

信义山证汇通天下

证券研究报告

计算机

AI 算力专题报告

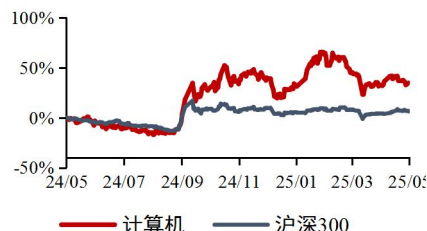
领先大市-A(维持)

AI 产业高景气持续，算力国产化大势所趋

2025 年 5 月 27 日

行业研究/行业深度分析

计算机板块近一年市场表现



资料来源：最闻

首选股票		评级
688041.SH	海光信息	买入-A
000977.SZ	浪潮信息	买入-A
603019.SH	中科曙光	买入-A

相关报告：

【山证计算机】机器人产业趋势持续强化，重视上下游投资机会 2025.5.14

【山证计算机】AutoGLM 沉思重磅发布，集深度思考与强工具使用能力于一体 2025.4.1

分析师：

方闻千

执业登记编码：S0760524050001

邮箱：fangwenqian@sxzq.com

研究助理：

邹昕宸

邮箱：zouxinchen@sxzq.com

投资要点：

- **互联网和智算中心两大下游 AI 算力需求持续，国产化采购快速推进。**1) 互联网：24 年以来互联网大厂仍持续投入基础大模型训练，并在今年开始布局推理大模型，推动算力资本开支保持快速增长，随着 DeepSeek 的突破大幅降低应用端门槛以及 Agent 技术和产品日趋成熟，推理侧需求有望逐渐成为 AI 算力支出的主驱动力。同时，互联网大厂正在加速推进国产 AI 芯片采购；2) 智算中心：中央和地方政策明确国产 AI 算力建设目标，奠定 AI 算力建设的政策基调，24 年以来各地方智算中心项目招投标数量仍在持续增加，同时多个全国产化标杆项目已跑通，未来智算中心的 AI 算力国产化率有望加速提升。
- **AI 芯片：国产替代空间广阔，第一梯队芯片厂商正加速突破性能、产能、生态瓶颈。**目前华为昇腾、海光信息、寒武纪、百度昆仑芯国产 AI 芯片第一梯队厂商正加速追赶英伟达，目前华为昇腾 910B 已经基本可对标 A100，华为昇腾 910C 预计性能可对标 H100，已成为国内互联网厂商国产训练芯片的首选，同时寒武纪、海光信息、昆仑芯新一代主力产品思元 590 芯片、深算三号、昆仑芯 3 代未来有望对英伟达 H20 形成替代。除了硬件性能的快速提升，我们认为一方面，国产芯片厂商正在加快解决国内晶圆代工等供应链问题，另一方面，在生态上，国产 AI 芯片厂商自研及兼容策略，加快突破 CUDA 的生态壁垒。由于目前英伟达在国内 AI 芯片市场仍占据主导地位，国产化替代存在巨大空间，考虑美国芯片出口管制政策存在进一步加码的可能，将倒逼国内客户加快推进国产 AI 芯片采购。
- **AI 服务器：AI 驱动服务器市场格局变化，昇腾系厂商份额持续提升。**国内主要互联网厂商对 AI 服务器仍将保持旺盛需求，根据 IDC 预测，到 2028 年中国加速服务器市场规模将达到 253 亿美元，2024-2028 年均复合增速超过 20%。24 年以来，H20 仍在国内互联网大厂及智算中心采购中占据较高份额，推动以浪潮信息为代表的服务器领军业绩高速增长，同时随着昇腾出货量的快速增长，以超聚变和华鲲振宇为代表昇腾系服务器厂商市场份额随之持续提升，在运营商及智算中心采购项目中占据较高份额。除了传统服务器厂商及昇腾系厂商之外，以华勤技术为代表的本土 ODM 厂商，凭借在消费电子领域积累的研发体系和供应链优势正在快速切入到 AI 服务器市场。
- **投资策略：**芯片出口管制趋严叠加国产芯片性能提升，两大下游互联网和智算中心的国产 AI 算力需求将持续高景气。在芯片端，看好新一代产品性能有望大幅提升的第一梯队厂商，推荐海光信息，关注寒武纪、景嘉微等；



请务必阅读最后股票评级说明和免责声明

1

在服务器端，看好 H20 放量对传统服务器领军的业绩拉动，推荐浪潮信息，关注华勤技术等，同时看好昇腾芯片份额提升带来的投资机会，关注四川长虹、神州数码、拓维信息等。

风险提示：互联网对国产 AI 芯片的采购不及预期，智算中心项目 AI 算力国产化率提升放缓，国产 AI 芯片技术研发不及预期，行业竞争格局加剧风险。

目录

1. AI 产业高景气持续，算力国产化大势所趋.....	6
1.1 算力需求及国产化两大因素共同推动国产算力发展.....	6
1.2 互联网：训练算力需求持续，推理侧需求将逐渐占据主导.....	8
1.3 智算中心：全国产化项目已跑通，国产化率有望持续提升.....	11
2. AI 芯片：国产替代空间广阔，第一梯队国产芯片正加速替代.....	15
2.1 英伟达仍占据国内市场主导，政策加码创造国产替代空间.....	15
2.2 国产 AI 芯片加速追赶，互联网及智算国产化采购快速提升.....	17
2.3 国产厂商基于兼容及自研两大路径逐渐突破 CUDA 生态壁垒.....	21
3. AI 服务器：AI 驱动服务器市场格局变化，昇腾系份额持续提升.....	26
3.1 AI 服务器高速增长，市场规模迅速追赶通用服务器.....	26
3.2 AI 上游芯片及下游需求驱动服务器市场竞争格局剧变.....	27
3.3 昇腾系计算产业生态成型，两大领先级部件伙伴将占主要份额.....	30
4. 盈利预测及投资建议.....	32
5. 风险提示.....	35

图表目录

图 1： 2023 和 2024 年的生成式 AI 支出方向及变化.....	6
图 2： 互联网和智算中心是国内 AI 算力两大下游.....	7
图 3： 国产 AI 算力需求框架.....	8
图 4： AI 驱动国内互联网大厂资本开支持续回暖（单位：亿元）	9
图 5： 近半年来文心一言、豆包、通义千问、腾讯元宝的应用月活变化情况（月活单位：万）	10
图 6： 国内云厂商高端 AI 芯片中 80%来自英伟达（24 年初）	11
图 7： 未来 5 年英伟达 AI 芯片在国内云厂商高端 AI 芯片中占比将降至 50-60%.....	11

图 8: 智算中心建设的参与方众多.....	13
图 9: 2022 年英伟达主导国内 AI 芯片出货量 85%.....	15
图 10: 2024 年国产 AI 芯片约占 30%的份额.....	15
图 11: 英伟达特供版芯片 H20、L20、L2 对比.....	17
图 12: 寒武纪、海光信息、昆仑芯等厂商的新一代主力产品有望对英伟达 H20 形成替代.....	21
图 13: 英伟达 CUDA 架构.....	22
图 14: 海光信息 DTK 高度兼容 CUDA.....	24
图 15: 昇腾 AI 计算体系更完整.....	25
图 16: 国内服务器总市场规模及增速.....	26
图 17: 国内 AI 服务器市场规模及增速.....	26
图 18: 全球服务器总市场规模及增速.....	27
图 19: 全球 AI 服务器市场规模及增速.....	27
图 20: 2023 年国内 x86 服务器市场规模及份额.....	28
图 21: 2023 年国内 AI 服务器市场规模及份额.....	28
图 22: 22Q1-24Q4 浪潮信息营收变化情况.....	29
图 23: 22Q1-24Q4 浪潮信息存货变化情况.....	29
图 24: 中国移动 2024 年至 2025 年新型智算中心采购项目中标份额情况（标包 1）.....	30
图 25: 中国电信 AI 算力服务器（2023-2024 年）集中采购项目中标份额情况.....	30
图 26: 昇腾计算产业生态中主要有 4 类合作伙伴.....	31
图 27: 2019-2025Q1 海光信息收入及增速.....	33
图 28: 2019-2025Q1 海光信息归母净利润及增速.....	33
图 29: 2019-2025Q1 浪潮信息收入及增速.....	34
图 30: 2019-2025Q1 浪潮信息归母净利润及增速.....	34

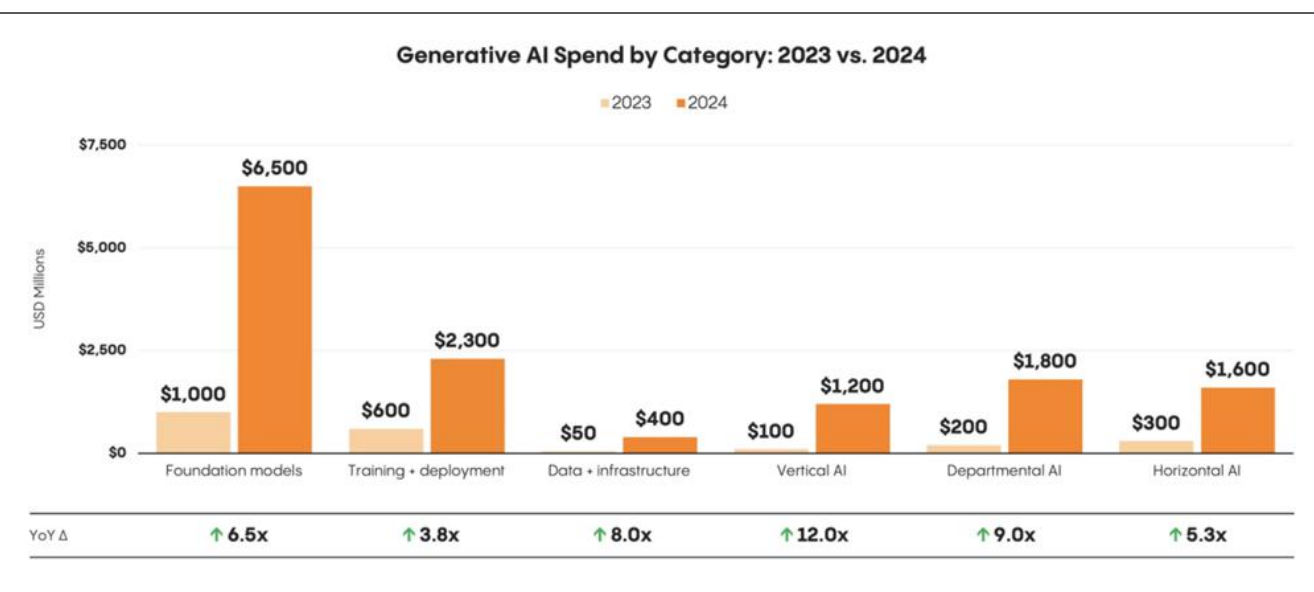
表 1: AI 算力建设纳入全国性及各省市的政策目标中.....	12
表 2: 各地方仍不断推进智算中心建设.....	14
表 3: 美国芯片出口管制政策持续加码.....	16
表 4: AI 芯片厂商可分为 GPGPU、FPGA 和 ASIC 三大阵营.....	18
表 5: 近年来国产 AI 芯片厂商快速涌现.....	19
表 6: 英伟达、AMD 及主要国产 AI 芯片对比.....	20
表 7: 国产 AI 芯片厂商主要通过兼容 CUDA 以及自研两种方式构建生态.....	23
表 8: 华鲲振宇（四川长虹）和昆仑技术（超聚变）是唯二的领先级昇腾部件伙伴.....	31
表 9: 主要公司盈利预测和估值情况.....	32

1. AI 产业高景气持续，算力国产化大势所趋

1.1 算力需求及国产化两大因素共同推动国产算力发展

全球算力需求持续高景气，目前基础模型训练仍是推动生成式 AI 需求增长的主要动力，而应用层需求正加速增长。1) 训练侧：目前大模型仍在沿着 Scaling Law 的路径持续升级，头部大模型厂商正在投入更大规模的模型，根据 Menlo Ventures 数据，对基础模型的投资支出从 2023 年的 10 亿美元增加至 2024 年的 65 亿美元，提升了 5.5 倍。同时，推理大模型的加速布局也将抬升后训练阶段的算力投入，并进一步推动训练侧需求增长；2) 推理侧：AI 应用有望迎来爆发式增长，一方面，ChatGPT 等 AI 原生应用近期的用户月活流量持续增长，同时 DeepSeek R1 的突破将大幅降低应用开发和部署门槛，进而加速 AI 应用渗透；另一方面，Agent 与 RAG 正加速企业级 AI 落地，代表产品包括 Palantir 的 AIP 平台、Glean 的 Assistant、Manus 等。

图 1：2023 和 2024 年的生成式 AI 支出方向及变化



资料来源：Menlo Ventures，山西证券研究所

国内市场互联网和智算中心两大下游仍保持旺盛的 AI 算力需求。1) 互联网：无论在国内市场还是全球市场，互联网/云厂商都是 AI 算力的主要下游，根据 Trendforce 数据，2023 年国际大型 CSP 占全球高端 AI 服务器需求的 65%，同时，根据 IDC 数据，2024 年国内互联网厂商占国内 AI 服务器需求的比重超过 65%。23-24 年国内互联网及云厂商的算力采购仍主要用于大模型训练，随着大模型军备竞赛进一步升级以及推理模型和混合模型布局加速，预计 25 年训练侧算力需求仍将持续，同时随着 DeepSeek 大幅降低应用开发和

部署门槛，推理侧需求将成为推动 AI 算力增长的重要增量，有望逐渐成为 AI 算力增长的主要驱动力；2）智算中心：国内智算中心建设从 23 年开始显著加速，2024 年以来相关招投标项目仍在持续增加。目前智算中心的下游用户主要集中于政府机构、院校及科研院所、大模型企业等，而政府机构、院校及科研院所、大模型企业都有持续且明确的训练或推理需求，因此预计 2025 年智算中心建设仍将持续。

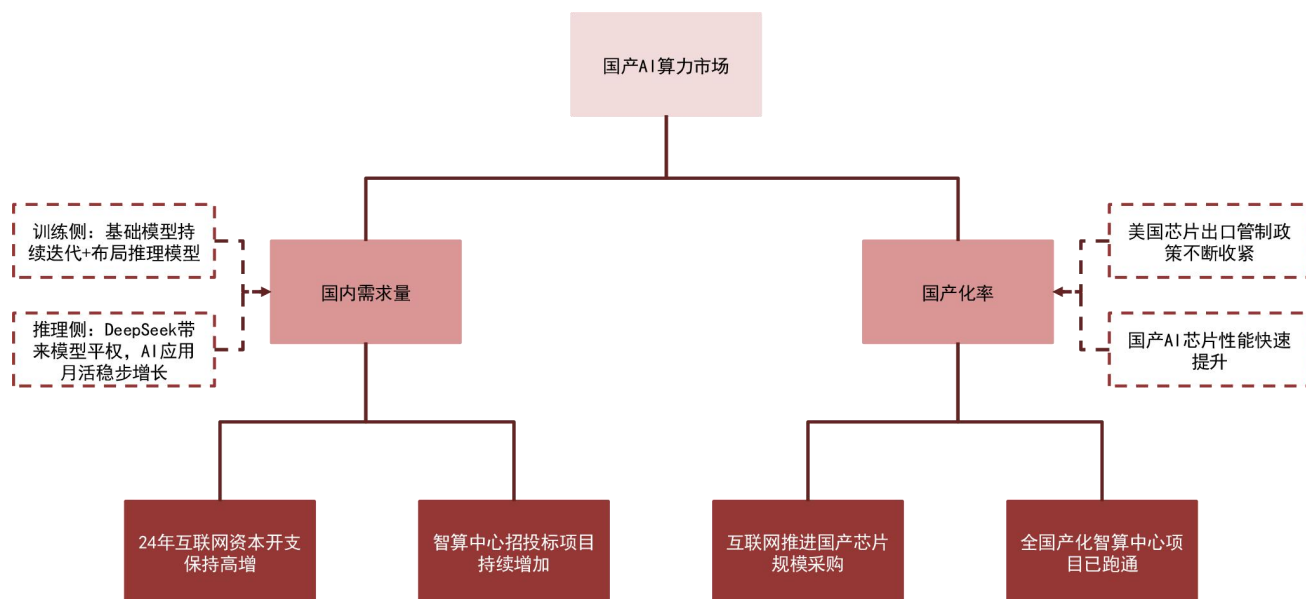
图 2：互联网和智算中心是国内 AI 算力两大下游

下游	需求	AI 芯片	AI 服务器厂商
互联网	AI 大模型持续研发训练（赋能自身业务、服务下游客户） 推理侧需求爆发（DeepSeek 大幅降低应用开发及部署门槛、Agent 快速成熟）	英伟达占主导，开始批量采购昇腾芯片	浪潮信息、新华三、宁畅、华勤技术等
智算中心	各地大力投入智算中心建设	英伟达占比较高，昇腾、海光等国产芯片占据一定份额	浪潮信息、新华三等主流服务器厂商
企业	AI 大模型推理需求、本地化大模型研发（金融、大型国企等）	英伟达占比较高，昇腾、海光等国产芯片占据一定份额	浪潮信息、新华三、联想、超聚变等

资料来源：山西证券研究所绘制

在 AI 算力总需求持续增长的同时，国产化是推动国产 AI 算力快速成长的主要驱动力。目前互联网和智算中心的国产 AI 芯片采购占比仍处于较低水平，而随着美国针对高性能 AI 芯片的出口管制政策逐步趋严，同时以华为昇腾、海光信息、寒武纪、昆仑芯为代表的国产厂商的新一代 AI 芯片正快速突破性能和生态瓶颈，加速追赶并替代海外竞品。与此同时，头部互联网厂商正逐步推进国产芯片的规模化采购，百度、阿里、腾讯、字节等厂商正不断增加国产 AI 芯片的采购。同时，在中央和地方政策助力以及全国产化项目的标杆效应下，智算中心的 AI 芯片国产化率也有望加速提升。

图 3：国产 AI 算力需求框架

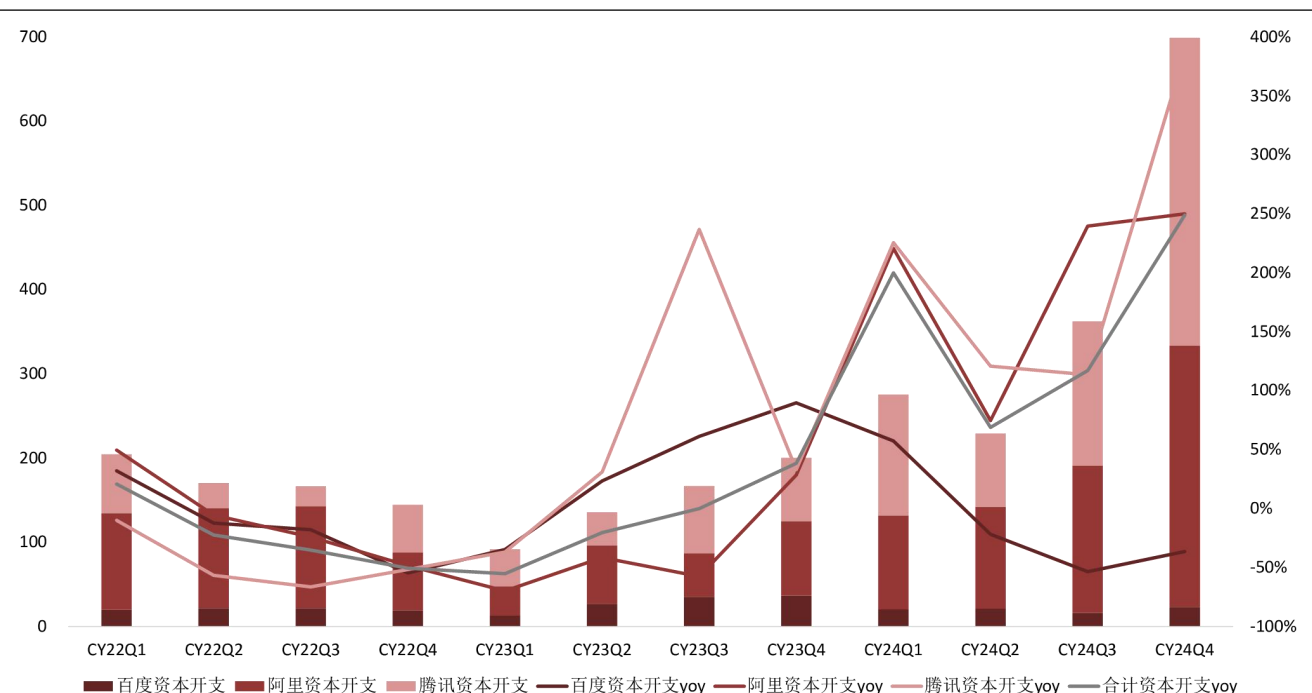


资料来源：山西证券研究所绘制

1.2 互联网：训练算力需求持续，推理侧需求将逐渐占据主导

国内头部互联网厂商仍持续投入大模型训练，全年资本开支保持快速增长。2024 年以来互联网大厂仍在持续迭代大模型，并在 DeepSeek R1 发布后纷纷加快布局推理模型：1）百度：2025 年 3 月 16 日正式发布新一代原生多模态基础大模型文心大模型 4.5 以及深度思考模型文心大模型 X1；2）腾讯：2025 年 2 月 27 日，腾讯发布新一代快思考模型混元 Turbo S，随后在 3 月 21 日，深度思考模型混元 T1 正式推出，创新采用 Hybrid-Mamba-Transformer 融合架构，推理性能业界领先的同时可比参数量下的解码性能高于行业 2 倍；3）阿里：今年 4 月 29 日将 Qwen2.5 升级至 Qwen3，其中旗舰模型 Qwen3-235B-A22B 在推理、指令遵循、工具调用、多语言能力等方面均实现大幅增强，同时部署成本显著下降，显存占用仅为性能相近模型的 1/3，并且 Qwen3 还创新性引入快思考模式与慢思考模式的混合设计；4）字节：今年 4 月 17 日豆包 1.5 深度思考模型正式发布，将视觉理解能力融入模型中，具备图片推理功能。受 AI 需求驱动，24 全年 BAT 合计资本开支同比增长 163%，预计 25 年互联网大厂的 AI 资本开支仍有望持续增加。

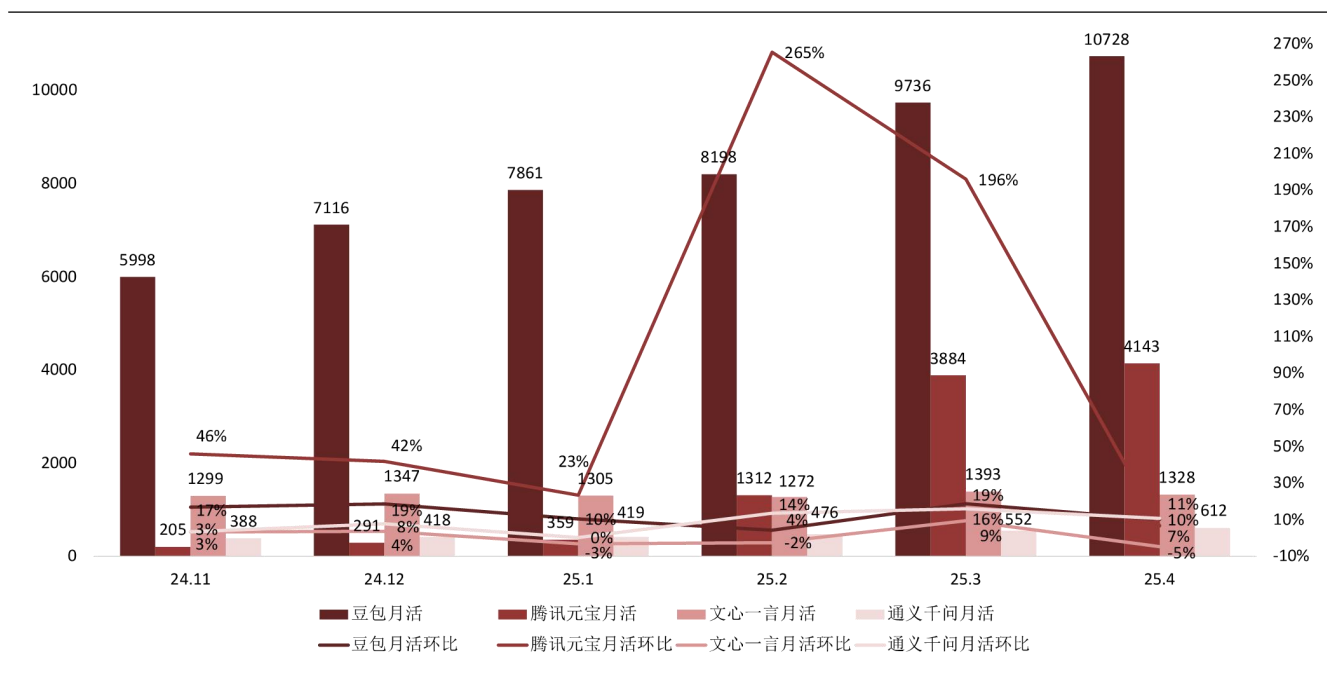
图 4：AI 驱动国内互联网大厂资本开支持续回暖（单位：亿元）



资料来源：腾讯 FY24Q4、FY24Q1、FY23Q3、FY22Q4、FY22Q2、FY21Q3、FY21Q1 财报，阿里 FY25Q3 财报电话会，阿里 FY24Q2-25Q2、FY22Q2-23Q1 财报，百度 FY24Q4、FY24Q2、FY23Q3、FY23Q1、FY22Q2、FY21Q4、FY21Q1 财报，山西证券研究所

头部互联网厂商 AI 应用流量持续增长，同时 DeepSeek 推动应用开发及部署成本下降，AI 应用正加速落地，推理侧需求持续旺盛，未来随着 AI Agent 应用不断成熟，推理需求有望进一步提升。近半年来，豆包、腾讯元宝、通义千问等互联网大厂的 AI 原生应用月活用户总体仍在持续增加，其中，豆包月活从 24 年 11 月的 5998 万持续增加，到今年 5 月已破亿，而腾讯元宝自接入 DeepSeek 以来月活流量快速增加，今年 2 月和 3 月的月活分别环比增长 265%和 196%，4 月月活已达到 4143 万。与此同时，DeepSeek R1 作为新一代国产开源推理模型，性能比肩国际顶尖大模型，同时训练及推理成本显著下降，将极大推动 AI 应用的开发部署，DeepSeek APP 月活流量也已从今年 1 月 3370 万增加到 4 月 9688 万。而展望未来，随着 A2A、MCP 等 Agent 相关技术的不断成熟，AI Agent 应用有望逐步爆发，从而进一步推升推理侧算力需求，今年 3 月推出的通用智能体 Manus 的 4 月月活用户数已达到 71 万。

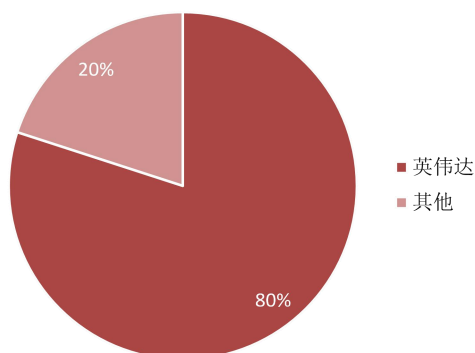
图 5：近半年来文心一言、豆包、通义千问、腾讯元宝的应用月活变化情况（月活单位：万）



资料来源：AI 产品榜，山西证券研究所

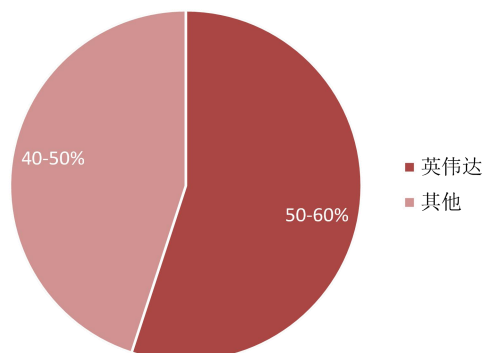
目前英伟达仍占据国内互联网厂商 AI 芯片采购的主要份额，以字节、阿里为代表的互联网大厂正不断加大国产 AI 芯片采购。一直以来国内互联网厂商大模型训练主要基于英伟达 A、H 等系列芯片，根据 Trendforce 在 2024 年初披露的数据，阿里、腾讯等国内头部互联网厂商采购的高端 AI 芯片中有 80%来自英伟达，国产芯片整体份额仍然较低。但随着去年以来美国政府逐渐收紧高端 AI 芯片出口，头部互联网厂商开始寻求英伟达 GPU 的替代方案以保障供应链安全，从当前头部互联网厂商采购情况来看，字节、阿里、百度已经在推进以昇腾为代表的国产 AI 芯片采购，根据路透社报道，24 年以来字节已订购超过 10 万颗昇腾 910B，同时阿里、百度、腾讯均开始批量采购。根据 Trendforce 预测，未来 5 年英伟达芯片在国内云厂商使用的高端 AI 芯片中的占比将降至 50-60%，相应地，国产 AI 芯片份额有望显著提升。

图 6：国内云厂商高端 AI 芯片中 80%来自英伟达
(24 年初)



资料来源：Trendforce，山西证券研究所

图 7：未来 5 年英伟达 AI 芯片在国内云厂商高端 AI 芯片中占比将降至 50-60%



资料来源：Trendforce，山西证券研究所

1.3 智算中心：全国产化项目已跑通，国产化率有望持续提升

《算力基础设施高质量发展行动计划》明确 AI 算力基础设施建设总体目标，各地方政府随即快速跟进，且部分省市提出国产算力占比要求。2023 年 10 月中央发布《算力基础设施高质量发展行动计划》，明确提出了未来 2 年内全国 AI 算力建设目标，包括到 2025 年算力规模超过 300EFLOPS，并且智算占比达到 35%。《行动计划》从顶层明确了全国范围内 AI 算力建设的目标和节奏，随后深圳、上海、江苏、北京等地方政府发布针对当地的“行动计划”，将全国的算力建设目标分解到当地，比如深圳市政府提出到 2025 年全市通用算力、智能算力、超算算力分别达到 14EFLOPS、25EFLOPS 以及 2EFLOPS。其中，部分省市补充了算力国产化率的相关目标，比如上海市提出到 2025 年新建智算中心国产算力芯片使用占比超过 50%的目标。

表 1：AI 算力建设纳入全国性及各省市的政策目标中

政策	发布时间	发布单位	主要内容
算力基础设施高质量发展行动计划	2023.10.8	工信部、中央网信办等 6 个部门	到 2025 年，算力规模超过 300EFLOPS，其中智能算力占比达到 35%
深圳市算力基础设施高质量发展行动计划（2024-2025）	2023.12.1	深圳市工信局	到 2025 年，深圳市通用算力达到 14EFLOPS，智能算力达到 25EFLOPS，超算算力达到 2EFLOPS
上海市智能算力基础设施高质量发展“算力浦江”智算行动实施方案（2024-2025 年）	2024.3.21	上海市通信管理局等 11 个部门	到 2025 年，上海市智算规模超过 30EFLOPS，占比达到总算力的 50%以上,并且新建智算中心国产算力芯片使用占比超过 50%
广东省算力基础设施高质量发展行动暨“粤算”行动计划（2024-2025 年）	2024.3.22	广东省通信管理局等 9 个部门	到 2025 年，广东省算力规模达到 38EFLOPS，其中智能算力占比达到 50%；建成智算中心 10 个
江苏省算力基础设施发展专项规划	2024.4.22	江苏省工信厅等 4 个部门	到 2030 年，全省数据中心机架规模达 120 万标准机架，在用总算力超过 50EFLOPS，智能算力占比超过 45%，智算中心数量突破 20 个；新建算力中心国产算力芯片使用占比达 70%以上
北京市算力基础设施建设实施方案（2024-2027 年）	2024.4.24	北京市经信局、北京市通信管理局	到 2025 年，北京市智算供给规模达到 45EFLOPS；到 2027 年，具备 100%自主可控智算中心建设能力

资料来源：银川深改公众号，中国政府网，广东网信办官网，广东省通信管理局官网，上海通信圈公众号，国际科技创新中心官网，北京市通信与互联网协会公众号，深圳发布公众号，江苏通信业公众号，江苏省通信管理局官网，北京市经济和信息化局官网，山西证券研究所

国内智算中心建设和运营主要采取政府主导、企业承建、联合运营的模式，考虑到智算中心对安全自主可控的需求，更倾向于国产 AI 算力采购。智算中心具有算力公共基础设施的定位，建设层面，政府主导下的政企合作模式占据较大比重，由企业具体承建智算中心，同时兼顾智算中心公共属性、技术安全、投资规模巨大等特殊要求，目前运营商是智算中心的主要承建方，其他承建方包括城投企业、IT 大厂、IDC 企业等，根据数据湖北对 2024 年前 7 个月的数据统计，三大运营商在 2024 年的中标项目中占据超过 40% 的份额。在建成后，智算中心服务模式包括租赁服务、数据服务、运营服务等，主要应用场景包括政府机构的智慧城市建设、大模型企业的模型训练、院校及科研院所的大数据处理和复杂模拟计算等。由于智算中心具有公共属性及技术安全等特点，政府通常会更倾向于采用国产优质品牌，涉及芯片、交换机、路由器等产品。

图 8：智算中心建设的参与方众多



资料来源：IDC，山西证券研究所

目前部分智算中心已实现全国产化，在标杆项目逐渐跑通之后，未来智算中心的 AI 算力国产化进程有望进一步加速。2024 年以来，全国大力投入智算中心建设，相关招投标项目持续增加，2024 全年智算中心相关项目中标公告已发布超 458 个，其中至少 61 个项目中标金额超亿元。与此同时，智算中心项目中国产 AI 算力占比正快速提升。23 年 12 月，国内首个采用全国产 AI 芯片的千卡智算中心正式落地，芯片供应商为摩尔线程；2024 年 4 月，中国联通建设的北京联通自主创新人工智能算力中心正式启用，提供全国产化 AI 基础软硬件的 300P 公共算力能力；在 2024 年 8 月份，中国移动建成中国移动智算中心（哈尔滨）并正式投产使用，成为全球运营商最大单集群智算中心，该智算中心部署超过 1.8 万张 AI 加速卡，AI 芯片国产化率达 100%，智算规模达 6.9EFLOPS。我们认为，随着采用全国产化 AI 芯片的标杆智算中心项目陆续跑通，未来智算中心的 AI 算力国产化率有望加速提升。

表 2：各地方仍不断推进智算中心建设

项目名称	项目地址	规划智算算力规模	项目进度	时间	投资金额
镇赉县智算中心	吉林省白城市	合计 4750P 一期 750P	一期基本完工	2025.5.6	6 亿元
通榆云针智算中心	吉林省白城市	5000P（混合精度）	一期展厅完工	2025.4.29	31 亿元
数聚城堡宜宾智算中心	四川省宜宾市	不低于 4000P	一期投产	2025.4.18	8 亿元
华池中科超级云智算中心	甘肃省庆阳市	一期 2000P 二期 2000P	开工	2025.4.18	一期 20 亿元 二期 8 亿元
中国移动湖北智算中心	湖北省鄂州市	860P 计算能力 185P 存储能力	开工	2025.4.9	3 亿元
阿勒泰市绿色低碳智算中心	新疆阿勒泰市	20000P	开工	2025.4.3	25 亿元
阿坝州低空经济智算中心	四川省阿坝州	12000P	开工	2025.3.31	50 亿元

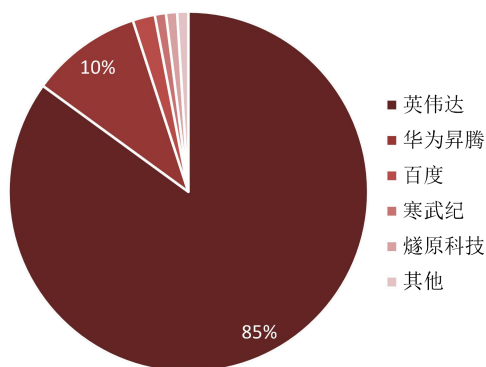
资料来源：IDC 圈，川观新闻，央广网，吉林省人民政府官网，中国拟在建项目网，白城市政府，四川在线，山西证券研究所

2. AI 芯片：国产替代空间广阔，第一梯队国产芯片正加速替代

2.1 英伟达仍占据国内市场主导，政策加码创造国产替代空间

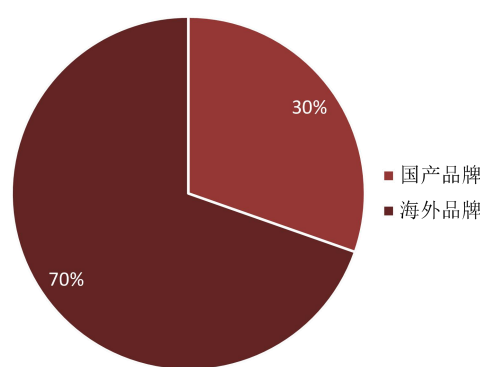
当前英伟达在国内 AI 芯片市场占据主导地位，而国产 AI 芯片份额仍处于较低水平。根据 IDC 数据，2022 年国内加速芯片出货量约 107 万张，其中英伟达占据 85% 的份额，华为昇腾占 10% 的份额，百度占 2% 的份额，寒武纪和燧原科技均占 1% 的份额。到了 2024 年，虽然国产 AI 芯片占比有所提升，但仍处于相对较低水平，根据 IDC 数据，2024 年国内加速芯片出货量超过 270 万张，其中本土品牌出货量已超过 82 万张，约占整个国内市场份额的 30%。

图 9：2022 年英伟达主导国内 AI 芯片出货量 85%



资料来源：IDC，芯智讯，199IT，山西证券研究所

图 10：2024 年国产 AI 芯片约占 30% 的份额



资料来源：IDC，山西证券研究所

近年来美国芯片出口管制政策持续加码，英伟达主力 AI 芯片出口中国大陆受限，且未来管制政策仍有进一步加码的可能。近年来美国不断加码对于高性能 AI 芯片的出口管制：1）2022 年 10 月更新出口管理条例，提出总处理性能（TPP）和 I/O 带宽传输速率两大判断指标，禁止 TPP 不小于 4800 且 I/O 带宽不小于 600GB/s 的 AI 芯片向中国等地区出口，导致英伟达 A100、H100 等高端芯片出口受限；2）2023 年 10 月美国商务部再次更新出口管理条例，新增性能密度（PD）指标，并设置 1.6、3.2 和 5.92 三个门槛，A800、H800、RTX4090 等芯片进入出口管制名单；3）2024 年美国再次对出口管制条例进行修订，将管制范围从 AI 芯片扩大到所有搭载 AI 芯片的电子设备。今年 4 月 9 日，英伟达收到美国政府通知，即向中国（包括香港和澳门）出口的 H20 芯片也需要获得许可，4 月 14 日，美国政府又通知英伟达该许可要求将无限延期。考虑到全球地缘政治风险的持续加剧，未来美国芯片出口管制政策仍存在进一步收紧的可能。

表 3：美国芯片出口管制政策持续加码

政策	颁布时间	颁布机构	生效时间	指标	管制范围	涉及 AI 芯片
出口管理条例 (更新)	2022.10.7	美国商务部	2022.10.21	总处理性能 (TPP)、I/O 带宽	$TPP \geq 4800$ 且 I/O 带宽 $\geq 600GB/s$	A100、H100
出口管理条例 (更新)	2023.10.17	美国商务部	2023.11.17	总处理性能 (TPP)、性能密 度 (PD)	禁止：1) $TPP \geq 4800$ 或 2) $TPP \geq 1600$ 且 $PD \geq 5.92$ 受限：1) $2400 \leq TPP < 4800$ 且 $1.6 \leq PD < 5.92$ 或 2) $TPP \geq 1600$ 且 $3.2 \leq PD < 5.92$	A800、H800、 RTX4090、 MI250、 MI250X、 MI300X
出口管理条例 (更新)	2024	美国商务部	2024.4.4	-	AI 芯片出口限制同样适用于 内载此类芯片的所有电子设备	-

资料来源：美国驻华大使馆官网，美国商务部官网，华尔街见闻，上观新闻，中国社会科学院工业经济研究所，《产业经济评论》2024 年 02 期《美对华半导体管制的趋势、实施要点与中国因应》，山西证券研究所

英伟达推出的中国特供版 AI 芯片 H20 自 2024 年以来在国内市场仍占据较高份额。虽然 H20 的训练算力只有 H100 的 15% 左右，但其在 CUDA 生态、通信互联等方面相较于其他 AI 芯片仍具备显著优势，其中英伟达 CUDA 生态为开发者提供了全面的并行计算框架和高效的卷积操作，支持高效的深度学习模型训练和推理，国内大模型开发者切换到其他生态具备较高的转换成本。同时，H20 保留了 H200 完整的 NVLink 和 PCIe 带宽性能，NVLink 互联带宽可达 900GB/s，PCIe 双向带宽可达 128GB/s，在组建大规模 GPU 集群方面优势显著；此外，相比于 H100，H20 还增配了 HBM3 显存容量（96GB）和带宽（4.0TB/s），能够加速模型推理过程中的 Decode 阶段，导致 H20 的推理性能优于 H100，甚至接近 H200。因此，短期内 H20 在国内市场仍占据较高份额。

图 11：英伟达特供版芯片 H20、L20、L2 对比

	HGX H20	L20 PCIe	L2 PCIe
GPU Architecture	NVIDIA Hopper	NVIDIA Ada Lovelace	NVIDIA Ada Lovelace
GPU Memory	96 GB HBM3	48 GB GDDR6 w/ ECC	24 GB GDDR6 w/ ECC
GPU Memory Bandwidth	4.0 TB/s	864 GB/s	300 GB/s
INT8 FP8 Tensor Core*	296 296 TFLOPS	239 239 TFLOPS	193 193 TFLOPS
BF16 FP16 Tensor Core*	148 148 TFLOPS	119.5 119.5 TFLOPS	96.5 96.5 TFLOPS
TF32 Tensor Core*	74 TFLOPS	59.8 TFLOPS	48.3 TFLOPS
FP32	44 TFLOPS	59.8 TFLOPS	24.1 TFLOPS
FP64	1 TFLOPS	N/A	N/A
RT Core	N/A	Yes	Yes
MIG	Up to 7 MIG	N/A	N/A
L2 Cache	60 MB	96 MB	36 MB
Media Engine	7 NVDEC 7 NVJPEG	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	2 NVENC (+AV1) 4 NVDEC 4 NVJPEG
Power	400 W	275W	TBD
Form Factor	8-way HGX	2-slot FHFL	1-slot LP
Interconnect	Pcie Gen5 x16: 128 GB/s NVLink: 900GB/s	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s
Availability	PS: Nov 2023 MP: Dec 2023	PS: Nov 2023 MP: Dec 2023	PS: Dec 2023 MP: Jan 2024

资料来源：派欧算力云公众号，山西证券研究所

2.2 国产 AI 芯片加速追赶，互联网及智算国产化采购快速提升

AI 芯片分为 GPGPU、FPGA 和 ASIC 三大阵营，国产芯片厂商以 GPGPU 和 ASIC 阵营为主。1) GPGPU：具有通用性强的特点，广泛应用于图像处理、AI 训练等领域，可支持多种数据类型的运算，英伟达、AMD 均采用 GPGPU 架构。国产代表厂商包括海光信息、壁仞科技、摩尔线程等，这类厂商核心技术团队大多具备 AMD、英伟达等国际芯片大厂背景；2) ASIC：具有定制化、专用性强的特点，通常应用于某些需求量较大的特定领域，目前海外的谷歌以及国内的百度、阿里等互联网大厂自研 AI 芯片大多采用 ASIC 架构，主要应用于自身业务场景，且华为昇腾、寒武纪等厂商也均采用 ASIC 架构；3) FPGA：属于半定制化芯片，具有功能可修改、一次性成本低但量产成本高的特点，适用于算法更新频繁或市场规模较小的专用领域，主要厂商包括海外的赛灵思、Altera 以及国内的复旦微电子、紫光同创等。

表 4：AI 芯片厂商可分为 GPGPU、FPGA 和 ASIC 三大阵营

	GPGPU	FPGA	ASIC
主要国产厂商	 中 科 海 光 摩尔线程 MOORE THREADS	 紫光同创 FUDAN MICRO	 Ascend Cambricon 寒 武 纪
主要海外厂商	 NVIDIA AMD	 XILINX ALTERA	 Google Intel
定制化程度	通用型	半定制化	定制化
研发成本	低	中	高
一次性成本	-	-	高
量产成本	高	高	低
编程语言	CUDA、OpenCL 等	Verilog 和 VHDL 等硬件描述语言、 OpenCL	-
功耗	大	中	小
主要优点	并行处理能力强	可重构，适应大规模并行操作	专用性强，能达成最佳的性能和功耗 平衡
主要缺点	功耗高，可修改不强	开发时间较长，顺序操作性能差	开发时间长，不可修改

资料来源：《Automatic Parallelization for Heterogeneous Embedded Systems》，《New FPGA Design Tools and Architectures》，海光信息招股说明书，AspenCore，EET 电子工程专辑，The Next Platform，山西证券研究所

近 10 年来国产 AI 芯片厂商快速涌现，同时互联网大厂均开始推出自研 AI 芯片以满足自身业务需求。其中：1) 互联网大厂：包括阿里平头哥、百度昆仑芯等，互联网大厂自研芯片多强调推理性能，主要应用于满足自身业务场景，包括淘宝产品搜索及推荐、百度智能搜索、腾讯云游戏等；2) 华为昇腾：2018 年华为发布昇腾 310 和昇腾 910，正式切入 AI 芯片领域，其中，昇腾 310 主打终端低功耗 AI 场景，采用 12nm 工艺，2018 年便实现量产；昇腾 910 采用 7nm 工艺，可用于 AI 训练，于 2019 年正式上市。目前昇腾 910 系列的主力产品 910B 综合性能处于国产 AI 芯片第一梯队，并且 910C 正在加速生产中，而华为在 4 月 10 日合作伙伴大会上又发布了下一代的昇腾 920 芯片；3) AI 芯片初创公司：包括海光信息、寒武纪、壁仞科技、摩尔线程、沐曦等，这类厂商于近 10 年成立，核心团队大多拥有英伟达、AMD 等国内外芯片或 IT 大

厂的技术背景，其中部分厂商自成立以来便专注于 AI 芯片研发，也有部分厂商此前从事 CPU 或其他芯片研发，近年开始切入 AI 芯片领域。

表 5：近年来国产 AI 芯片厂商快速涌现

厂商	成立时间	创始团队主要背景	地点
华为昇腾	2018（首款芯片发布）	华为等	深圳
海光信息	2014	中科曙光、国科控股等	北京
寒武纪	2016	中科院计算所等	北京
壁仞科技	2019	商汤科技、英伟达、AMD、英伟达、阿里云等	上海
摩尔线程	2020	英伟达、AMD 等	北京
沐曦	2020	AMD 等	上海
燧原科技	2018	AMD、锐迪科等	上海
平头哥	2018	阿里等	杭州
昆仑芯	2018	百度等	北京

资料来源：澎湃新闻，国际电子商情，海光信息官网，Wind，海光信息招股说明书，海光信息 2024 年报，寒武纪官网，中国科学院官网，壁仞科技官网，电子工程专辑，摩尔线程官网，证券时报网，沐曦官网，燧原科技官网，平头哥官网，昆仑芯官网，山西证券研究所

以华为昇腾、海光信息、寒武纪、昆仑芯为代表的国产 AI 芯片厂商正加速追赶英伟达，目前华为昇腾 910B 已经基本可对标 A100，而新一代的 910C 也有望成为 H100 的有力竞争对手。目前华为在人工智能领域的量产芯片有昇腾 310 与昇腾 910 两款产品，华为在 2019 年推出昇腾 910 芯片，并在 2023 年推出了升级版昇腾 910B。昇腾 910B 作为华为主力 AI 训练芯片，在单卡性能上可对标英伟达 A100，其 FP16 稠密算力达到 320 TFLOPS，超越 A100 的 312 TFLOPS，与英伟达 H20 相比，910B 在计算能力和能效比方面均表现出色，但在显存配置、GPU 互联带宽方面目前仍存在差距，同时正在进行生产的 910C 的 FP16 稠密算力达到 780TFLOPS，内存带宽为 3.2TB/s，综合性能接近 H100 的 80%。此外，华为基于 384 块 910C 发布了 CloudMatrix 384 超节点，在系统层面上与 GB200 NVL72 展开竞争，CloudMatrix 384 包含 12 个计算机架和 4 个交换机机架，FP16 稠密算力达 300PFLOS，是 GB200 NVL72 的 1.7 倍，总内存容量和内存带宽也分别是 GB200 NVL72 的 3.6 倍和 2.1 倍，目前已通过昇腾云正式商用。

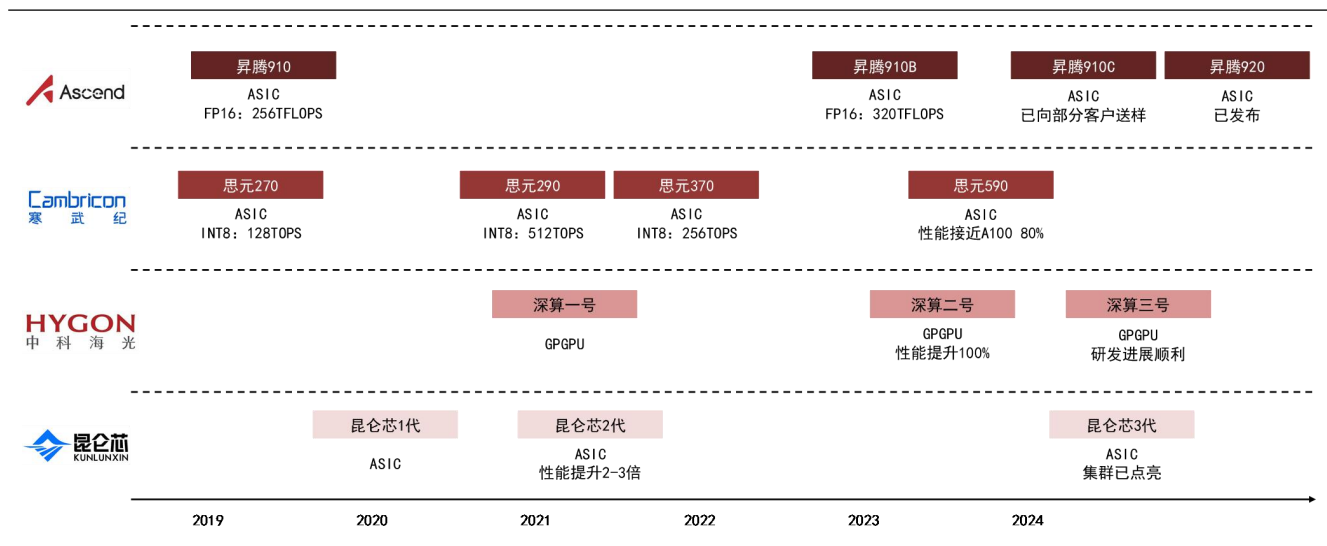
表 6：英伟达、AMD 及主要国产 AI 芯片对比

公司	芯片	算力 (TFLOPS)			技术	显存配置		GPU 互联		TDP (W)
		FP32	FP16	INT8		容量 (GB)	带宽 (TB/s)	技术	带宽 (GB/s)	
英伟达	B200	1100	2250	4500	-	192	8	NVLink	1800	1000
	H100	494.5	989.5	1979	HBM3	80	3.35	NVLink	900	700
	A100	156	312	624	HBM2e	40/80	1.555/2.039	NVLink	600	400
	H20	37	74	148	HBM3	96	4	NVLink	900	400
AMD	MI300X	653.7	1307.4	2614.9	HBM3	192	5.3	Infinity Fabric	896	750
华为	昇腾 910C	-	780	-	HBM 技术	128	3.2	-	-	-
	昇腾 910B	-	320	640	-	-	-	-	-	310
海光信息	深算一号	-	-	-	HBM2	32	1.024	xGMI	184	350
寒武纪	思元 370-X8	24	96	256	LPDDR5	48	0.6144	MLU-Link	200	250
昆仑芯	R200	32	128	256	GDDR6	16/32	0.512	-	-	150
平头哥	含光 800	-	205	825	-	-	-	-	-	276
摩尔线程	MTTS4000	25	100	200	-	48	0.768	MTLink	240	450

资料来源：Trendforce, Semianalysis, 海光信息招股说明书, 寒武纪官网, 电子工程专辑, 皇华电子元器件 IC 供应商, 平头哥官网, 摩尔线程官网, 英伟达官网, 派欧算力云公众号, AMD 官网, 山西证券研究所

寒武纪、海光信息、昆仑芯处于国产 AI 芯片第一梯队，其新一代主力产品思元 590 芯片、深算三号、昆仑芯 3 代未来有望对英伟达 H20 形成替代。1) 寒武纪：2020 年寒武纪推出首款 7nm 训练芯片思元 290，INT8 算力达到 512 TOPS，接近 A100 的 624 TOPS（非稀疏化），2023 年年底发布了 590 系列，综合性能进一步提升，接近 A100 的 80%；2) 海光信息：2021 年推出深算一号，2023 年三季度推出的深算二号性能相对于深算一号提升约 100%，且下一代深算三号进展顺利；3) 昆仑芯：2020 年昆仑芯 1 代实现量产，主要用于百度搜索引擎、小度等业务，2021 年量产的昆仑芯 2 代较 1 代性能提升 2-3 倍，并且百度已点亮基于新一代芯片昆仑芯 3 代组件的万卡集群。此外，摩尔线程、燧原科技、天数智芯等国产 AI 芯片厂商已有产品上市，并在部分场景中实现了批量应用。

图 12：寒武纪、海光信息、昆仑芯等厂商的新一代主力产品有望对英伟达 H20 形成替代

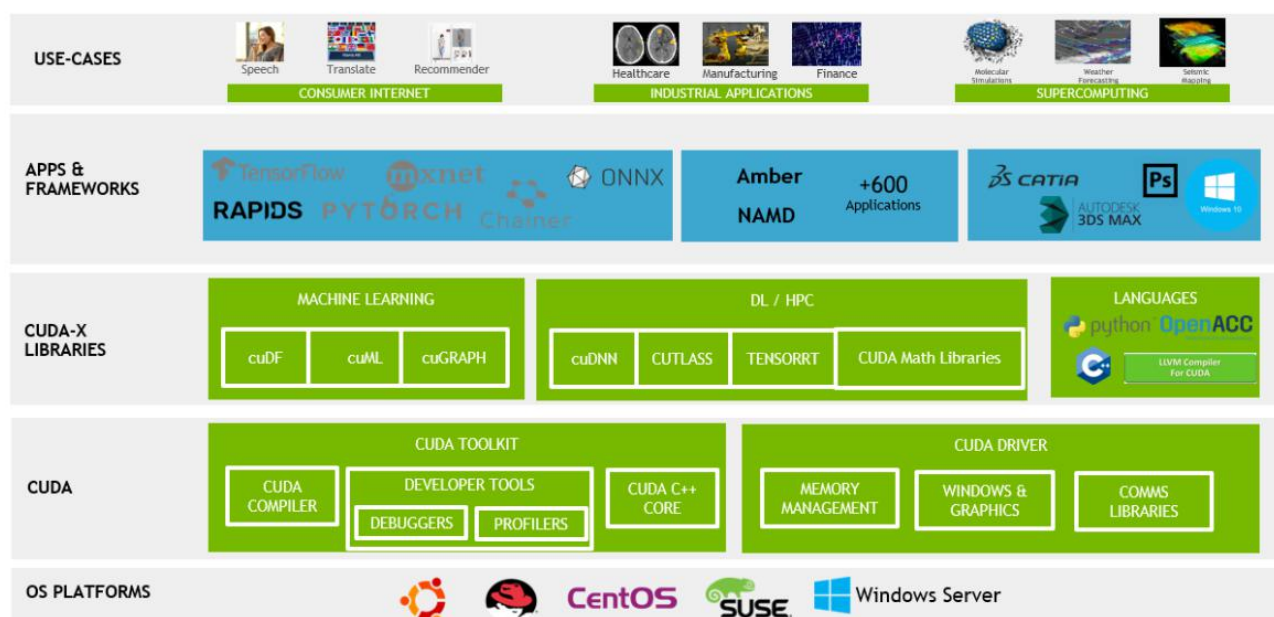


资料来源：寒武纪 2020 年报，寒武纪官网，ofweek，海光信息 2023 年 9 月 20 投资者调研纪要，海光信息招股说明书，昆仑芯官网，百度智能云官网，Trendforce，搜狐，雅虎，新京报，昇腾官网，Tom's Hardware，山西证券研究所

2.3 国产厂商基于兼容及自研两大路径逐渐突破 CUDA 生态壁垒

英伟达所打造的 CUDA 生态是目前国产 AI 芯片实现国产化替代的主要障碍。2006 年英伟达首次发布 CUDA 软件平台，随后进行市场推广，快速渗透人工智能开发者。同时英伟达持续投入软件研发，针对用户反馈的问题进行优化，CUDA 平台的软件栈丰富度以及好用程度不断提升，目前 CUDA 已全面支持 TensorFlow、PyTorch 等 AI 框架。通过长期市场培育，CUDA 框架已成为事实上的行业标准，根据英伟达在 2024 年 8 月底披露的数据，CUDA 已拥有 510 万开发人员，市场上共有 3700 个 CUDA 应用程序，涵盖深度学习、科学计算、图像处理等各个领域，同时 CUDA 的累计下载量达到 5300 万，考虑到绝大部分国内外大模型或 AI 应用开发者均在 CUDA 平台上进行软件研发，因此切换到新框架需要付出巨大的转换成本。

图 13：英伟达 CUDA 架构



资料来源：英伟达官网，山西证券研究所

国产 AI 芯片厂商目前主要通过兼容 CUDA 以及自研两种方式构建生态。1) 海光信息、壁仞科技、摩尔线程等厂商采用兼容 CUDA 生态的方式，一方面，这类厂商在硬件层面上大多采用与英伟达类似的 GPGPU 架构，另一方面，这类厂商投入研发具备较高转换质量的编译器，用于 CUDA 代码转译，因此能够降低 CUDA 迁移难度并减少效率损失。短期来看，通过兼容 CUDA 构建生态的优势在于能够实现 CUDA 应用的低成本迁移，从而使开发者在不改变使用习惯的情况下逐渐过渡为使用国产芯片；2) 华为昇腾、寒武纪等厂商致力于构建自主生态，这类厂商会针对自身 AI 芯片的特性通过自研构建软件栈，以通过软硬件协同更好发挥芯片性能，同时自主可控程度也更高。但在短期内，自建生态难度较大，这类厂商也会通过编译器实现对 CUDA 的兼容。而由于这类厂商多采用 ASIC 架构，CUDA 迁移难度相对较高，在一定程度上会导致更大的性能损失。

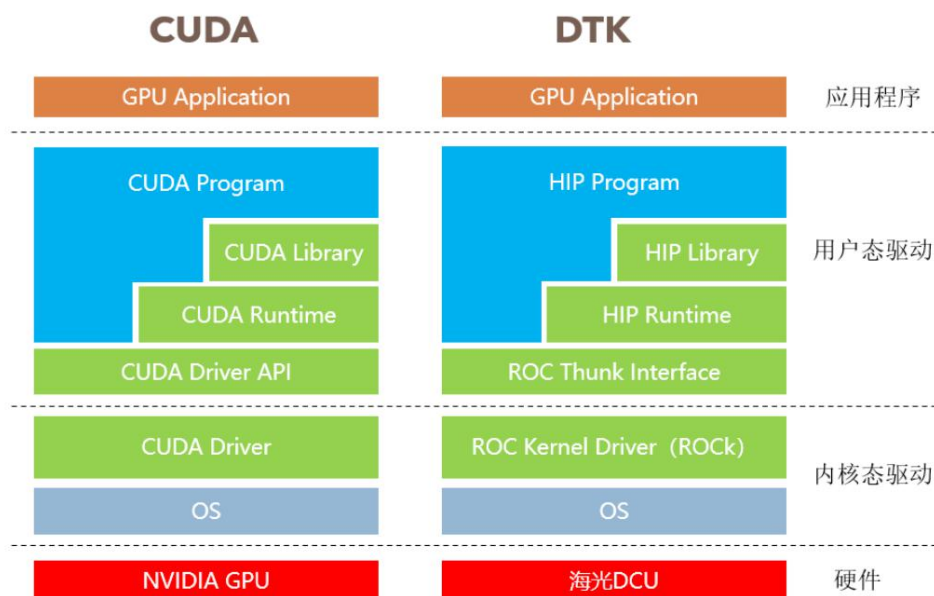
表 7：国产 AI 芯片厂商主要通过兼容 CUDA 以及自研两种方式构建生态

厂商	芯片架构	基本计算单元	软件平台	包含组件
华为昇腾	达芬奇	AI core	CANN	AscendCL 昇腾计算语言、AOL 算子加速库、Ascend C 算子开发语言等
海光信息	-	-	DTK	HIP 异构编程接口、hipBLAS 库、hipsparse 库等
寒武纪	MLUarch03/ MLUv02	-	Neuware	CNNL 算子库、CNCL 通信库、BANG 编程语言等
摩尔线程	MUSA	MUSA Processor	MUSA AI	TensorX 推理引擎、MUSA Toolkits（内置运行时、编译器、GPU 加速计算库）、计算库、神经网络加速库

资料来源：Trendforce，昇腾官网，飞桨，海光信息官网，寒武纪官网，摩尔线程官网，山西证券研究所

海光信息 DTK 开放软件平台封装了 AMD 的 ROCm 相关组件，能够高度兼容 CUDA 生态，同时，海光大量采用国际主流开源方案，较大程度上降低了生态拓展难度。AMD 的 ROCm 框架借鉴了 CUDA 的许多设计理念，同时经过多年兼容经验积累，在软件库、编译器、API 接口等主要模块上与 CUDA 的相似度非常高，因而 ROCm 对 CUDA 具有高度兼容性。而海光 DTK 软件平台封装了 ROCm 生态相关组件，并基于 DCU 硬件进行优化并提供完整的软件工具链，能够充分发挥 ROCm 高度兼容 CUDA 的优势，以更低成本将 CUDA 上运行的应用迁移到海光平台上运行。同时，海光 DTK 充分利用国际开源社区，MIOpen、Eigen、RCCL 等软件库均为目前的主流开源方案，一方面能够降低软件开发工作量，提升用户吸引力，另一方面也方便用户进行二次开发，从而提升用户体验。

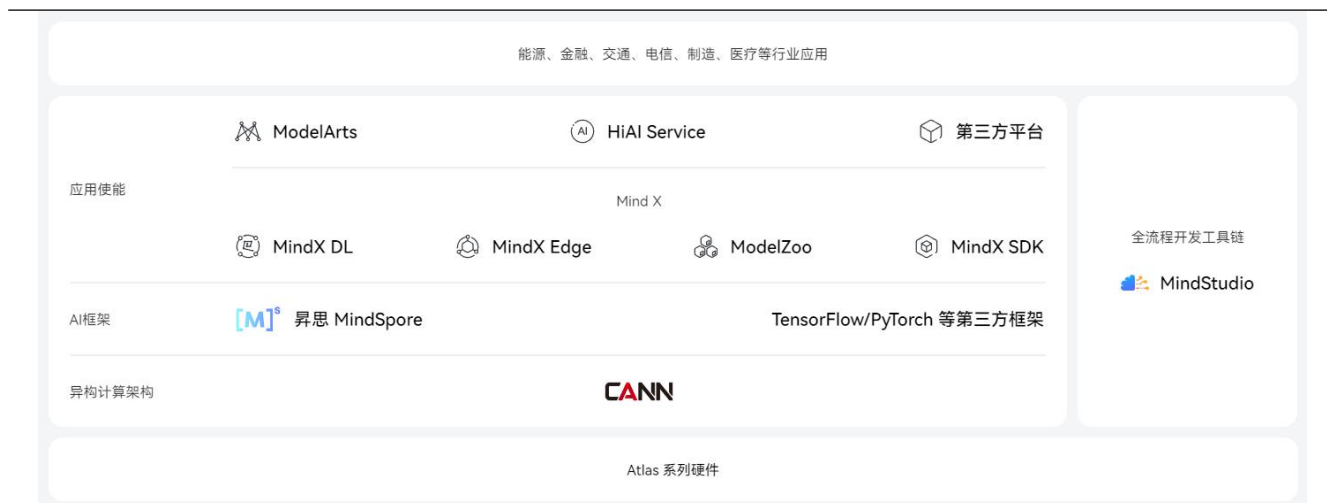
图 14：海光信息 DTK 高度兼容 CUDA



资料来源：趋动科技公众号，山西证券研究所

华为昇腾自研构建了全栈 AI 计算基础设施、行业应用及服务，包括昇腾系列硬件、CANN 异构计算架构、MindSpore AI 计算框架等，打造完整的国产化 AI 生态，并通过软硬件深度协同，充分发挥芯片性能。其中，CANN 是华为专为昇腾 AI 处理器设计的核心软件层，对标英伟达 CUDA 架构。CANN 于 2018 年 9 月推出，目前已迭代至 8.0.0 alpha001 版本，提供了超过 1400 个硬件亲和的高性能算子，可覆盖主流 AI 框架的算力加速需求，同时开放了自定义算子开发能力，满足开发者算法创新需求。CANN 具有三大优势，包括 1) 提供用户友好型 API，易于开发，2) 计算效率、数据处理效率、训练效率较高，3) 支持云边端全场景；而自研的 AI 框架昇思（MindSpore）在国内开发者中的使用率排名第四，仅次于三大第三方框架。目前昇腾生态已初具规模，收录了紫东太初、智谱天启等大模型，适用场景涵盖多模态等，同时，昇思社区用户已超过 1220 万，收藏项目超过 2.86 万。

图 15：昇腾 AI 计算体系更完整



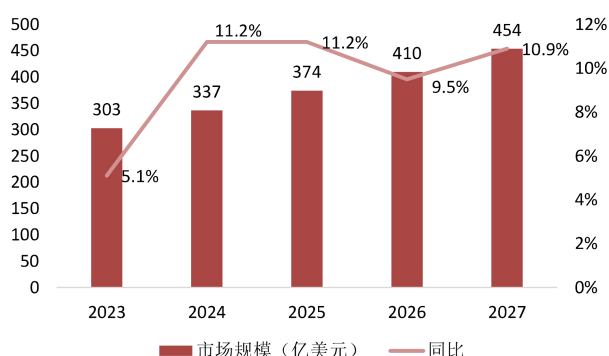
资料来源：昇腾官网，山西证券研究所

3. AI 服务器：AI 驱动服务器市场格局变化，昇腾份额持续提升

3.1 AI 服务器高速增长，市场规模迅速追赶通用服务器

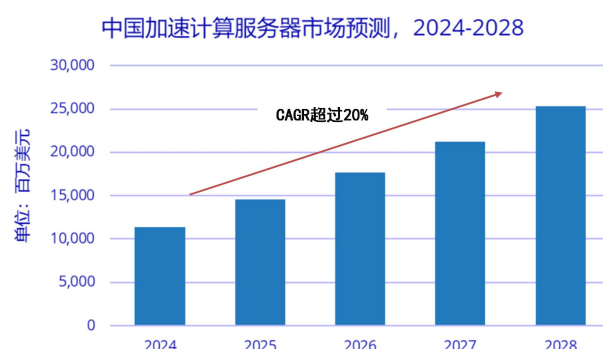
互联网、运营商、金融等各行业用户持续加大 AI 算力投入，国内 AI 服务器市场规模整体保持快速增长。根据 IDC 数据，未来 4 年国内服务器市场总规模将保持 10% 左右的增长。而国内 AI 服务器市场增速将明显快于通用服务器，2024 上半年中国加速服务器市场规模达到 50 亿美元，同比增长 63%。在各类 AI 服务器中，GPU 服务器依然占据主导，市场规模达 43 亿美元，占比达 86%，同时 NPU、ASIC 和 FPGA 等非 GPU 加速服务器以同比 182% 的增速达到近 7 亿美元市场规模。根据 IDC 预测，到 2028 年中国加速服务器市场规模将达到 253 亿美元，2024-2028 年均复合增速超过 20%，其中以 ASIC 为代表的非 GPU 服务器市场规模占比也将提升至接近 50%。

图 16：国内服务器总市场规模及增速



资料来源：IDC，山西证券研究所

图 17：国内 AI 服务器市场规模及增速



资料来源：IDC，山西证券研究所

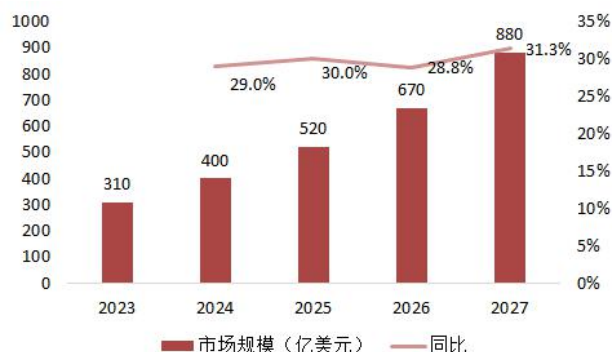
从全球看，随着大型 CSP 以及企业客户不断增加 AI 算力支出，全球 AI 服务器市场快速成长，市场规模追赶通用服务器。根据 The Next Platform 引用 IDC 数据，到 2027 年全球服务器市场总规模将达到 1891 亿美元，2023-2027 年 CAGR 为 10.2%。而根据 Statista 数据，未来 4 年全球 AI 服务器市场将保持 30% 左右的高速增长，预计到 2027 年市场规模将达到 880 亿美元，并逐步与通用服务器市场规模持平。

图 18：全球服务器总市场规模及增速



资料来源：The Next Platform, IDC, 山西证券研究所

图 19：全球 AI 服务器市场规模及增速

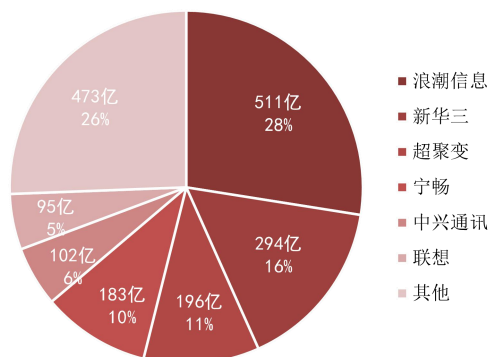


资料来源：Statista, 山西证券研究所

3.2 AI 上游芯片及下游需求驱动服务器市场竞争格局剧变

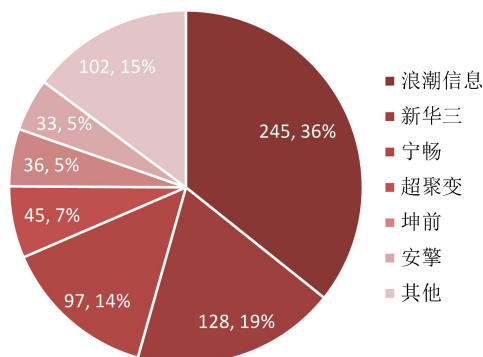
浪潮信息和新华三等服务器市场领军占据国内 AI 服务器市场的主导，同时以超聚变和华鲲振宇为代表的昇腾系服务器厂商份额持续提升。根据 IDC 数据统计，2023 年，头部服务器厂商浪潮信息、新华三、超聚变、宁畅在国内 x86 服务器市场份额分别为 28%、16%、11%、10%，同时 AI 服务器市场中，浪潮信息、新华三、超聚变、宁畅份额分别为 36%、19%、7%、14%。其中，浪潮信息在通用服务器和 AI 服务器均位列国内市场第一，公司凭借 JDM 模式（Joint Design Manufacture），深度参与到服务器的设计、研发和交付流程，占据了国内互联网客户云服务器和 AI 服务器的主要份额；而新华三在国内通用和 AI 服务器市场的份额均仅次于浪潮，公司背靠紫光股份，在政府、运营商、央企等行业具有较强优势；同时 23 年超聚变在国内 x86 服务器 AI 服务器市场份额分别排名三和第四，目前昇腾 910B 已经成为国内客户在国产 AI 训练芯片能够替代 A100 的主要选择，随着昇腾出货量的快速增长，以超聚变和华鲲振宇为代表昇腾系服务器厂商的份额也随之持续提升。除了传统服务器厂商及昇腾系厂商之外，以华勤技术为代表的本土 ODM 厂商正在切入到 AI 服务器市场，凭借在消费电子领域积累的研发体系和供应链优势，公司 22 及 23 年的服务器业务均达到数倍的增长。

图 20：2023 年国内 x86 服务器市场规模及份额



资料来源：IDC，云头条，山西证券研究所

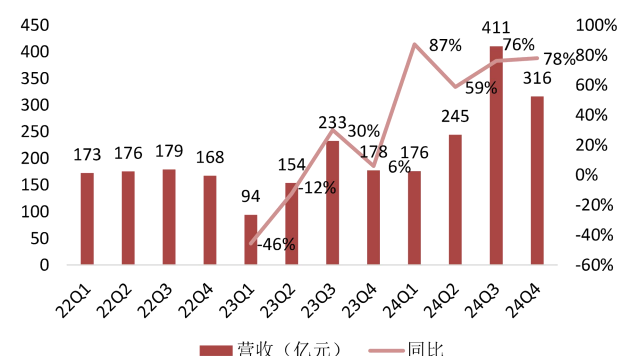
图 21：2023 年国内 AI 服务器市场规模及份额



资料来源：IDC，云头条，山西证券研究所

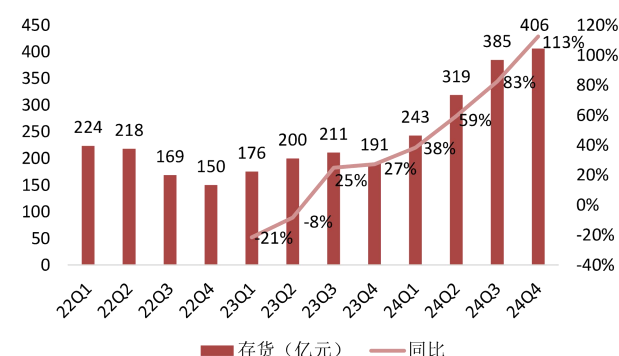
2024 年以来，以浪潮信息为代表的头部服务器厂商收入持续高增验证 AI 算力景气度。2024 年 H20 在头部互联网厂商及智算中心用户 AI 芯片采购中占据较高份额，出货量持续增长，其中，截止到 2024 年 9 月初，字节的 H20 订单已追加至 32-33 万张，并已交付 14-15 万张，对应服务器环节，以浪潮信息为代表的头部服务器厂商将率先受益。24Q3 和 24Q4，浪潮信息的营收分别为 411 亿元和 316 亿元，分别同比增长 76%和 78%，同时，24 年以来浪潮的存货金额持续增加，24Q4 存货金额达到 406 亿元，创近 12 个季度新高，预示着公司对未来几个季度的经营发展充满信心。与此同时，以华勤技术为代表的国产头部 ODM 厂商也开始切入到服务器市场，通过优秀的研发能力和供应链整合能力，市场竞争力快速提升，2023 年公司数据业务实现营收 3 倍增长，并推出首款 AI 服务器产品，2024 上半年数据业务以及其中的服务器板块均同比增长 150%，而下半年 AI 服务器也进入发货高峰期。

图 22：22Q1-24Q4 浪潮信息营收变化情况



资料来源：Wind，山西证券研究所

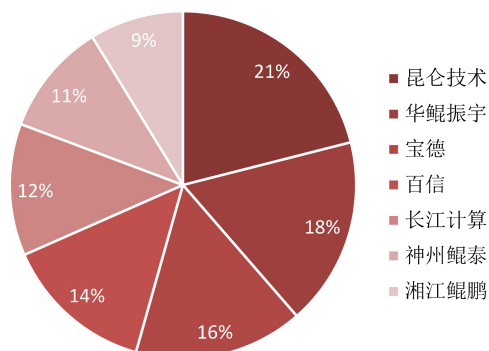
图 23：22Q1-24Q4 浪潮信息存货变化情况



资料来源：Wind，山西证券研究所

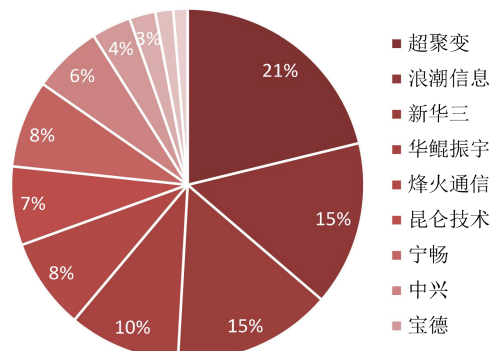
2024 年以来昇腾 AI 芯片放量带动昇腾系服务器厂商市场份额快速提升，运营商及智算中心采购项目持续验证。三大运营商作为国内智算中心最大的建设主体之一，在 2023-2024 发标多个大型 AI 服务器采购项目，其中昇腾系厂商占据运营商 AI 服务器采购主要份额，具体来看：1）中国移动：2023 年以来两次人工智能服务器集采规模合计超万台。2024 年 4 月 18 日，中国移动公布了《2024 年至 2025 年新型智算中心采购招标公告》，涉及 7994 台人工智能服务器及配套产品，从中标厂商来看，华为昇腾合作伙伴占据绝对主导，标包 1 的中标厂商均为昇腾合作伙伴。此前中国移动还发布了 2023 年至 2024 年新型智算中心（试验网）集采项目，12 个标包对应 AI 训练服务器采购量总计达到 2374 台；2）2024 年 3 月底，中国联通发标采购 2503 台 AI 服务器，目前在四位中标候选人中有三位是昇腾合作伙伴，包括昆仑技术、宝德、长江计算；3）中国电信：在 2023 年 10 月也采购了 4175 台 AI 服务器和 1182 台交换机，4 个服务器标包合计超过 80 亿元，共有超聚变、华鲲振宇、烽火通信、宁畅、中兴、新华三等 11 家厂商中标，其中昇腾系占比达 67%。

图 24：中国移动 2024 年至 2025 年新型智算中心采购项目中标份额情况（标包 1）



资料来源：IDC 圈，山西证券研究所

图 25：中国电信 AI 算力服务器（2023-2024 年）集中采购项目中标份额情况

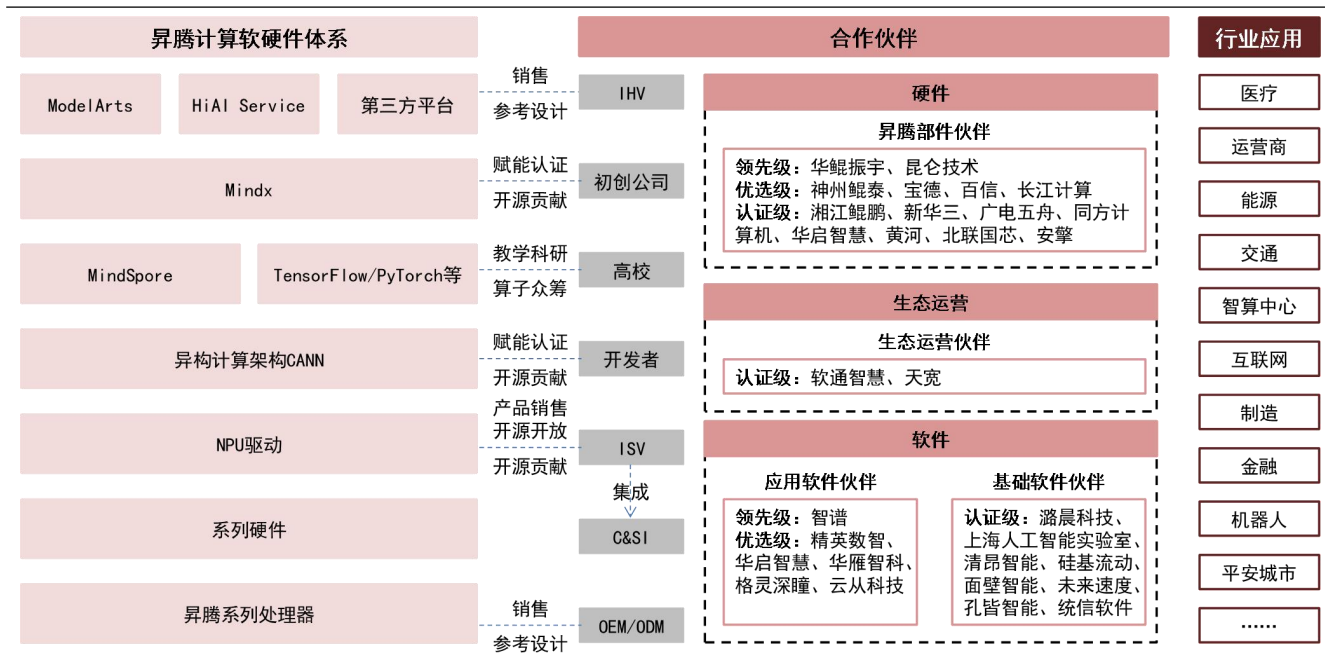


资料来源：芯智讯，山西证券研究所

3.3 昇腾系计算产业生态成型，两大领先级部件伙伴将占主要份额

目前昇腾 AI 芯片连同昇腾服务器、应用软件、生态运营等合作伙伴形成了完整的昇腾计算产业生态体系，其中共有 4 大类合作伙伴。1) 昇腾部件伙伴，即拥有自有品牌产品，并能在昇腾产品基础上二次开发或加工生产，销售至最终用户的合作伙伴，包括华鲲振宇、昆仑技术、神州鲲泰等；2) 应用软件伙伴，主要对接昇腾产品二次开发并销售自有知识产权的应用程序、软件、垂直细分应用，包括智谱、精英数智、华启智慧、云从科技、格灵深瞳等；3) 基础软件伙伴，主要基于昇腾开发算子、加速库、模型套件、工具链等产品，以及基于开源模型提供微调、后训练、优化加速等专业能力，包括硅基流动、面壁智能、潞晨科技等；4) 生态运营伙伴，即运营指定区域内智算中心、生态创新中心或创新实验室的合作伙伴，包括软通智慧、天宽等。

图 26：昇腾计算产业生态中主要有 4 类合作伙伴



资料来源：昇腾官网，《昇腾计算产业白皮书》，山西证券研究所绘制

昆仑技术（超聚变）和华鲲振宇（四川长虹）两大领先级部件伙伴将占据昇腾服务器较大份额。今年华为升级了合作伙伴政策，在加速增长型伙伴中新增产品增值伙伴，主要分为整机伙伴、部件伙伴和白牌伙伴。目前华鲲振宇和昆仑技术是唯二的领先级昇腾部件伙伴，并均在此前荣获“华为昇腾万里伙伴计划-战略级整机硬件伙伴”认证，将最大程度受益于昇腾产业的崛起，其中，华鲲振宇还是昇腾战略级 OEM 硬件伙伴，并且据公司高管透露，华鲲振宇在鲲鹏和昇腾整机生态中出货量位于第一。

表 8：华鲲振宇（四川长虹）和昆仑技术（超聚变）是唯二的领先级昇腾部件伙伴

类别	级别	合作伙伴
昇腾部件伙伴	领先级	华鲲振宇（四川长虹持股 3%）、昆仑技术（超聚变持股 100%）
	优选级	神州鲲泰（神州数码持股 100%）、宝德、百信、长江计算（烽火通信持股 71%）
	认证级	湘江鲲鹏（拓维信息持股 100%）、新华三（紫光股份持股 81%）、广电五舟（广电运通持股 16%）、同方计算机（软通动力持股 100%）、华启智慧（大华股份持股 100%）、黄河、北联国芯、安擎

资料来源：企查查，Wind，紫光股份 2024 年三季报，山西证券研究所

4. 盈利预测及投资建议

互联网和智算中心两大下游国产 AI 算力需求将快速增长，其中，24 年以来互联网大厂算力需求仍以训练侧为主，随着 DeepSeek 的突破大幅降低应用端门槛以及 Agent 技术和产品日趋成熟，推理侧需求将逐渐成为 AI 算力支出的主驱动力，同时，互联网大厂正加速推进国产 AI 芯片采购；2024 年以来各地方智算中心项目招投标数量仍在持续增加，同时多个全国化标杆项目已跑通，未来智算中心的 AI 算力国产化率有望加速提升。

1) 芯片端：考虑美国芯片出口管制政策存在进一步加码的可能，将倒逼国内客户加快推进国产 AI 芯片采购。目前第一梯队国产芯片厂商正加速突破性能、产能、生态瓶颈，在性能上，其新一代主力产品未来有望对 H20 形成替代；在产能上，国产厂商正加快解决国内晶圆代工等供应链问题；在生态上，国产厂商通过兼容及自研突破 CUDA 生态壁垒。AI 芯片环节推荐海光信息，建议关注寒武纪。

2) 服务器端：2024 年以来，H20 仍在国内互联网大厂及智算中心采购中占据较高份额，推动以浪潮信息为代表的服务器领军业绩高速增长。同时，随着昇腾出货量的快速增长，以超聚变和华鲲振宇为代表昇腾系服务器厂商市场份额随之持续提升。服务器环节推荐浪潮信息，建议关注华勤技术、四川长虹、神州数码、拓维信息。

表 9：主要公司盈利预测和估值情况

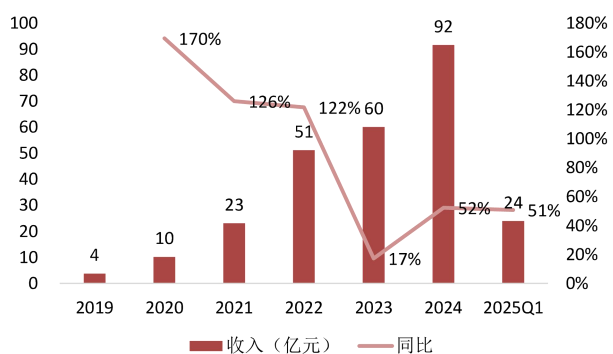
证券代码	证券名称	收盘价	EPS (25E)	EPS (26E)	EPS (27E)	PE (25E)	PE (26E)	PE (27E)
688041.SH	海光信息	136.1	1.6	2.3	3.1	85.8	59.0	43.5
688256.SH	寒武纪	651.0	2.9	5.3	8.0	225.2	121.9	81.6
000977.SZ	浪潮信息	50.0	2.5	2.9	3.0	20.0	17.4	16.6
603019.SH	中科曙光	61.9	1.5	1.8	2.0	40.5	34.2	30.3
000938.SZ	紫光股份	24.3	0.8	1.0	1.1	31.5	25.4	21.8
0992.HK	联想集团	9.3	0.9	0.9	1.1	10.8	9.9	8.4
600839.SH	四川长虹	10.1	-	-	-	-	-	-
000034.SZ	神州数码	38.2	1.9	2.3	2.7	20.6	16.9	14.1
002261.SZ	拓维信息	30.8	0.1	0.2	-	285.4	200.1	-
600498.SH	烽火通信	20.4	0.9	1.0	1.2	23.8	19.7	16.8
301236.SZ	软通动力	53.9	0.4	0.6	0.8	130.3	94.0	70.0
002152.SZ	广电运通	11.7	0.4	0.5	0.6	27.0	23.3	20.5
603296.SH	华勤技术	68.4	3.6	4.3	5.2	19.1	15.8	13.0

资料来源：Wind，山西证券研究所；注：采用 2025 年 5 月 26 日的收盘价，海光信息、浪潮信息、中科曙光的预测 EPS 基于自己的估值模型，其他公司预测 EPS 采用 Wind 一致预期数据，其中联想集团预测时间区间为 2025-2027 财年，收盘价和 EPS 单位均为人民币元

海光信息：公司将充分受益于 AI 芯片国产化替代加速。随着美国芯片出口管制政策持续加码，英伟达主力 AI 芯片出口中国大陆受限，以海光为代表的国产 AI 芯片市场份额预计将进一步提升。目前，海光 DCU 位于国产 AI 芯片第一梯队，在产品迭代上，公司新一代主力产品深算三号未来将成为华为昇腾的有力竞争者，同时，海光 DCU 系列产品以 GPU 架构为基础，兼容通用的“类 CUDA”环境，因此 CUDA 上运行的应用能够也低成本迁移到海光平台上运行，公司在软件生态上具有优势。

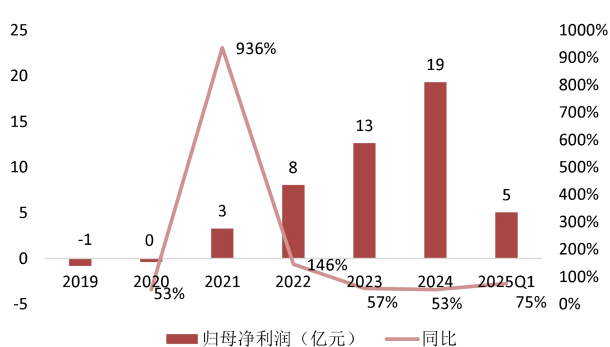
风险提示：客户集中度较高的风险；供应链风险；主要原材料价格上涨的风险；新产品研发进展不及预期。

图 27：2019-2025Q1 海光信息收入及增速



资料来源：Wind，海光信息 2025 年一季报，山西证券研究所

图 28：2019-2025Q1 海光信息归母净利润及增速

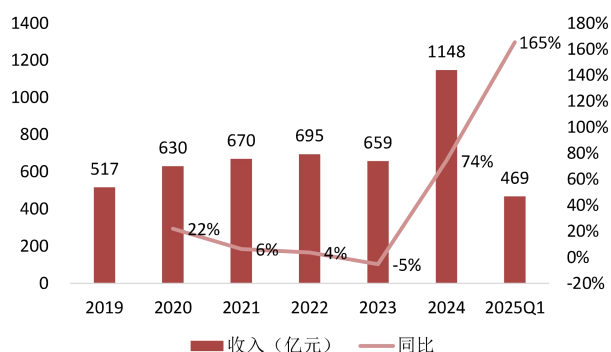


资料来源：Wind，海光信息 2025 年一季报，山西证券研究所

浪潮信息：公司是国内 AI 服务器行业龙头，有望充分受益于 AI 算力需求的持续高景气。根据 IDC 数据，2023 年公司在国内 AI 服务器市场占据 36% 的份额，大幅领先于其他厂商。目前公司与产业链上下游厂商拥有稳固的合作关系，包括上游的国内外芯片厂以及下游的互联网、运营商等行业客户。同时，公司与互联网客户联合推出了 JDM（联合设计制造）模式，具备较高的产品研发、生产和交付效率，能够快速满足下游客户对 AI 服务器的定制化需求。

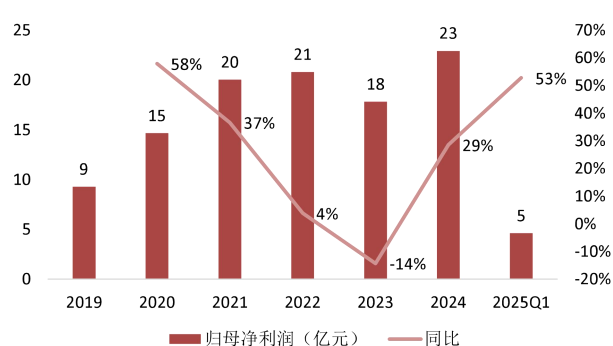
风险提示：技术研发风险；宏观经济持续下行风险；市场竞争加剧风险；汇率波动的风险。

图 29：2019-2025Q1 浪潮信息收入及增速



资料来源：Wind，浪潮信息 2025 年一季报，山西
证券研究所

图 30：2019-2025Q1 浪潮信息归母净利润及增速



资料来源：Wind，浪潮信息 2025 年一季报，山西
证券研究所

5. 风险提示

互联网对国产 AI 芯片的采购不及预期。若未来百度、阿里、腾讯、字节等头部互联网厂商减少对国产 AI 芯片的采购占比，可能会对国产 AI 芯片厂商的业绩产生负面影响。

智算中心项目 AI 算力国产化率提升放缓：2024 年以来，国内智算中心项目的 AI 算力国产化率持续提升，且中国联通已建成并投运全国产化智算中心项目，若未来智算中心的 AI 算力国产化率提升速度放缓，可能会对国产 AI 芯片及对应服务器厂商的经营发展产生不利影响。

国产 AI 芯片技术研发不及预期：目前国产 AI 芯片的性能快速提升，正加速追赶英伟达，若未来下一代芯片性能提升幅度较小，或研发及测试进展不及预期，可能会使 AI 芯片的国产化替代进程放缓，进而对国产芯片厂商的发展产生负面影响。

行业竞争格局加剧风险：目前国产 AI 算力需求保持高景气，若 AI 芯片和服务器市场出现大量新进入者，现有企业的利润率可能会受到不利影响。

分析师承诺：

本人已在中国证券业协会登记为证券分析师，本人承诺，以勤勉的职业态度，独立、客观地出具本报告。本人对证券研究报告的内容和观点负责，保证信息来源合法合规，研究方法专业审慎，分析结论具有合理依据。本报告清晰准确地反映本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点直接或间接接受到任何形式的补偿。本人承诺不利用自己的身份、地位或执业过程中所掌握的信息为自己或他人谋取私利。

投资评级的说明：

以报告发布日后的 6--12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。

无评级：因无法获取必要的资料，或者公司面临无法预见的结果的重大不确定事件，或者其他原因，致使无法给出明确的投资评级。（新股覆盖、新三板覆盖报告及转债报告默认无评级）

评级体系：**——公司评级**

- 买入： 预计涨幅领先相对基准指数 15%以上；
- 增持： 预计涨幅领先相对基准指数介于 5%-15%之间；
- 中性： 预计涨幅领先相对基准指数介于-5%-5%之间；
- 减持： 预计涨幅落后相对基准指数介于-5%- -15%之间；
- 卖出： 预计涨幅落后相对基准指数-15%以上。

——行业评级

- 领先大市： 预计涨幅超越相对基准指数 10%以上；
- 同步大市： 预计涨幅相对基准指数介于-10%-10%之间；
- 落后大市： 预计涨幅落后相对基准指数-10%以上。

——风险评级

- A： 预计波动率小于等于相对基准指数；
- B： 预计波动率大于相对基准指数。

免责声明：

山西证券股份有限公司(以下简称“公司”)具备证券投资咨询业务资格。本报告是基于公司认为可靠的已公开信息，但公司不保证该等信息的准确性和完整性。入市有风险，投资需谨慎。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，公司不对任何人因使用本报告中的任何内容引致的损失负任何责任。本报告所载的资料、意见及推测仅反映发布当日的判断。在不同时期，公司可发出与本报告所载资料、意见及推测不一致的报告。公司或其关联机构在法律许可的情况下可能持有或交易本报告中提到的上市公司发行的证券或投资标的，还可能为或争取为这些公司提供投资银行或财务顾问服务。客户应当考虑到公司可能存在可能影响本报告客观性的利益冲突。公司在知晓范围内履行披露义务。本报告版权归公司所有。公司对本报告保留一切权利。未经公司事先书面授权，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，或以任何侵犯公司版权的其他方式使用。否则，公司将保留随时追究其法律责任的权利。

依据《发布证券研究报告执业规范》规定特此声明，禁止公司员工将公司证券研究报告私自提供给未经公司授权的任何媒体或机构；禁止任何媒体或机构未经授权私自刊载或转发公司证券研究报告。刊载或转发公司证券研究报告的授权必须通过签署协议约定，且明确由被授权机构承担相关刊载或者转发责任。

依据《发布证券研究报告执业规范》规定特此提示公司证券研究业务客户不得将公司证券研究报告转发给他人，提示公司证券研究业务客户及公众投资者慎重使用公众媒体刊载的证券研究报告。

依据《证券期货经营机构及其工作人员廉洁从业规定》和《证券经营机构及其工作人员廉洁从业实施细则》规定特此告知公司证券研究业务客户遵守廉洁从业规定。

山西证券研究所：**上海**

上海市浦东新区滨江大道 5159 号陆家嘴滨江中心 N5 座 3 楼

太原

太原市府西街 69 号国贸中心 A 座 28 层
电话：0351-8686981
<http://www.i618.com.cn>

深圳

广东省深圳市福田区金田路 3086 号大百汇广场 43 层

北京

北京市丰台区金泽西路 2 号院 1 号楼丽泽平安金融中心 A 座 25 层

