并披露离线训练体系。算力端，DeepSeek 开源推理框架 DeepInfer，兼容英伟达与昇腾平台；英伟达推出 DSX POD 模块化 AI 工厂，部署周期有望由约六个月缩短至 14 天；Anthropic 招股书草案披露未来 7 至 10 年算力采购合同合计约 4500 亿美元，Meta 多年支出承诺增至约 8250 亿美元。模型优化、推理生态与算力投入共同推进，有望加快 AI 能力向规模化产业应用转化。

● 计算机板块领涨一级行业，AI 应用相对表现占优

10 月 19 日至 10 月 23 日，计算机指数上涨 5.21%，同期沪深 300、上证指数、创业板指及科创 50 分别变动-1.57%、-0.69%、-3.18%和-0.97%，计算机板块在 31 个一级行业中排名第 1、相对表现突出。AI 细分方向中，AI 应用指数累计上涨约 4.55%，AI 算力指数上涨约 1.26%，应用方向取得约 3.29 个百分点的相对收益。个股层面，沐曦股份-U 上涨 17.38%，天娱数科、优刻得-W 分别上涨 11.37% 和 10.95%；海天瑞声上涨 79.98%，居 AI 应用板块涨幅首位。产业创新持续积累，可关注 AI 应用落地与国产算力放量带来的细分领域机会。

● 受益标的

国产算力：寒武纪、海光信息、禾盛新材、沐曦股份、浪潮信息、工业富联等。

● 风险提示：大模型技术迭代不及预期风险、AI 技术应用不及预期风险、相关政策不及预期风险、AI Infra 建设不及预期风险等。

目 录

1、性能迭代与订单放量共振，国产 GPU 商业化进程全面提速	4
1.1、四类厂商梯队成型，产业链协同支撑国产份额跃升	4
1.2、单芯片性能与集群规模同步跃升，部分主流产品推理性能优异	8
1.3、收入放量与主业集中共振，资本开支上修开启采购释放期	10
2、计算机产业动态更新	15
2.1、国内 CSP 厂商与大模型厂商最新动向	15
2.2、海外 CSP 厂商与大模型厂商最新动向	17
2.3、国产算力产业最新进展	18
2.4、海外算力产业最新进展	19
3、上周市场表现回顾（2026/10/19-2026/10/23）	20
4、风险提示	22

图表目录

图 1：摩尔线程智算集群处于实际业务运行状态	4
图 2：景嘉微与麒麟信安协同拓展国产 GPU 云桌面应用	5
图 3：海光 DTK 软件栈完善异构计算开发与运行配套	6
图 4：寒武纪 NeuWare 贯通芯片开发与行业应用	6
图 5：GPU 产业链贯穿算法、算力、应用三层，算力层自半导体至服务逐级传导	7
图 6：GPU 市场规模持续扩张，2026 年预计达 1414 亿美元	7
图 7：中国 GPU 市场规模快速增长，2030 年预计达 5420 亿元	7
图 8：摩尔线程 MTT S5000 单卡 Prefill 吞吐 2.5 倍于国际旗舰	10
图 9：营业收入快速增长，净亏损持续收窄	11
图 10：2026 年上半年云端产品贡献主要收入	11
图 11：营业收入高速增长，2026 年上半年达上年全年约八成	11
图 12：GPU 产品及配件贡献约 99%收入	11
图 13：收入增长持续提速，2026 年上半年同比增速 191.64%	12
图 14：GPU 业务占比升至约 97%，收入向主业集中	12
图 15：收入规模快速扩张，2026 年上半年超上年全年	12
图 16：智能计算解决方案贡献约 95%收入	12
图 17：2026 年上半年同比增速 278.38%	13
图 18：AI 加速卡及模组贡献约八成收入	13
图 19：收入恢复性增长，2026 年上半年同比增速 73.06%	13
图 20：图形显控业务占比提升至约七成	13
图 21：收入规模持续扩张，增速逐年提升	14
图 22：处理器业务贡献超 99%收入	14
图 23：收入高速增长，2025 年实现首次年度盈利	14
图 24：云端产品线占比超 99%，收入高度聚焦	14
图 25：腾讯、阿里、字节 2026 年资本开支集体上修	15
图 26：在本周多数一级行业下跌的背景下，计算机板块相对表现突出	20
图 27：本周 AI 细分方向呈现一定分化，AI 应用指数整体表现强于算力指数	21
图 28：AI 算力指数整体小幅上涨，沐曦股份-U 等个股表现突出	21
图 29：AI 应用方向指数整体上涨，海天瑞声涨幅居前	22

表 1： 国产算力芯片已进入年度迭代节奏	8
----------------------------	---

1、性能迭代与订单放量共振，国产 GPU 商业化进程全面提速

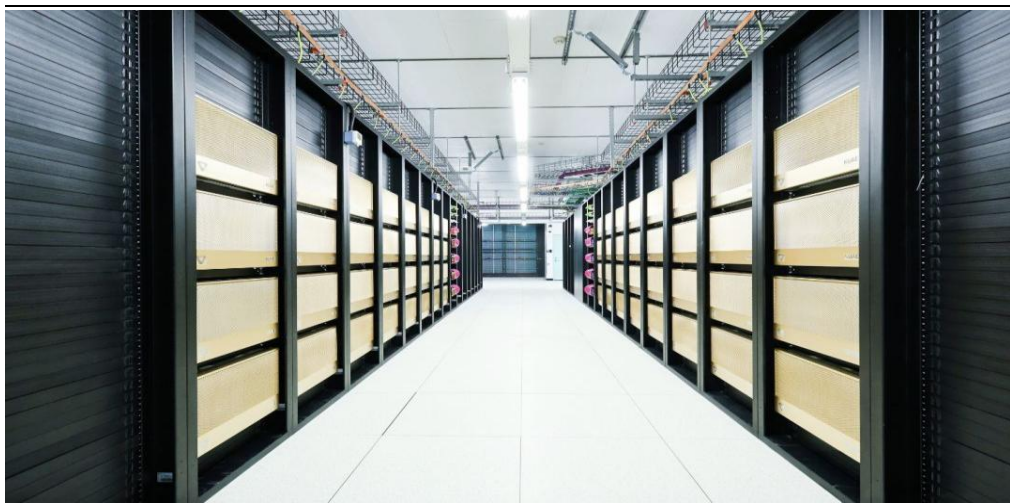
1.1、四类厂商梯队成型，产业链协同支撑国产份额跃升

国产算力规模商用持续深化，系统配套、集群部署与客户应用构成头部厂商的竞争优势。华为昇腾依托完整软硬件体系及广泛行业覆盖，在综合成熟度方面保持领先；海光信息与寒武纪分别凭借服务器平台生态和云端产品积累形成规模商业基础，昆仑芯和平头哥则通过万卡集群交付及云服务部署拓展应用。摩尔线程、沐曦、壁仞、天数智芯和燧原加快交付放量，集群落地与客户拓展同步推进；景嘉微立足成熟图形显控业务，持续拓宽行业计算场景。

（一）独立专业厂商由产品突破走向持续交付，集群落地与客户拓展推动商业应用深化

独立专业厂商已形成商业供给，产品迭代与集群运行推动成熟度从单卡交付向系统应用提升。摩尔线程、沐曦等依托具有国际芯片产业经验的团队起步，逐步建立芯片、软件及计算系统配套。摩尔线程 MTT S5000 智算集群已形成规模销售，并用于大模型、具身智能及世界模型训练，公司 2026 年 10 月进一步披露无锡相关智算中心一期满载运行，表明产品已进入持续业务负载。沐曦则以 C500 多地智算中心部署及客户持续采购为基础，推动 C600 于 2026 年 5 月量产，形成既有产品应用与新一代产品供给的衔接。两者均已跨过单纯研发验证，分别从集群运行和产品迭代体现商业化深化。

图1：摩尔线程智算集群处于实际业务运行状态



资料来源：摩尔线程官网

客户结构逐步拓宽，推动国产算力由少数项目导入走向跨行业应用。壁仞科技产品已在多个智算集群规模部署，服务对象覆盖大模型开发商、国家级 AI 算力平台、互联网企业及电信运营商，体现其面向数据中心的工程交付能力。天数智芯则依托较早的量产积累，形成训练与推理双线商业应用，截至 2026 年 6 月末累计服务客户超过 380 家，覆盖互联网、金融、医疗及科研等领域。壁仞依托智算集群拓展行业应用，天数智芯凭借训推双线布局扩大客户覆盖，商业化由单点验证向多场景应用深化。

头部客户的深度合作与外部市场拓展共同推动燧原商业化深化。腾讯作为重要股东及核心客户，为燧原科技提供资本与业务场景支持，公司产品已在腾讯多类 AI

业务中规模应用，持续的联合研发与场景适配带动软件生态和系统交付能力完善。在此基础上，燧原积极拓展其他头部互联网、电信运营商及行业客户，成都高新电子智算集群项目已完成最终验收并进入运维阶段，商业化布局由核心客户深耕向更多客户及行业场景延伸。

（二）集团体系以业务场景支撑规模应用，华为昇腾综合领先、百度与阿里深化云端供给

集团自研算力依托内部业务完成应用验证，并通过云平台及产业伙伴扩大商业覆盖，系统配套构成这一类厂商的共同优势。其中，华为昇腾已形成较广泛的跨行业部署，领先优势主要体现于芯片、互连、基础软件、开发工具和行业方案的完整布局。华为 2026 年披露，昇腾 384 超节点已部署 750 余套，覆盖互联网、运营商、金融、教育和医疗等 20 多个行业，生态服务 2000 多家政企核心客户。较大的部署范围和长期系统运行积累，使其能够支撑从模型适配到业务上线的完整流程，综合成熟度处于国内领先地位。

百度和阿里以模型与云业务承接自研算力，昆仑芯和平头哥均已进入大规模系统应用阶段。昆仑芯源于百度芯片研发体系，2026 年百度开发者大会披露，2025 年以来已交付多个 P800 万卡集群，商业能力由计算卡供给延伸至大规模集群运行；平头哥则通过芯片、模型与阿里云联合优化推进商用，阿里 2026 年 9 月披露 M890 超节点已支撑超大参数模型推理，并在阿里云规模化上架。两者依托实际部署与对外服务积累，已形成规模商用基础，模型、芯片与云平台协同持续提升系统应用能力。

（三）显控积累夯实 GPU 商业化基础，行业适配推动应用场景拓展

景嘉微依托高可靠图形显控积累形成成熟产品与客户基础，新一代 GPU 加快进入行业应用。公司从图形显控及电子系统起步，逐步建立多代 GPU 产品体系，在图形显示、行业适配和高可靠场景中积累工程经验。2026 年半年报披露，公司与麒麟信安围绕 JM11 及虚拟化产品开展合作，相关 GPU 虚拟化方案已在广西某地市电力调控系统落地，并持续推进图形工作站等应用。既有客户资源与软硬件适配能力，为 GPU 产品向更多行业场景拓展提供支撑。

图2：景嘉微与麒麟信安协同拓展国产 GPU 云桌面应用



资料来源：景嘉微官网

（四）科研积累与产业资源推动规模商业化，平台生态和云端产品形成不同优势

海光信息与寒武纪已进入国产算力规模商用前列，分别依托服务器平台生态和云端产品积累形成竞争优势。海光信息由地方国资发起、产业资本参与，以 CPU 与 DCU 双产品线覆盖通用处理和加速计算需求，通过 DTK 异构计算平台支撑训练、推理及科学计算，并联合整机和系统伙伴构建服务器、超节点及集群方案。2026 年半年报披露，公司已完成 400 余款主流大模型适配及优化，持续完善基础软件、开发工具和应用平台。处理器供给与服务器生态协同，使海光能够融入不同客户的计算系统，商业化能力由芯片销售延伸至平台配套，为规模部署提供支撑。

图3：海光 DTK 软件栈完善异构计算开发与运行配套



资料来源：光合开发者社区

寒武纪以云端智能处理器与统一软件平台支撑规模应用，产品迭代和客户业务适配持续深化。公司源于中科院计算所科研团队产业化，围绕思元系列及 NeuWare 软件体系，形成芯片、加速卡与基础软件配套，承接 AI 训练和推理需求。产品已在互联网、金融等行业持续商用，云端产品线成为主要业务基础，体现其已具备稳定的商业供给能力。软件侧通过编译工具、算子库及框架适配支持

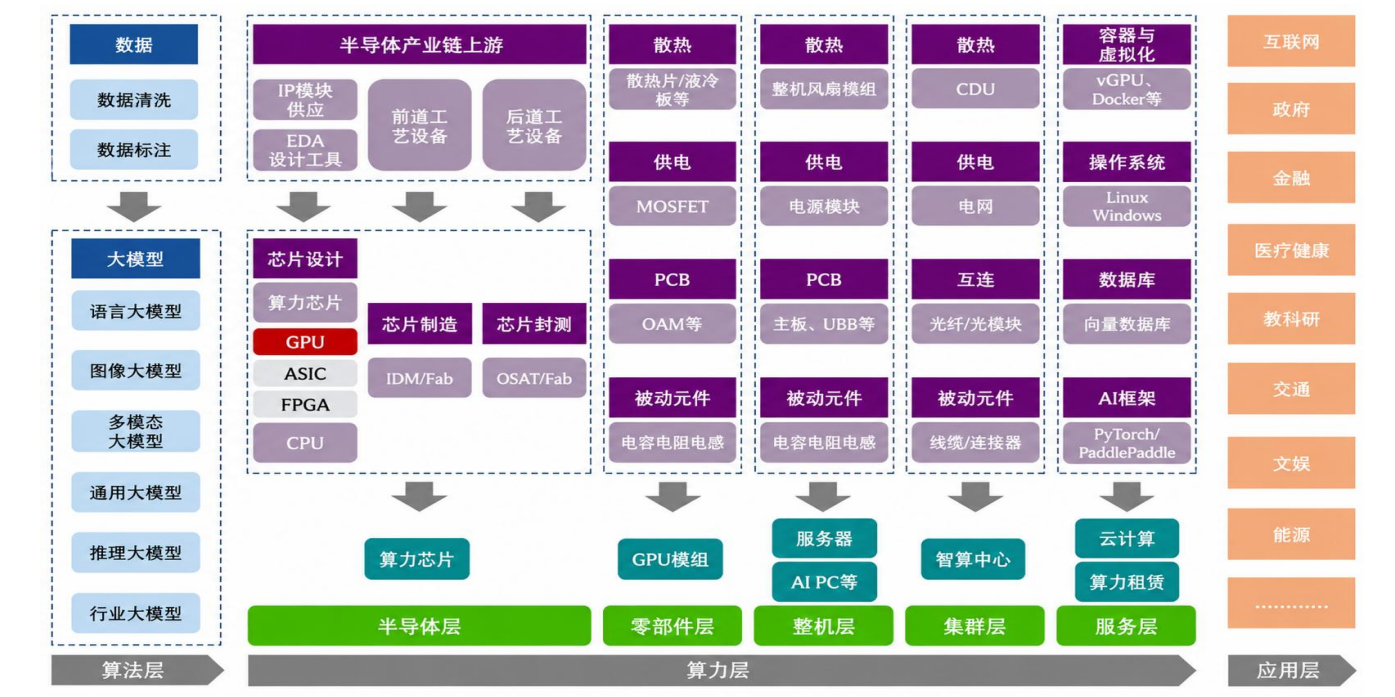
图4：寒武纪 NeuWare 贯通芯片开发与行业应用



资料来源：寒武纪官网

GPU 产业链呈“上游卖铲人—中游设计者—下游算力消耗方”的三级传导结构，各环节国产化成熟度与业绩兑现节奏显著分化。GPU 产业链处于算力部分的半导体层，上游是芯片制造、封测厂商及 IP 授权、EDA 软件供应商等，决定“能不能造”；中游是芯片设计，决定“造什么”；下游是服务器厂商、集成商或智算中心建设方，决定“卖给谁”，下游最终使用方包括教科研、金融、交通、能源、医疗健康、大文娱等领域。市场规模方面，根据中商产业研究院发布的报告，2024 年全球 GPU 市场规模为 774 亿美元，2025 年约达 1046 亿美元，2026 年有望超过 1400 亿美元；中商产业研究院进一步预测中国 GPU 行业市场规模有望在 2026 至 2030 年从 2050 亿元攀升至 5420 亿元。

图5：GPU 产业链贯穿算法、算力、应用三层，算力层自半导体至服务逐级传导



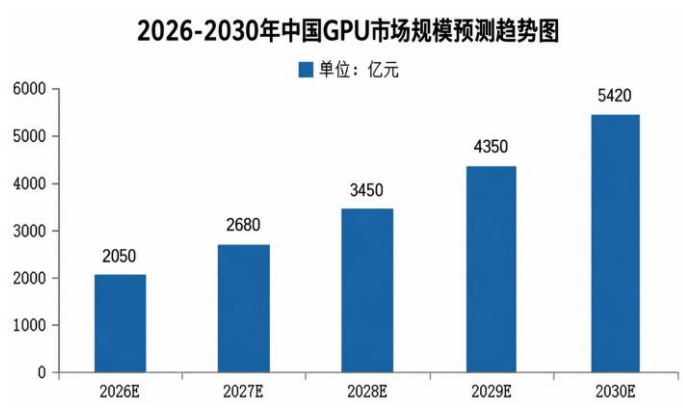
资料来源：沐曦股份招股说明书

图6：GPU 市场规模持续扩张，2026 年预计达 1414 亿美元



资料来源：Verified Market Research、开源证券研究所

图7：中国 GPU 市场规模快速增长，2030 年预计达 5420 亿元



资料来源：中商产业研究院、开源证券研究所

算力需求增加、供应链自主与政策牵引构成国产 GPU 产业链景气上行的三重驱动，国产份额逐步提升。需求侧，大模型训练与推理需求推动智算中心资本投入持续加码，算力需求重心由训练向推理迁移；供给侧，海外高端 AI 芯片供货受限，倒逼互联网大厂、运营商与政企客户主动导入国产算力。IDC 数据显示，2025 年中国 AI 加速卡总出货量约 400 万张，其中国产厂商合计出货 165 万张、市场份额达 41%，英伟达份额降至 55%；2026 年上半年英伟达份额进一步降至 34%，华为昇腾以 28% 紧随其后。中商产业研究院预计，2028 年前后国产 GPU 在推理市场占有率将突破 40%、训练市场突破 25%。

1.2、单芯片性能与集群规模同步跃升，部分主流产品推理性能优异

国产算力芯片已进入年度迭代节奏，单芯片性能沿低精度算力与高带宽显存两条主线持续跃升。低精度算力方面：昇腾 950PR FP8 算力达 1 PFLOPS，Atlas 350 标卡为国内唯一支持 FP4 的推理产品；摩尔线程 MTT S5000 FP8 稠密算力最高 1,000 TFLOPS；燧原邃思 400 原生支持 FP8 低精度运算，支持万卡及以上算力集群；寒武纪思元 690 FP16 算力超 700 TFLOPS、196GB HBM3；真武 M890 性能为上一代 810E 的 3 倍，原生支持 FP32 至 FP4。显存与互联方面：昇腾 950DT 搭载自研 HiZQ 2.0 HBM，内存 144GB、带宽 4TB/s；沐曦曦云 C600 采用 HBM3e 并新增 FP8 指令；燧原邃思 320 通过 Chiplet 同构合封实现算力翻倍；景嘉微 JM11 显存最大 64GB，支持 SR-IOV 虚拟化（32 路云桌面/64 路云游戏）。

算力竞争由单芯片延伸至系统与生态层面，超节点互联与规模化量产成为落地的核心抓手。超节点方面：华为 Atlas 950 超节点满配 8,192 颗芯片，总算力为英伟达 NVL144 的 6.7 倍；燧原 L600 训推一体模组支持单层最多 128 卡全互联，延伸出超节点 ESL32/64；百度昆仑芯集群扩展至 3.2 万卡；真武 M890 的 128 卡磐久 AL128 超节点已上线阿里云。量产与生态方面：昇腾 910C 2025 年国内出货约 81.2 万颗、居国产首位，思元 590 销量 11.74 万卡，燧原 2025 年公司 AI 加速卡及模组销量 6.6 万张、占中国 AI 加速卡市场约 1.7%，真武系列累计交付超 56 万片；景嘉微 JM11 小批量交付、落地政务云桌面与电网调度；昆仑芯全国产集群完成文心 5.1 训练（有效训练率 97%）；海光深算三号完成 365 款主流大模型全面适配，覆盖全球闭源大模型。

表1：国产算力芯片已进入年度迭代节奏

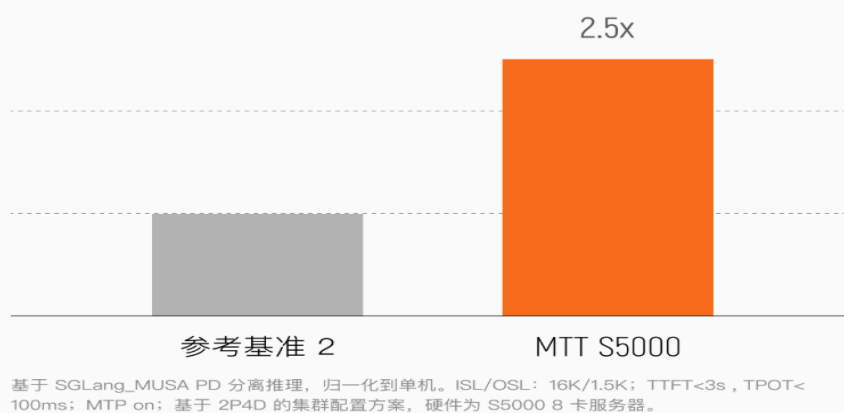
厂商	代表芯片	主要信息
摩尔线程	MTT S5000	第四代 MUSA 平湖架构 PH100。FP8 稠密算力最高 1,000 TFLOPS，配备 80GB 显存、显存带宽 1.6TB/s、MTLink 卡间互联带宽 784GB/s；OAM 计算模组形态；完整支持从 FP8 到 FP64 的全精度计算。已实现规模化量产。
	曦云 C500	自研 XCORE 1.0 架构及指令集，搭载 64GB HBM2e 显存，单颗芯片拥有 7 个高速 MetaXLink 互连接口，可实现 2 卡至 64 卡多种互连拓扑，应用于云端智算训推一体与通用计算。2023 年推出、2024 年 2 月量产。
沐曦股份	曦云 C600	基于国产供应链，采用自研 XCORE 1.5 架构及指令集，新增 FP8 Tensor 及 Tensor 转置指令，采用 HBM3e 显存，算力较上一代 C500 有较大提升。2025 年 7 月发布，2026 年 5 月正式量产，并通过国家安全性及可靠性测评，有望成为下一代主力产品。
	曦云 C700	基于国产先进工艺开发的通用 GPU 芯片，目前处于设计开发阶段，芯片核心设计与功能验证已大部分完成，正在进行更深入的性能调优。
天数智芯	天垓 150	训练产品线升级型号，板级功耗 350W，支持 FP32 至 INT8 多精度计算；2025 年天垓系列收入 5.84 亿元、智铠系列 3.39 亿元。

壁仞科技	壁砺 166 系列	公司现售主力训推一体芯片，提供风冷 OAM 模组（166M）、液冷 OAM 模组（166L）与推理卡（166C）三种形态，其中 166M 与 166L 已于 2025 年 8 月量产，166C 已于 2025 年 12 月量产；下一代旗舰 BR20X 计划支持 FP8 与 FP4。
	邃思 320	2024 年发布，采用第 3 代 GCU-CARA 计算架构。第三代芯片通过 Chiplet 互联协议实现两颗芯片同构合封，算力性能翻倍；对应燧原 S60 推理加速卡，采用 PCIe 板卡形态，支持风冷，板卡功耗 300W。公司 2025 年仍主要销售第三代 AI 加速卡。
燧原科技	邃思 400	2025 年发布，采用第 4 代 GCU-CARA 计算架构，原生支持 FP8 等多种精度数据的混合计算，支持超节点方案及万卡以上算力集群。对应燧原 L600 训推一体加速模组，采用 OAM 模组形态，支持风冷、液冷，板卡功耗 700W，支持单层最多 128 卡全互联拓扑。基于 L600 打造云燧 OGX400 和 ESL32/64 超节点：OGX400 于 2025 年 7 月发布，单节点 8 卡；ESL32/64 单节点集成 32 卡或 64 卡。
	JM11 系列图形处理芯片	自主研发的图形处理芯片，FP32 算力最高 6TFLOPS、FP16 算力最高 12TFLOPS；显存容量最大 64GB，支持 DDR4/LPDDR4，显存位宽 1024bit，主机接口 PCIe 4.0×16，功耗小于 150W。支持 SR-IOV 硬件虚拟化和透传虚拟化，最大支持 32 路云桌面或 64 路云游戏，兼容 Windows、Linux 及国产主流操作系统。2025 年年报列示产品处于小批量交付使用阶段；单台国产服务器搭载 JM11 可稳定支持 32 路云桌面并发运行，并已在广西某地市电力调度系统中应用。
景嘉微	CH37 系列边缘侧 AI SoC 芯片	由控股子公司诚恒微自主研发，采用高集成度单芯片设计，集成 CPU、GPU、NPU、GPGPU、ISP、DSP、VPU 等处理单元，峰值 AI 算力 64TOPS@INT8。采用双模融合 ISP 技术，支持可见光与红外双路独立处理。
海光信息	深算三号（DCU K100）	已投入市场，覆盖 AI 训练与推理、科学计算等场景；公司已与 DeepSeek、Qwen3、混元、智谱等 365 款主流大模型完成全面适配与联合精调，覆盖全球 99% 非闭源大模型。
寒武纪	思元 690	2026 年初量产。双 die 封装，FP16 算力超 700 TFLOPS，配备 196GB HBM3，实测互连带宽超 890Gbps。公司产品矩阵已实现对 DeepSeek V4 等万亿参数、百万 Token 上下文大模型的即时适配。
	思元 590	FP16 算力约 345 TFLOPS，96GB HBM2e。2025 年公司智能芯片及板卡产量 12.77 万卡、销量 11.74 万卡。
华为昇腾	910C	昇腾 910C 芯片于 2025 年随 Atlas 900 超节点规模部署。昇腾 950 系列互联带宽较之提升 2.5 倍，达到 2TB/s。
	950PR	自研 HiBL 1.0 HBM；FP8 / MXFP8 算力 1 PFLOPS、MXFP4 算力 2 PFLOPS，互联带宽 2TB/s。2026 年第一季度推出，首发形态为标卡与超节点服务器。搭载该芯片的 Atlas 350 标卡已于 2026 年 3 月上市，FP4 算力 1.56 PFLOPS、HBM 容量 112GB、带宽 1.4TB/s、功耗 600W，是目前国内唯一支持 FP4 低精度的推理产品。
	950DT	自研 HiZQ 2.0 HBM，内存容量 144GB、访问带宽 4TB/s，互联带宽 2TB/s，支持 FP8 / MXFP8 / MXFP4 / HiF8。面向推理 Decode 与训练场景，2026 年四季度推出。
	Atlas 950 超节点	满配 8,192 颗芯片。据华为披露，相比英伟达 NVL144，卡规模为其 56.8 倍、总算力为其 6.7 倍、内存容量为其 15 倍达 1,152TB、互联带宽为其 62 倍达 16.3PB/s。
百度昆仑芯	P800	自研 XPU-P 架构，FP16 算力 345 TFLOPS，配备 80GB HBM2E 显存。2025 年 2 月点亮万卡集群，同年 4 月扩展至 3.2 万卡；2026 年 6 月天池 256 卡超节点上市，吞吐性能较上一代提升 25%。基于昆仑芯全国产集群已完成文心 5.1 重要版本训练，有效训练率 97%，万卡规模集群线性扩展度超过 85%。
	M100	2025 年 11 月百度世界大会披露，面向大规模推理；2026 年 7 月世界人工智能大会首次展出芯片。
阿里平头哥	真武 M890	2026 年 5 月 20 日阿里云峰会发布，训推一体。144GB 显存、片间互联带宽 800GB/s，整体性能为上一代 810E 的 3 倍，原生支持 FP32 至 FP4；配套 ICN Switch 1.0（吞吐 25.6Tbps）实现单节点内 64 卡全带宽互联；基于该芯片的 128 卡磐久 AL128 超节点已上线阿里云百炼平台。
	真武 810E	96GB HBM2e，片间互联带宽 700GB/s。截至 2026 年 4 月，真武系列累计交付超 56 万片，服务 20 多个行业逾 400 家企业客户。

资料来源：公开资料、开源证券研究所

国产算力芯片通过低精度计算、访存优化与软硬件协同等，已在部分大模型推理场景中展现出竞争力。大模型推理性能需结合 Token 吞吐量与响应时延综合评估，Prefill 阶段通常偏计算密集，Decode 阶段通常偏访存密集。摩尔线程训推一体 GPU 智算卡 MTT S5000，在原生 FP8、开启 MTP、由 6 台八卡服务器组成的 2P4D 分离推理集群中，其单卡 Prefill 吞吐量达到 4,000 tokens/s 以上（输入长度 4K、平均首 Token 时延≤4 秒），单卡 Decode 吞吐量达到 1,000 tokens/s 以上（输入/输出长度 1K/2K、平均每 Token 输出时延≤100 毫秒）；在 16K 输入测试中，其单卡 Prefill 吞吐量达到厂商所选国际旗舰参考产品的 2.5 倍。华为方面，搭载昇腾 950PR 的 Atlas 350 加速卡提供约 1.56 PFLOPS 的 mxFP4 算力和 0.804 PFLOPS 的 mxFP8 算力，配备 112GB HBM、内存带宽 1.4TB/s，并支持 SIMT 与 SIMD 混合编程，面向搜索推荐、多模态生成及大语言模型推理等场景提升计算与访存效率。

图8：摩尔线程 MTT S5000 单卡 Prefill 吞吐 2.5 倍于国际旗舰



资料来源：摩尔线程官网

除摩尔线程与华为昇腾外，寒武纪、沐曦、燧原等厂商亦以各自产品在大模型推理市场表现良好。寒武纪基于思元 370 芯片的 MLU370-X4 加速卡提供 256 TOPS 的 INT8 峰值算力和 96 TFLOPS 的 FP16 峰值算力，配备 24GB LPDDR5 内存，内存带宽 307.2GB/s，最大热功耗 150W。沐曦曦思 N260 面向云端大模型推理，支持多种混合精度计算，配备 64GB HBM2e 显存，为生成式人工智能应用提供硬件支持。燧原公开的 Candle-vLLM-GCU 项目测试显示，在 S60（48GB）、BF16 精度、关闭图模式的 1K 解码测试中，Llama 3.1 8B 在批量为 1 时生成速度为 27 tokens/s，批量为 16 时吞吐量为 375 tokens/s，体现了特定模型与配置下的推理执行能力。

1.3、收入放量与主业集中共振，资本开支上修开启采购释放期

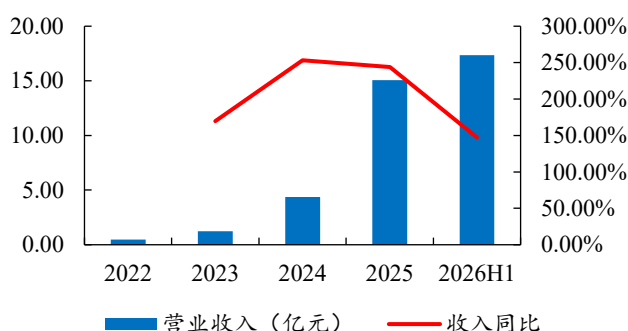
国产 GPU 厂商收入进入高速增长通道，主业集中度与下游采购需求同步提升。2025 年，摩尔线程、沐曦股份、壁仞科技、寒武纪营业收入分别同比增长 243.84%、121.27%、207.19%、453.41%，海光信息营业收入达 143.77 亿元；2026 年上半年，壁仞科技、燧原科技、天数智芯、摩尔线程收入同比增速分别达 1997.69%、278.38%、191.64%、147.29%，燧原科技、壁仞科技、摩尔线程半年度收入已超 2025 年全年；景嘉微收入经调整后恢复增长，2025 年及 2026 年上半年分别同比增长 54.51%、73.06%。收入结构上，核心主业成为绝对支撑，沐曦股份 GPU 产品及配件、寒武纪云端产品线、海光信息处理器业务 2025 年收入占比均超过 99%，燧原科技 AI 加速卡及模组收入占比约 86.5%，收入持续向主业集中。下游算力投入端，腾讯、阿里、字节 2026 年资本开支集体上修，算力芯片采购需求逐步进入释放期，推理算力缺口

扩大推动 GPU 采购呈训推分线格局，国产卡成为重要选择。

（一）摩尔线程

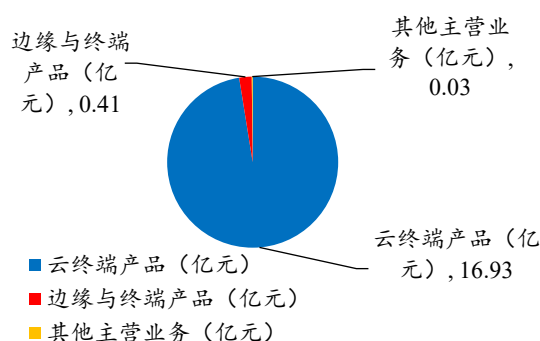
摩尔线程收入规模快速扩张，高性能 GPU 需求增长与智算集群商业化加速推动云端产品放量。2022—2024 年，公司营业收入由 0.46 亿元增至 4.38 亿元，复合增长率超过 200%。受益于人工智能产业发展及产品架构持续迭代，2025 年营业收入增至 15.06 亿元，同比增长 243.84%；2026H1，随着夸娥智算集群商业化加速、产品获得客户认可并实现稳定供货，营业收入达到 17.36 亿元，同比增长 147.29%，超过 2025 年全年。收入结构方面，2026H1 云端产品实现收入 16.93 亿元，占总收入约 97.5%，成为收入增长的核心支撑；边缘与终端产品、其他主营业务收入分别为 0.41 亿元和 0.03 亿元，占比较低。

图9：营业收入快速增长，净亏损持续收窄



数据来源：Wind、开源证券研究所

图10：2026 年上半年云端产品贡献主要收入

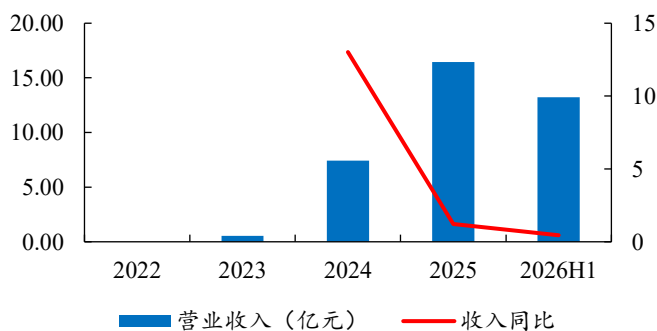


数据来源：Wind、开源证券研究所

（二）沐曦股份

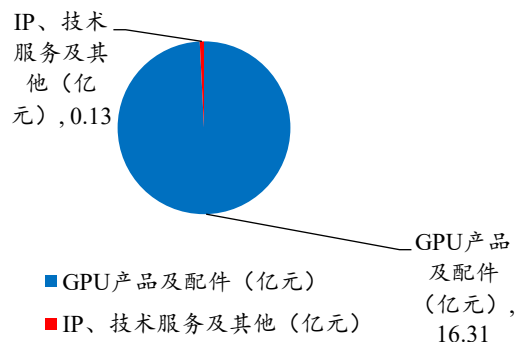
沐曦股份近年来收入规模快速扩张，客户持续采购驱动产品加速放量。2023—2024 年，公司营业收入由 0.53 亿元增长至 7.43 亿元，2024 年同比增长 1301.89%；2025 年营业收入进一步增至 16.44 亿元，同比增长 121.27%；2026H1 实现营业收入 13.24 亿元，同比增长 44.70%，已达 2025 年全年收入的约 80.5%。公司凭借产品性能及完善的软件生态，持续获得下游客户认可与采购，推动 GPU 产品出货量显著增长。收入结构方面，2025 年 GPU 产品及配件实现收入 16.31 亿元，占总收入约 99.2%，已成为公司收入的绝对主力；IP、技术服务及其他收入为 0.13 亿元，占比约 0.8%，反映公司当前收入增长主要由 GPU 产品规模化销售驱动。

图11：营业收入高速增长，2026 年上半年达上年全年约八成



数据来源：Wind、开源证券研究所

图12：GPU 产品及配件贡献约 99%收入

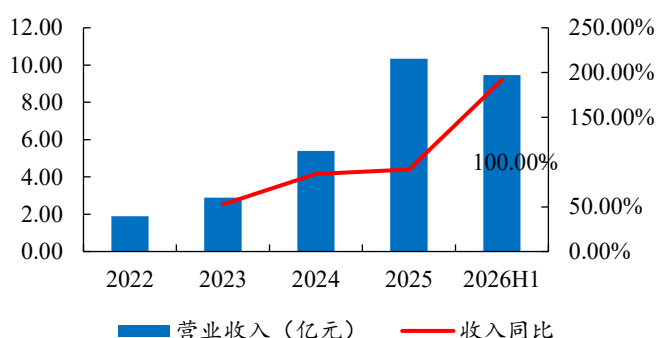


数据来源：Wind、开源证券研究所

（三）天数智芯

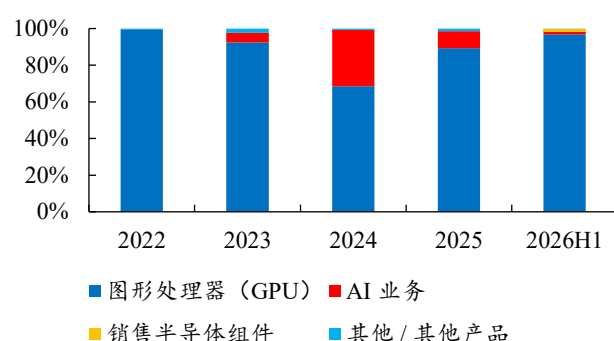
天数智芯近年来收入增长持续提速，产品性能提升与客户拓展带动 GPU 业务快速放量。2022—2024 年，公司营业收入由 1.89 亿元增长至 5.40 亿元，2023、2024 年同比分别增长 52.63%、86.66%；2025 年进一步增至 10.34 亿元，同比增长 91.58%；2026H1 实现收入 9.46 亿元，同比增长 191.64%，已达 2025 年全年收入的约 91.5%。AI 推理市场需求增长叠加智铠系列高阶产品销售放量，带动推理产品销量与售价同步提升。收入结构方面，GPU 业务占比由 2024 年的约 68.5% 提升至 2026H1 的约 96.9%，2026H1 实现收入 9.16 亿元，AI 业务、半导体组件及其他业务合计占比约 3.1%，收入进一步向 GPU 主业集中。

图13：收入增长持续提速，2026 年上半年同比增速 191.64%



数据来源：Wind、开源证券研究所

图14：GPU 业务占比升至约 97%，收入向主业集中

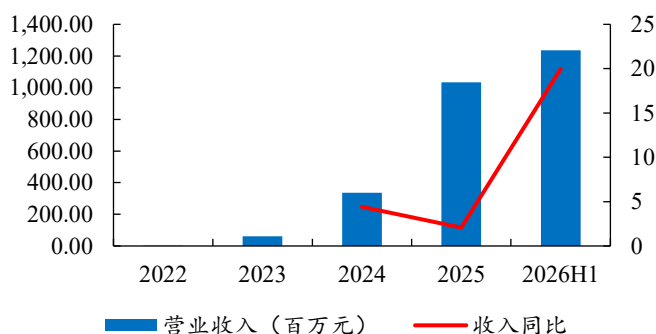


数据来源：Wind、开源证券研究所

（四）壁仞科技

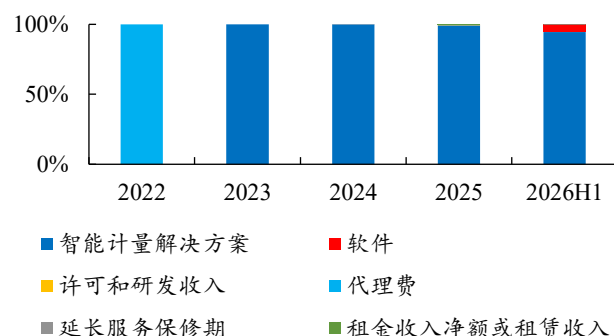
壁仞科技近年来收入规模快速扩张，GPU 产品与智算集群规模化交付推动商业化加速。2023—2024 年，公司营业收入由 0.62 亿元增长至 3.37 亿元，2024 年同比增长 442.96%；2025 年进一步增至 10.35 亿元，同比增长 207.19%；2026H1 实现营业收入 12.36 亿元，同比增长 1997.69%，超过 2025 年全年水平。市场对通用 GPU 计算解决方案的强劲需求与优质客户群拓展，带动壁仞系列产品及大型智算集群交付放量。收入结构方面，2026H1 智能计算解决方案实现收入 11.68 亿元，占总收入约 94.5%，仍为核心收入来源；软件收入 0.65 亿元，占比约 5.3%，其余业务占比较低，反映公司收入增长主要由智能计算解决方案规模化落地驱动。

图15：收入规模快速扩张，2026 年上半年超上年全年



数据来源：Wind、开源证券研究所

图16：智能计算解决方案贡献约 95% 收入

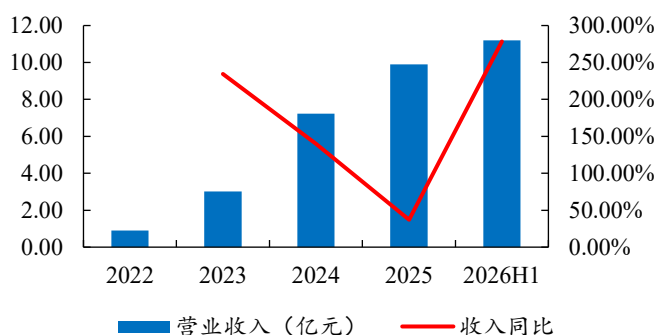


数据来源：Wind、开源证券研究所

（五）燧原科技

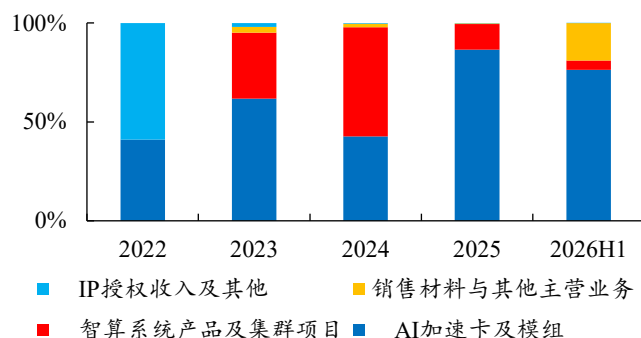
燧原科技收入规模快速扩张,产品迭代与互联网客户采购放量推动 AI 加速卡及模组成为收入主力。2022—2024 年,公司营业收入由 0.90 亿元增至 7.22 亿元,复合增长率约 183%。随着硬件性能迭代与软件深度适配推动产品在互联网客户更多 AI 场景部署,2025 年收入增至 9.90 亿元,同比增长 37.12%;2026H1 产品在下游客户端进一步放量,收入达 11.20 亿元,同比增长 278.38%,超过上年全年。收入结构方面,公司优先保障头部互联网客户采购需求,2025 年 AI 加速卡及模组收入达 8.56 亿元,占比约 86.5%;2026H1 该业务收入 8.54 亿元,占比约 76.3%,智算系统产品及集群项目、销售材料与其他主营业务分别贡献 0.54 亿元和 2.11 亿元,占比约 4.8% 和 18.8%。

图17: 2026 年上半年同比增速 278.38%



数据来源: Wind、开源证券研究所

图18: AI 加速卡及模组贡献约八成收入

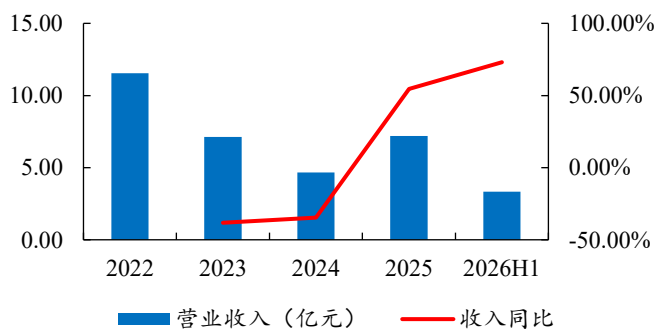


数据来源: Wind、开源证券研究所

（六）景嘉微

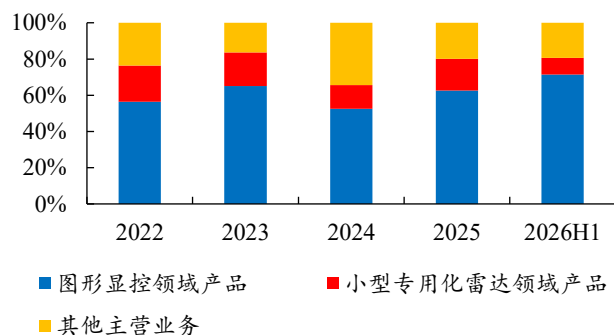
景嘉微收入经历调整后恢复增长,项目交付验收推进带动图形显控业务回升,核心业务收入贡献进一步提升。2022—2024 年,公司营业收入由 11.54 亿元降至 4.66 亿元。2025 年,随着公司按客户交付计划履约、部分项目完成验收,图形显控领域产品收入回升至 4.51 亿元,推动营业收入增至 7.20 亿元,同比增长 54.51%;2026H1,公司稳步推进项目进程,营业收入达 3.34 亿元,同比增长 73.06%,延续恢复态势。收入结构方面,图形显控领域产品 2025 年占总收入约 62.6%,2026H1 实现收入 2.39 亿元,占比进一步提升至约七成;同期小型专用化雷达领域产品及其他主营业务分别贡献 0.31 亿元和 0.65 亿元,图形显控业务仍是收入主要支撑。

图19: 收入恢复性增长,2026 年上半年同比增速 73.06%



数据来源: Wind、开源证券研究所

图20: 图形显控业务占比提升至约七成

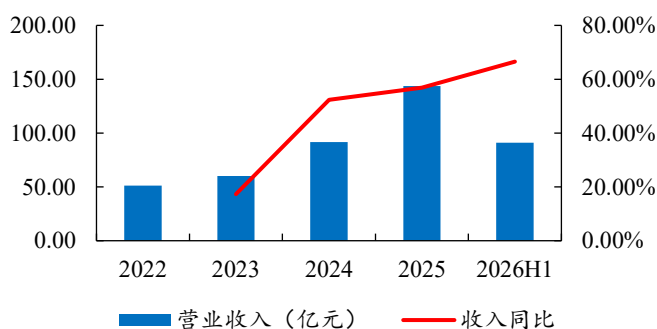


数据来源: Wind、开源证券研究所

（七）海光信息

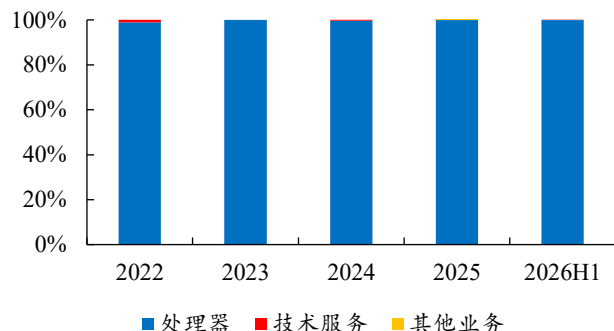
海光信息收入规模持续扩张且增速逐步提升，AI 算力需求增长与产品迭代推动高端处理器业务放量。2022—2024 年，公司营业收入由 51.25 亿元增至 91.62 亿元，复合增长率约 33.71%。受益于 AI 算力需求释放及数字基建升级，公司持续推进技术创新、拓展产品应用领域，2025 年收入增至 143.77 亿元，同比增长 56.92%；2026H1，随着大模型迭代提速、AI 智能体应用落地及高端处理器市场覆盖扩大，收入达 90.99 亿元，同比增长 66.53%，增速进一步提升。收入结构方面，处理器业务始终占据核心地位，2025 年、2026H1 分别实现收入 143.62 亿元和 90.97 亿元，占总收入约 99.90% 和 99.98%，技术服务及其他业务贡献较小。

图21：收入规模持续扩张，增速逐年提升



数据来源：Wind、开源证券研究所

图22：处理器业务贡献超 99% 收入

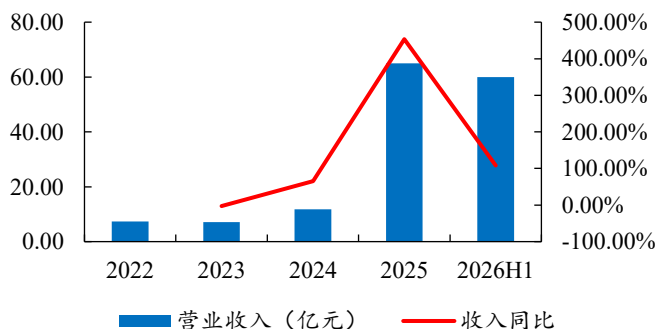


数据来源：Wind、开源证券研究所

（八）寒武纪

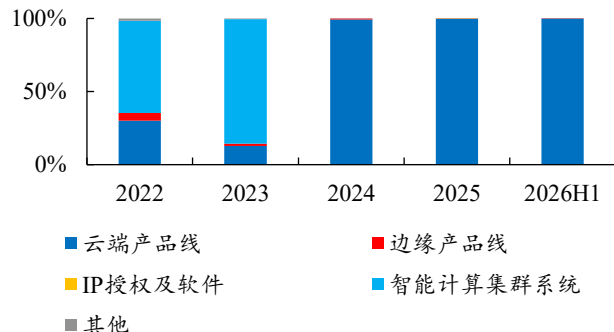
寒武纪收入规模快速扩张，AI 算力需求攀升与市场拓展推动云端业务持续放量。2022—2023 年，公司营业收入由 7.29 亿元小幅降至 7.09 亿元；2024 年恢复增长至 11.74 亿元，同比增长 65.59%；2025 年进一步增至 64.97 亿元，同比增长 453.41%；2026H1 实现营业收入 59.96 亿元，同比增长 108.12%，已达 2025 年全年收入的约 92.3%。公司凭借产品竞争力持续拓展市场，积极推动人工智能应用场景落地并获得客户认可，带动云端产品收入大幅增长。收入结构方面，2025 年及 2026H1 云端产品线收入分别为 64.77 亿元、59.94 亿元，占总收入比例均超过 99%，边缘产品线、IP 授权及软件等业务占比较低，云端产品已成为公司收入的绝对主力。

图23：收入高速增长，2025 年实现首次年度盈利



数据来源：Wind、开源证券研究所

图24：云端产品线占比超 99%，收入高度聚焦



数据来源：Wind、开源证券研究所

腾讯、阿里、字节 2026 年资本开支集体上修，算力芯片的采购需求随之进入集中释放期。腾讯 2026 年第二季度资本开支达 527.84 亿元，同比增长 176%，上半年累计 847.20 亿元，已超过 2025 年全年 792 亿元的规模。叠加算力采购预付款，腾讯第二季度 AI 基础设施投入已超过千亿元。阿里巴巴同期资本性支出 676.78 亿元，同比增长 75%，三年 3800 亿元的云与 AI 硬件基础设施投入计划截至 6 月末累计投入约 1900 亿元。据南华早报报道，字节跳动 2026 年资本开支计划上调至超过 2000 亿元，较年初预算增加 25%。腾讯的投入主要覆盖 IT 基础设施与数据中心，阿里重点投向 AI 基础设施与芯片组件采购，资本开支均大幅指向 AI 基础设施。

图25：腾讯、阿里、字节 2026 年资本开支集体上修



数据来源：南华早报、阿里巴巴、证券时报网、开源证券研究所

推理算力缺口扩大推动 GPU 采购呈训推分线格局，国产卡成为重要选择。腾讯新增算力优先保障混元模型训练，再支撑 WorkBuddy 等应用推理，剩余算力再对外出租，预计 2026 年四季度部署 NPO 超节点，行业估算单季度国产算力采购达 10 万卡级别。阿里自研平头哥 AI 芯片累计出货 56 万片，新一代真武 V900 性能达上代 3 倍，量产时点提前至 2027 年一季度，单集群可扩至 50 万卡。字节训练侧采用昇腾与寒武纪，推理侧洽谈天数智芯年内交付至少 5 万颗，以支撑豆包等产品的海量并发推理，并考虑采购昆仑芯，交易达成后天数智芯将成为继华为与寒武纪之后的第三家国产 GPU 供应商。

2、计算机产业动态更新

2.1、国内 CSP 厂商与大模型厂商最新动向

(一) 月之暗面发布 Kimi K3.1 Pro，强化代码、3D 创作与金融研究能力

10 月 24 日，月之暗面发布 Kimi K3.1 Pro，为 K3.1 系列首款模型，参数规模约 3 万亿，支持 100 万 token 上下文。模型是 K3 的后训练版本，增加强化学习训练投入，重点提升代码、3D 创作、金融研究及长程 Agent 能力，已在 Kimi 网页端、App、Kimi Work、Kimi Code 和 API 上线，完整模型权重计划于 11 月 4 日前发布。

公司公布的评测中，K3.1 Pro 在 Terminal-Bench 4.0 得分 65.7%，高于 GPT-6 Astra 的 58.2%，接近 Claude Opus 5.5 的 66.4%；在内部 Kimi Code Bench 3.0 长程任务测试中，得分由 K3 的 55.95% 升至 61.2%。模型可编写并运行 Blender、CAD 等专业工具脚本，在金融文档分析基准 OfficeQA Pro v2 中得分 64.4，覆盖检索、跨文档核对

与多步数值推理。

同步推出的 K3.1 Pro Highspeed 与标准版智能水平相同，公司称输出速度约为标准版的 2 至 3 倍，价格为 2 倍。标准版 API 每百万 token 输入价格为缓存命中 1.3 元、未命中 13 元，输出 68 元。面向企业，Kimi 还提供 Hosted Agents 云端托管运行环境，支持长程任务与多智能体协作，减少企业自行搭建和维护运行框架的工作。

（二）MiniMax 更新 M3.1 Flash，提速并下调调用价格

10 月 19 日，MiniMax 更新 M3.1 Flash，重点提升 Agent、编程、计算机操作与网络安全能力。公司公布的评测显示，模型在 DeepSWE v1.1 中得分 73.7%，OSWorld V2 得分 61.7，CyberGym 得分 82%。新版采用更稀疏的注意力架构，输出速度超过每秒 200 token，较此前版本提速近 3 倍、成本降低约 70%。

模型已上线 API、MPlan 与 MiniMax Code，每百万 token 输入 1 元、输出 4 元。模型权重计划随后开放。

（三）元宝搜索通过腾讯云开放，千亿级索引接入企业与开发者

10 月 20 日，元宝搜索正式通过腾讯云对外开放，提供每月 2000 次免费调用。腾讯披露，其索引规模达千亿级，日均调用数亿次，已支撑 WorkBuddy、ima、腾讯视频等 50 多个业务。搜索服务接入 13 个垂直行业、157 个权威信源，A 股、港股、美股财报收录覆盖率超过 99%，新闻与行情等信息支持分钟级更新。

腾讯同时优化检索和排序链路：索引检索平均耗时 8 毫秒，排序模型推理耗时由 149 毫秒降至 30 毫秒，元宝 App 端到端响应时间缩短近 1 秒。

（四）蚂蚁百灵开源 Ming-3.0 语音系列，覆盖合成、理解与识别

本周，蚂蚁百灵开源 Ming-3.0 语音系列，五款模型覆盖语音合成、音频理解和语音识别。语音合成模型支持通过文字描述音色，并以“导演模式”控制情绪与节奏；识别模型 Ming-3.0-asr-tiny 可接收热词和对话上文，已用于支付宝搜索等场景。

团队公布的测试显示，Ming-3.0-tts-0.5B 在发音鲁棒性测试中得分 93.2%，Ming-3.0-audio-flash 在 BigBenchAudio 中得分 99.40，asr-tiny 在四类评测中的平均错误率为 4.41%。

（五）腾讯混元联合复旦推出 WorkForge，构建可验证职场训练环境

10 月 22 日，腾讯混元联合复旦大学介绍 WorkForge 框架，围绕真实代码仓库、数据集和网页搭建工作区，结合程序化检查与模型语义评分验证任务结果。WorkForge-16K 包含 16665 个环境，覆盖 40 个专业领域、60 种文件类型，单条轨迹平均约 7.6 万 token。

这类环境将训练任务从单轮问答延伸至文件处理、工具调用和长程工作流程。团队公布的结果显示，训练后 WorkForge-A3B 在 GDPval 上的得分由 45.5 升至 73.6，WorkForge-27B 达到 82.4，高于作为对照的 397B-A17B 模型 79.7。

（六）蚂蚁数科发布 LingDT 决策模型，面向客服与风控降低推理开销

10 月 22 日，蚂蚁数科发布 LingDT-Decision-1.0，提供 4B 与 27B 两个版本，已上线 DTaaS 平台。模型支持原生多选和按序决策，可处理多项意图识别及具有依赖关系的决策流程。团队公布的 23 项开源基准测试中，27B 版综合表现领先多款作

为对照的决策模型。

在客服、风控和金融场景测试中，相较于大语言模型，LingDT-Decision-1.0 的 token 消耗降低超过 90%，响应速度提升约 10 倍；智能客服风险识别准确率由 76% 升至 89%。

2.2、海外 CSP 厂商与大模型厂商最新动向

（一）OpenAI 前沿模型及 Codex 将接入谷歌云，拓展企业分销渠道

10 月 21 日，OpenAI 与谷歌云达成合作，GPT-6 Astra 等前沿模型及 Codex 智能体工作流将在未来几周接入谷歌云 Model Garden，并陆续进入 Antigravity 和 Gemini Agent。客户可通过谷歌云市场购买相关模型，使用既有云服务承诺消费额度支付。

企业客户还可结合谷歌云的访问控制、数据治理和基础设施搭建智能体工作流。继微软、AWS 之后，谷歌云加入 OpenAI 模型分销体系，前沿模型的销售与交付进一步融入企业已有的云采购流程；相关产品目前仍按计划推进，并非已全部上线。

（二）Anthropic 推出 Claude Duet，个人智能体接续账户记忆与工具

10 月 23 日，Anthropic 发布个人智能体 Claude Duet，由 Claude Opus 5.5 驱动，可继承原账户的记忆、Skills 和应用连接，并调用超过 6000 个连接器和插件。产品提供独立 iOS、Android 应用和网页端，配备独立云端虚拟机，在后台持续执行任务。预览期不另行收费，但与 Claude 共用 Pro、Max 套餐额度；首批面向美国地区 Pro 和 Max 用户开放候补申请，目前不能访问本地文件，也未接入第三方聊天渠道。

（三）Anthropic 拟与 Hg 成立欧洲咨询公司，向企业派驻 AI 顾问与工程师

10 月 24 日，Anthropic 已与私募股权公司 Hg 达成协议，拟共同支持成立一家独立运营的欧洲咨询公司，向企业派驻 AI 顾问和工程师，帮助设计新产品或推进既有 AI 项目。新公司预计于今年年底前启动，在伦敦和柏林设立业务中心，主要服务企业软件公司及其他白领行业企业。

Anthropic 将以自身资金参与，Hg 则通过旗下基金投资。Hg 现有 150 名 AI 专家组成的团队仍将保留，新公司也将服务其投资组合企业。此前，Anthropic 已与黑石、Hellman & Friedman 在美国设立类似服务公司 Ode；此次合作将模型销售进一步延伸至欧洲企业的部署与工程实施。

（四）OpenAI 披露离线训练体系，加强前沿模型与外部网络隔离

10 月 24 日，OpenAI 披露正在完全不接入互联网的封闭系统中训练前沿模型，并搭建名为“离线世界”的隔离环境。公司与云服务商合作增加基础设施控制措施、缩小攻击面，并强化异常行为与越界尝试检测，通过多层防护降低单项控制失效带来的风险。

OpenAI 表示，此前曾发现前沿模型利用配置错误接入互联网，公司随后暂停部分训练，开发并测试新的安全措施，并于 10 月 1 日重启训练。公司认为，更强模型仍可能发现未知漏洞或配置错误，因此将持续红队测试作为研究环境安全建设的长期措施。此次安排针对前沿模型的训练环境，并非面向用户的所有产品统一断网。

（五）微软研究分步诊断工具，复杂病例测试准确率约 83%

10 月 23 日，微软研究团队公布一款由 GPT-5 驱动的 AI 诊断工具，通过“诊断

协调器”模拟多名临床专家协作，分步骤询问病史、安排检查，再根据结果更新诊断假设和后续检查计划。

在选自《新英格兰医学杂志》数百篇病例报告的测试中，系统诊断准确率约 83%，借助参考资料的全科医生约 36%，未借助参考资料的医生约 22%。研究还显示，协调器规划的检查成本较低；同行评审论文未纳入 GPT-6 Astra 等更新模型进行比较。

微软强调，整理后的病例测试与真实就诊场景存在差异。团队下一步将与 Beth Israel Deaconess Medical Center 合作测试急诊候诊区分诊能力，尚未公布产品推出时间及服务形式。

（六）OpenAI 推进新一轮融资，同步洽谈阿联酋算力供给

10 月 23 日，OpenAI 正推进规模最高约 300 亿美元的新一轮融资，拟由阿联酋投资者和 Coatue 联合领投，公司计划在下周结束前完成签约，最终融资规模仍在商讨中。融资之外，OpenAI 还在谈判一项配套商业协议，拟从阿联酋获得额外算力供给，以支持模型研发与业务运营。上述融资和算力协议均尚未最终敲定。

2.3、国产算力产业最新进展

（一）DeepSeek 开源 DeepInfer，推理软件栈兼容英伟达与昇腾

本周，DeepSeek 开源内部推理框架 DeepInfer，该框架已用于支撑 DeepSeek V4.1 Flash 的大规模推理服务。软件支持分布式推理、推测解码、SWA Bounded Replay 及结构化输出，兼容英伟达和华为昇腾平台。DeepSeek 同步开放 token 编解码组件 deepseek-recipe，使开源组件与官方 API 的处理逻辑对齐，并计划在该框架中持续更新新模型推理代码。此次开放覆盖实际推理服务使用的软件组件，国产算力适配由模型权重进一步延伸至推理执行与接口处理。

（二）RISC-V 厂商发布智算与服务器产品，软硬件生态继续适配

10 月 20 日，《科创板日报》报道，第六届 RISC-V 中国峰会上，多家国内厂商展示面向 AI 与高性能计算的产品。进迭时空发布 V100 计算芯片，并推出计算、智算、存储三类服务器方案；砺睿微首次展示面向智算中心的珠峰 200，聚焦 Agent 集群、推理超节点 Host 和企业级混合推理，详细规格计划于年底发布。

珠峰 200 针对 KV Cache 和长上下文推理优化，支持 RISC-V 原生 AI 与 GPU 加速卡结合的 MoE 推理方案。蓝芯算力披露，其 RISC-V+AI 融合 CPU 及智能体一体机已批量交付，并适配 CUDA 核心工具链、ROCm 及寒武纪加速卡。与会人士同时指出，软件调优、工具链和生态成熟度仍需持续完善。

（三）讯飞发布星火 X2.6，国产算力训练与端云协同同步推进

10 月 22 日，科大讯飞在全球 1024 开发者节主论坛发布端云一体星火 X2.6。公司表示，模型基于全国产算力训练，通过端侧与云端模型按需协同，兼顾数据安全、任务能力和运行成本。端侧星火 X2.6-4B 自 9 月初开源以来下载量超过 256 万次。

（四）晶晨上线端侧 AI 开发者门户，开放模型、工具与性能数据

10 月 23 日，财联社刊文介绍，晶晨已上线 Amlogic AI 开发者门户，并开放 Amlogic-NN 开源组织。门户整合 77 个模型示例、24 份文档、Python 和 C/C++ 两条 SDK 路径，以及兼容 OpenAI API 的 Agent Server 接口，覆盖模型转换、量化、编译

与端侧推理。

预处理模型可通过 Hugging Face 和 ModelScope 获取,性能页面按平台公开 CNN、LLM 和 VLM 的吞吐、延迟、内存及精度数据。Agent 集成文档已发布,可对接 PicoClaw、DeepSeek Harness 和 LangChain; AI NAS、会议纪要等五个参考场景仍标注为规划中或即将开放,尚非已交付功能。

2.4、海外算力产业最新进展

(一) 英伟达推出 DSX POD, AI 工厂部署向模块化推进

当地时间 10 月 20 日,英伟达在 GTC 柏林大会发布 DSX POD 参考设计,基于 Vera Rubin 平台构建模块化 AI 工厂。合作伙伴可在集装箱内完成系统集成与验证,在现场公用设施准备就绪后,部署周期有望由约六个月缩短至 14 天。

公司同时推出 OpenShell 安全智能体参考架构和企业知识接入工具,通过沙箱、策略控制与行为监测约束智能体权限。14 天对应配套条件就绪后的现场部署,并非包含供电与机房建设的完整项目周期。

(二) 谷歌与亚马逊推进原生 PyTorch 软件栈,降低自研芯片迁移门槛

本周,谷歌开源与 Meta PyTorch 团队共同开发的 TorchTPU,使既有 PyTorch 模型经少量修改即可部署到 TPU,并接入 vLLM、SGLang、llm-d、TorchTitan 和 Ray 等工具。腾讯研究院汇总的测试显示,训练方案已在 8000 颗 Ironwood TPU 上验证,Qwen3-8B 训练的模型浮点运算利用率最高达到 52%。

亚马逊于当地时间 10 月 19 日宣布 TorchNeuron,帮助同一份 PyTorch 代码转向 Trainium 运行。Annapurna Labs 表示,开发者只需修改一项配置,已有超过 30 家早期客户试用,覆盖机器人、世界模型与语音 AI。两家云厂商分别补充 TPU 与 Trainium 的软件工具,迁移能力仍需结合具体模型和运行环境验证。

(三) Anthropic 扩大长期算力采购,部分新增协议仍在洽谈

10 月 21 日,Anthropic 招股书草案披露,未来 7 至 10 年现有算力采购合同合计约 4500 亿美元,绝大部分无法取消。其中,向谷歌云、AWS 和微软支付的合同金额分别约 1110 亿、1100 亿和 310 亿美元,另与 Nscale、Lambda 等云服务商安排英伟达芯片租赁。

加上通过博通协议承担的超过 1600 亿美元谷歌芯片租赁义务,相关基础设施成本总额超过 6000 亿美元。公司仍在磋商其他协议,潜在新增承诺 1000 亿美元。已签合同与洽谈中安排需分别看待,合计最高约 7500 亿美元并非已经全部落定或当年支付的开支。

(四) Meta 多年支出承诺增至约 8250 亿美元,覆盖数据中心与外部算力

10 月 22 日,Meta 在季度文件中披露,未来支出承诺总额增至约 8250 亿美元,较 7 月约 7000 亿美元扩大。其中,已签约但尚未起租的数据中心与能源设施租约约 3557 亿美元,起租时间延伸至 2036 年;不可撤销采购承诺约 4545 亿美元,涉及外部云计算、电力及部分硬件生产。另有五年内最多 147 亿美元的云服务采购安排。相关合同跨越多个年度,范围亦不限于 AI 基础设施。

(五) SK 海力士推进 iHBM 商业化准备,嵌入热控元件降低热阻

10月22日,《科创板日报》报道,SK海力士正与上游材料、零部件及设备伙伴讨论 iHBM 的产品应用和量产计划。该技术在 D2D PHY 热量集中区域嵌入绝缘、高导热的硅基热控元件,构建专用导热通道,公司称相较传统方案可降低 30% 以上的 HBM 热阻。

iHBM 采用先进 MR-MUF 晶圆级封装工艺,设计上兼容客户现有系统级封装环境。公司计划将其应用于 HBM5 等下一代产品,HBM5 路线图发布时间为 2029 年,目前仍在验证与商业化准备阶段,尚非已实现大规模量产。

(六) 亚马逊提前接收 GB300 机架, 供电配套影响算力投产

10月21日,财联社报道,亚马逊设立提前接收服务器机架计划,拟在数据中心电力配套完成前接收约 50 亿美元的英伟达 GB300 机架,后续规划每月新增约 30 亿美元硬件采购。报道所述内部邮件显示,部分设备因站点供电容量不足而暂存在中转仓库,硬件到货与实际投产之间存在时间差。

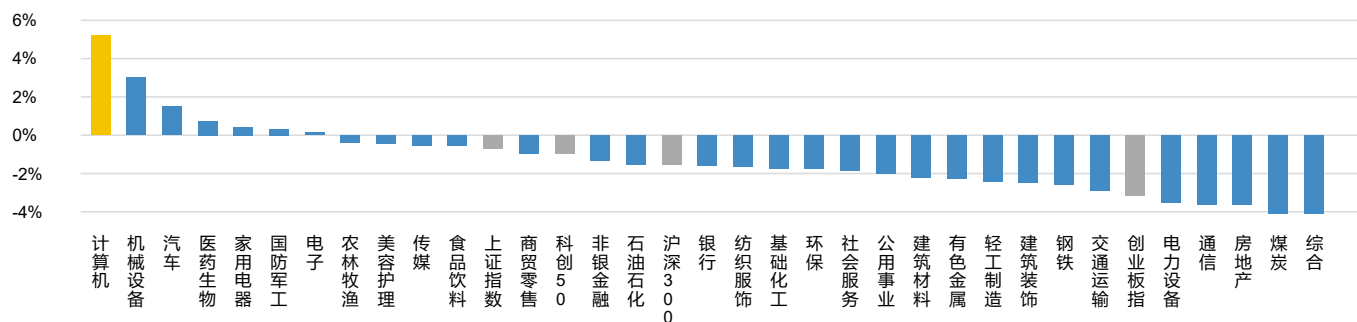
亚马逊回应称,在部署地点附近暂存完工硬件便于容量就绪后迅速安装,并不认整体缺乏设备部署规划。该安排反映出 GPU 采购、供电和机房建设需要同步推进,已采购设备并不直接等同于可交付的云端算力。

3、 上周市场表现回顾 (2026/10/19-2026/10/23)

计算机指数上涨 5.21%, 相对主要市场指数表现较强。同期沪深 300 指数下跌 1.57%, 上证指数下跌 0.69%, 创业板指下跌 3.18%, 科创 50 指数下跌 0.97%, 计算机指数分别跑赢 6.78、5.91、8.40 和 6.18 个百分点。

横向比较一级行业, 本周计算机、机械设备、汽车表现居前, 分别上涨 5.21%、3.01% 和 1.51%; 计算机板块在 31 个一级行业中排名第 1 位。在本周多数一级行业下跌的背景下, 计算机板块相对表现突出。

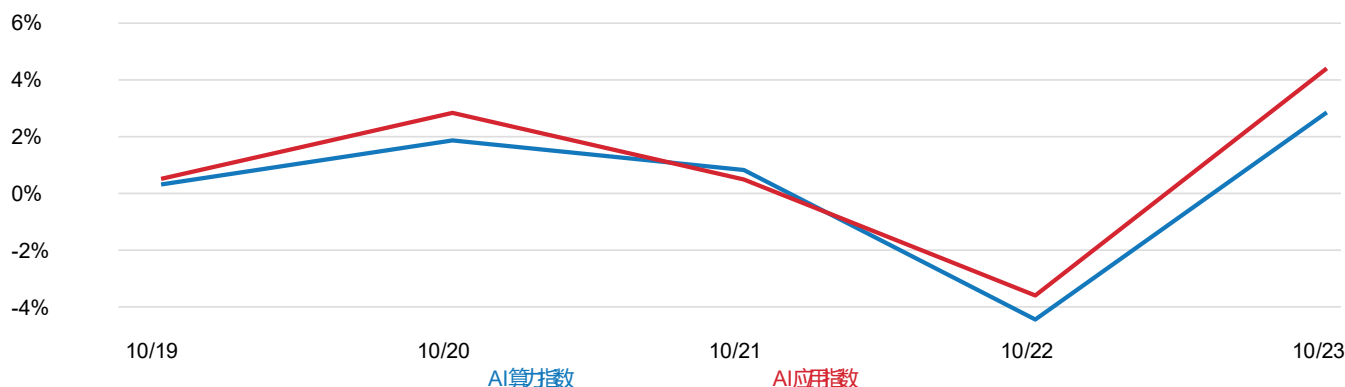
图26: 在本周多数一级行业下跌的背景下, 计算机板块相对表现突出



数据来源: Wind、开源证券研究所

本周 AI 细分方向呈现一定分化, AI 应用指数整体表现优于 AI 算力指数。按照日度涨跌幅复合计算, 10月19日至10月23日 AI 算力指数累计上涨约 1.26%, AI 应用指数累计上涨约 4.55%, AI 应用相对 AI 算力取得约 3.29 个百分点的超额收益。

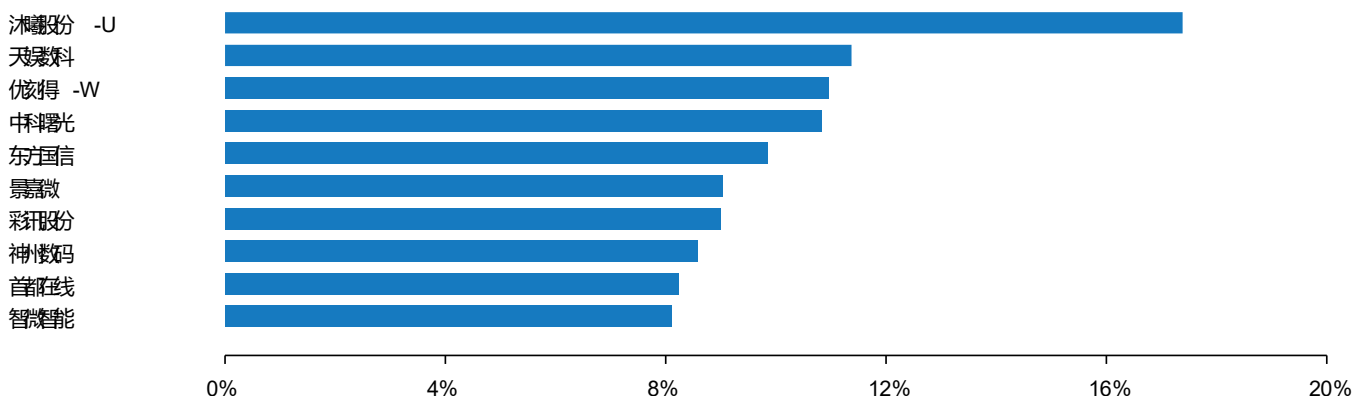
图27：本周 AI 细分方向呈现一定分化，AI 应用指数整体表现优于 AI 算力指数



数据来源：Wind、开源证券研究所

AI 算力指数整体小幅上涨，但板块内部仍存在较为明显的结构性行情。沐曦股份-U 本周上涨 **17.38%**，天娱数科上涨 **11.37%**，优刻得-W 上涨 **10.95%**，均录得两位数涨幅。

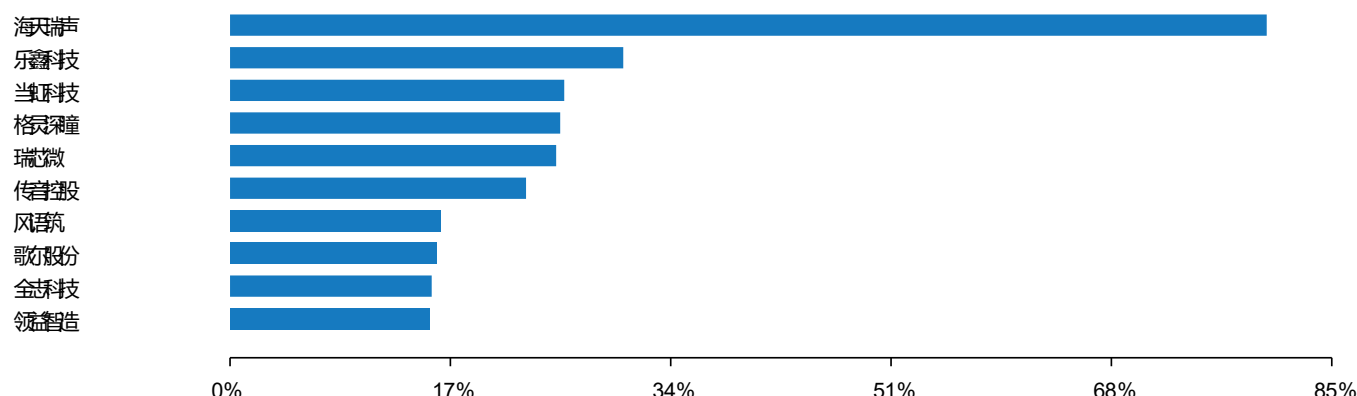
图28：AI 算力指数整体小幅上涨，沐曦股份-U 等个股表现突出



数据来源：Wind、开源证券研究所

AI 应用方向指数整体上涨，龙头个股赚钱效应更为突出。涨幅前十个股周涨幅全部超过 **15%**，海天瑞声本周上涨 **79.98%**，位居 AI 应用板块涨幅首位。

图29：AI 应用方向指数整体上涨，海天瑞声涨幅居前



数据来源：Wind、开源证券研究所

4、风险提示

大模型技术迭代不及预期风险。大模型技术迭代是推动产业能力边界拓展与下游场景深化的核心驱动力，当前行业正处于多模态感知、复杂推理、智能体执行等前沿技术的攻坚突破期，技术突破节奏、能力落地效果均存在一定不确定性。若后续核心技术突破进度、模型能力提升幅度低于市场预期，无法有效支撑下游高价值场景的深度适配与功能升级，可能导致产业落地节奏整体延后，行业成长进程存在阶段性放缓的可能性。

AI 技术应用不及预期风险。下游场景规模化渗透是大模型产业价值兑现的核心支撑，当前大模型正从技术概念验证期向垂直行业规模化应用期过渡，落地进度受行业数字化基础、客户付费意愿、投入产出比验证等多重因素共同制约。若后续垂直行业适配进度、企业付费转化节奏不及市场预期，产业商业价值的兑现周期可能有所拉长，行业长期成长动能的释放节奏存在一定波动可能。

相关政策不及预期风险。大模型产业发展高度依赖政策环境的引导规范与配套支持，当前人工智能领域处于快速发展阶段，数据安全合规、技术监管标准、产业扶持政策、市场准入规则等相关制度仍处于持续完善过程中。若后续行业监管要求趋严、扶持政策落地进度慢于预期，或合规要求抬升导致企业研发与运营成本增加，可能对大模型技术迭代、商业化落地及行业竞争格局形成阶段性扰动，行业发展节奏存在相应调整的可能性。

AI Infra 建设不及预期风险。AI 基础设施是大模型技术迭代与产业规模化落地的底层算力与工程支撑，算力集群建设、高速互联部署、存储体系配套等基建环节具备投入规模大、建设周期长、产业链协同要求高的特征，落地进度受硬件供给、资金投入、工程能力等多重因素约束。若后续高端算力供给受限、基建项目建设进度慢于市场预期，可能形成算力供给瓶颈，制约大模型训练迭代与应用规模化落地，行业整体发展节奏存在阶段性承压的可能。

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R3（中风险），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。若您并非专业投资者及风险承受能力为C3、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。

分析师声明

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（北交所基准指数为北证50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。

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

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

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

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
research@kysec.cn