公司研究

2025年06月29日

公司深度

电子

买入（首次覆盖）

寒武纪（688256）:云边端共铸国产算力脊梁，软硬件同迎寒武破晓时代

——公司深度报告

证券分析师

方霁 S0630523060001

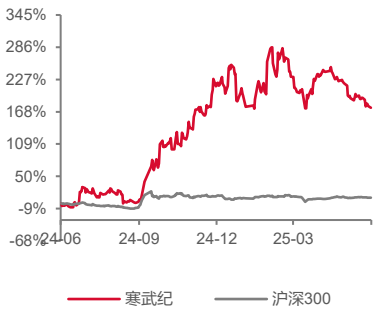
fangji@longone.com.cn

联系人

董经纬

djwei@longone.com.cn

数据日期	2025/06/27
收盘价	585.50
总股本(万股)	41,835
流通A股/B股(万股)	41,746/0
资产负债率(%)	15.97%
市净率(倍)	41.97
净资产收益率(加权)	6.32
12个月内最高/最低价	818.87/187.50



相关研究

1. AI大模型竞赛方兴未艾，OpenAI与DeepSeek引领行业生态重构——半导体行业深度报告（十二）
2. AI大模型风起云涌，半导体与光模块长期受益——半导体行业深度报告（十）

投资要点：

- **寒武纪是国内稀缺的云端AI芯片厂商，提供云边端一体、软硬件协同、兼顾训练与推理的系列化智能AI芯片产品和平台化基础系统软件。**公司业务主要分为云端产品线、边缘产品线、IP授权及软件三块，产品面向互联网、金融、交通、能源、电力和制造等领域的复杂AI应用场景提供算力，赋能产业升级。云端产品线主要提供云端AI芯片、加速卡、训练整机等，涵盖模型训练与推理，目前已迭代至思元590系列；边缘产品线以思元220为主，主要服务于智能制造、智能家居等边缘计算场景；IP授权及软件主要包括以寒武纪1M为主的终端智能处理器IP以及基础软件开发平台Cambricon Neuware，软硬件协同打造优质生态护城河。2024Q4起，公司营收同环比大增且归母净利润扭亏为盈，2025Q1公司营收同比增长4230.22%，归母净利润同比增长256.82%，主要系在当前政策背景下AI芯片国产替代需求强劲，以及公司云端产品国内性能领先、技术壁垒较高，在客户处大幅放量，此外目前公司存货充裕且预付账款大幅增加，为未来业绩兑现储备了强劲动能。

- **推理与训练算力需求爆发拉动AI芯片市场规模扩张，海外芯片龙头仍占据国内市场主导地位，但随着厂商加大研发及行业政策刺激，以寒武纪为代表的本土AI芯片品牌的国产替代正全面提速，2025年份额有望升至40%。**（1）需求面看，当前AI大模型训练和推理所需算力都在持续上升，海内外云厂商也在不断扩大AI基建资本开支，自2024年起未来5年全球算力规模增速将超过50%。中国算力规模全球份额位居第二，2023年占比31%。算力需求的暴增促进了AI芯片市场规模的扩张，2025年全球AI芯片规模有望达920亿美元，同比增长29.58%，中国AI芯片市场规模2024年有望达到1412亿元，占比全球约28%，目前国内GPU仍为主流，但ASIC和FPGA份额正加速增长，互联网厂商为主要采购商。（2）供给面看，受性能领先其他厂商1-2年、市场先发优势、软硬件护城河等因素影响，英伟达仍在全球GPU市场占有绝对领先地位，但以寒武纪、华为等厂商为代表的国产AI厂商正在加速布局，2024年国内本土AI芯片品牌的出货份额已达30%，预计2025年将升至40%。此外，在美国AI芯片管制政策趋严的背景下，国产AI芯片自主可控进程也有望全面提速。

- **硬件层面，云端AI芯片是支撑公司营收的主要力量，MLU370-X8单卡性能与主流350W RTX GPU相当，思元590已进入国产供应链；软件层面，统一的平台级基础系统软件已成功支持大多头部AI模型，自研架构与指令集保持了核心技术的自主可控。**公司硬件产品涵盖云、边、端全场景，云端AI芯片2024年营收占比高达99.30%，公司已迭代发布了四代产品，思元370最大算力达256TOPS(INT8)，MLU370-X8单卡性能与主流350W RTX GPU相当，覆盖训练与推理需求，思元590目前有望成为营收的新支撑。边缘端产品围绕思元220推出了相应加速卡及智能模组。终端智能处理器1A、1H、1M系列芯片覆盖0.5TOPS-8TOPS内不同档位的AI算力需求，可集成于手机或IoT类SoC芯片中。软件层面，公司为硬件产品提供统一的平台级基础系统软件Cambricon Neuware，训练与推理软件平台均成功支持并优化了如DeepSeek、Llama、Qwen等主流AI大模型。此外，公司软硬件均基于自研处理器架构，且构建于自研MLU指令集基础之上，有助于保持核心技术的自主可控。

- **投资建议：首次覆盖，给予“买入”评级。**作为国内稀缺的云端AI芯片标的公司，公司云端产品线正大幅放量中，思元系列芯片产品受益于国内各产业算力需求的提升以及国产替代的趋势，有望带动公司整体营收高增以及归母净利润的持续盈利。我们预计公司2025-2027年营业收入分别为84.43、161.71和251.05亿元，同比增速分别为618.91%、91.52%和55.25%；归母净利润分别为15.95、38.60和69.13亿元，同比增速分别为452.69%、

141.96%和79.09%。对应2025-2027年的PE分别为153、63、35倍，PS分别为29、15、10倍。

- **风险提示**：产品研发及验证进度不及预期风险；地缘政治风险；宏观经济下行风险。

盈利预测与估值简表

	2022A	2023A	2024A	2025E	2026E	2027E
主营收入(百万元)	729.03	709.39	1,174.46	8,443.34	16,170.54	25,104.50
同比增速(%)	1.11%	-2.70%	65.56%	618.91%	91.52%	55.25%
归母净利润(百万元)	-1,256.35	-848.44	-452.34	1,595.33	3,860.00	6,912.75
同比增速(%)	-52.3%	32.5%	46.7%	452.69%	141.96%	79.09%
毛利率(%)	65.76%	69.16%	56.71%	60.57%	62.96%	63.01%
每股盈利(元)	-3.01	-2.03	-1.08	3.82	9.25	16.56
ROE(%)	-25.9%	-15.0%	-8.3%	22.7%	35.5%	38.9%
PE(倍)	-	-	-	153.21	63.32	35.36
PS(倍)	335.27	344.55	208.11	28.95	15.12	9.74

资料来源：携宁，东海证券研究所（截至2025年6月27日）

正文目录

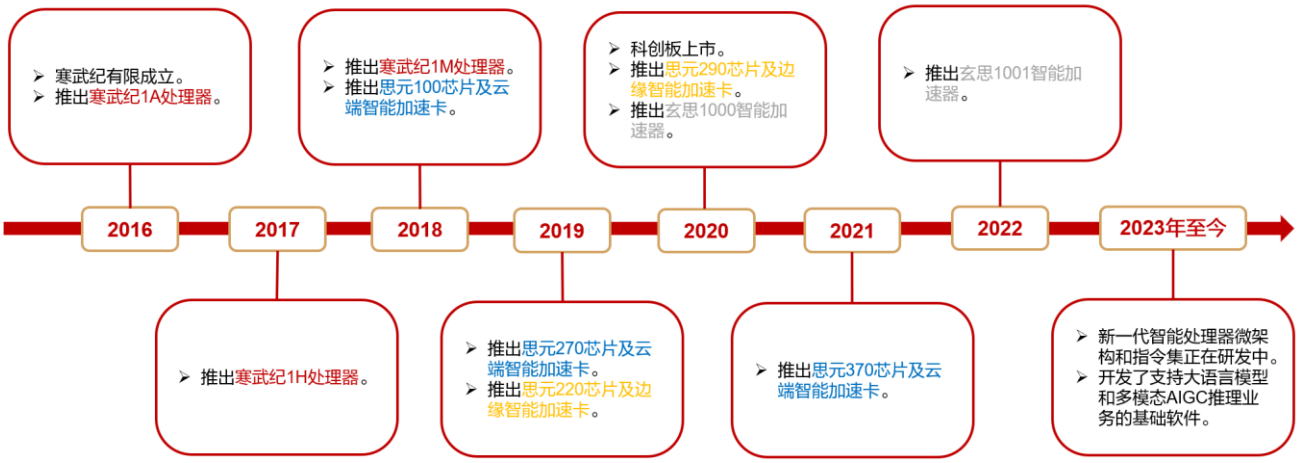
图表目录

1.国产耀眼“芯”光，业绩腾飞在即

1.1.云端 AI 芯片国产化稀缺标的，自研硬件与平台软件双轮驱动

（1）寒武纪是国内稀缺的 AI 芯片设计公司，提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。寒武纪成立于 2016 年，并于 2020 年上市，成立以来专注于人工智能芯片产品的研发与技术创新，致力于打造人工智能领域的核心处理器芯片，推出的产品包括但不限于思元 100、220、270、290 和 370 芯片等，涵盖云端、边缘端和终端产品，产品广泛应用于服务器厂商和产业公司，面向互联网、金融、交通、能源、电力和制造等领域的复杂 AI 应用场景提供充裕算力，推动人工智能赋能产业升级。

图1 寒武纪发展历程



资料来源：公司公告，东海证券研究所

（2）按产品线划分，寒武纪业务主要涵盖云端、边缘端产品线及 IP 授权及软件三大块。

1）云端产品线方面，主要为云端智能芯片及加速卡和训练整机。云端智能芯片及加速卡是云服务器、数据中心等进行人工智能处理的核心器件，其主要作用是为云计算和数据中心场景下的 AI 应用程序提供高性能、高计算密度、高能效的硬件计算资源，支撑该类场景下复杂度和数据吞吐量高速增长的 AI 处理任务；训练整机通常集成多颗 AI 芯片，以玄思 1000 智能加速器为例，玄思 1000 智能加速器整机在 2U 机箱内集成了 4 颗思元 290 智能芯片，2 台玄思 1000 加速器与 CPU 服务器可组成一套包括 8 张加速卡的整机系统，可实现 AI 算力多向扩展，满足性能、扩展性、灵活性、鲁棒性（Robustness，反映一个系统在面临着内部结构或外部环境的改变时也能够维持其功能稳定运行的能力）的要求。

2）边缘产品线指边缘智能芯片及加速卡，可支持边缘计算场景下的智能数据分析与建模、视觉、语音、自然语言处理等多样化的 AI 应用。边缘计算一方面可有效弥补终端设备计算能力不足的劣势，另一方面可缓解云计算场景下数据安全、隐私保护、带宽与延时等潜在问题，可应用于智能制造、智能零售、智能教育、智能家居、智能电网、智能交通等众多领域。

3）IP 授权及软件主要包含终端智能处理器 IP 和基础系统软件平台。终端智能处理器是终端设备中支撑 AI 处理运算的核心器件，例如近年来各品牌旗舰级手机上与图像视频、

语音、自然语言相关的智能应用均依靠终端智能处理器提供计算能力支撑。为了提升性能降低功耗，同时节省成本，终端智能处理器通常不是以独立芯片的形式存在，而是作为一个模块集成于终端设备的 SoC 芯片当中。公司的终端智能处理器 IP 产品主要有 1A、1H 和 1M 系列；基础系统软件平台方面公司主要提供统一的平台级基础系统软件 Cambricon Neuware（包含软件开发工具链等），打破不同场景之间的软件开发壁垒，无须繁琐的移植即可让同一人工智能应用程序便捷高效地运行在公司云边端系列化芯片与处理器产品之上。

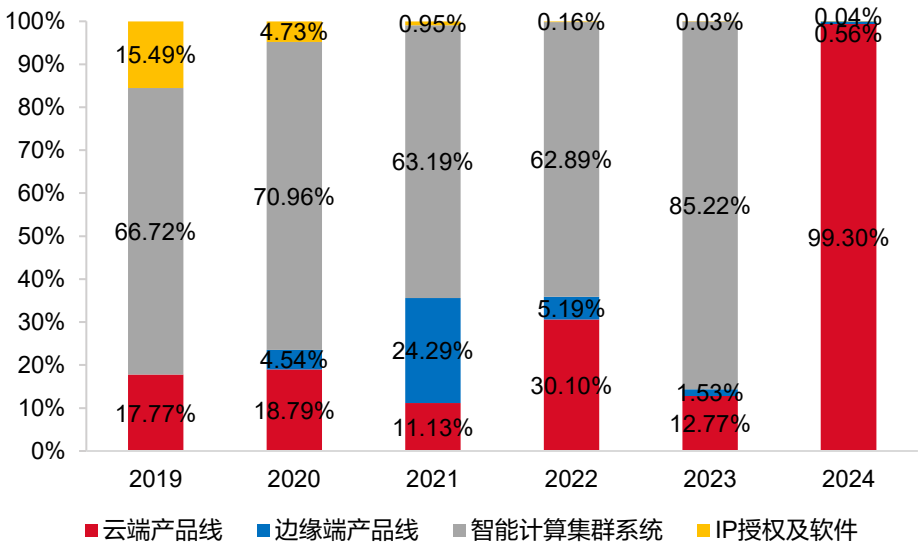
图2 寒武纪主要产品线及产品性能

产品线	产品类型	主要产品	相关性能
云端产品线	云端智能芯片及加速卡	思元100（MLU100）芯片及云端智能加速卡	-
		思元270（MLU270）芯片及云端智能加速卡	计算精度支持：INT16,INT8,INT4,FP32,FP16 峰值算力：128TOPS(INT8)；256TOPS(INT4)；64TOPS(INT16) 内存容量：16GB DDR4, ECC
		思元290（MLU290）芯片及云端智能加速卡	制程：7nm 峰值算力：自适应精度训练算力512TOPS(INT8);256TOPS(INT16);64TOPS(CINT32) 内存容量：32GB HBM2高带宽内存
		思元370（MLU370）芯片及云端智能加速卡（以MLU370-X8为例）	制程：7nm 计算精度支持：FP32,FP16,BF16,INT16,INT8,INT4 峰值算力： 256TOPS(INT8);128TOPS(INT16);96TFLOPS(FP16);96TFLOPS(BP16);24TFLOPS(FP32) 内存容量：48GB LPDDR5
	训练整机	玄思1000智能加速器	峰值算力：自适应精度算力 2.05 PetaOPS (INT8);1 PetaOPS (INT16);256 TOPS (CINT32) 内存容量：128GB
		玄思1001智能加速器	-
边缘产品线	边缘智能芯片及加速卡	思元220（MLU220）芯片及边缘智能加速卡	-
IP授权及软件	终端智能处理器IP	寒武纪1A处理器	-
		寒武纪1H处理器	较初代产品其能效比有着数倍提升，广泛应用于计算机视觉、语音识别、自然语言处理等人工智能处理关键领域
		寒武纪1M处理器	具备了更优性能、更低功耗和更强的完备性，混合支持fp32/fp16/int32/int16/int8/int4位宽，增加了压缩解压缩模块。在上代基础上，可支持个性化人工智能应用，也可使用于多路视频实时处理和自动驾驶等领域。
	基础系统软件平台	寒武纪基础软件开发平台（适用于公司所有芯片与处理器产品）	-

资料来源：公司公告，东海证券研究所

（3）以思元系列芯片为代表的公司云端产品线迅速起量，2024 年已开始显著贡献营收。2023 年及以前，公司绝大部分营收主要由智能计算集群系统提供，即公司以自有的思元系列芯片加速卡产品为核心，基于 Cambricon Neuware 基础系统软件平台，为客户提供定制化的软硬件整体解决方案，以科学地配置和管理集群的软硬件、提升运行效率，公司积极参与并中标国内各类智算中心项目，该部分营收占据整体营收 60%以上。同时，公司积极推进云端思元系列芯片研发，并持续拓展客户市场，助力 AI 应用落地，2024 年云端产品线收入大幅增长，达 11.66 亿元，同比大增 1187.78%，占总营收的份额也从 2023 年的 12.77%上升至 99.30%，体现出公司 AI 芯片市场竞争力不断上升，规模效应有望逐步显现。

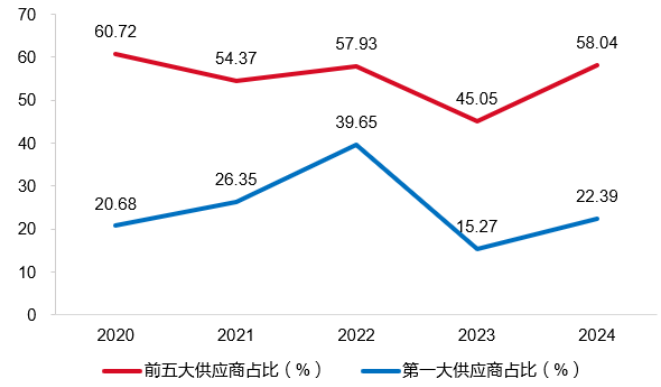
图3 2019-2024 年公司分业务营收占比



资料来源：公司公告，东海证券研究所

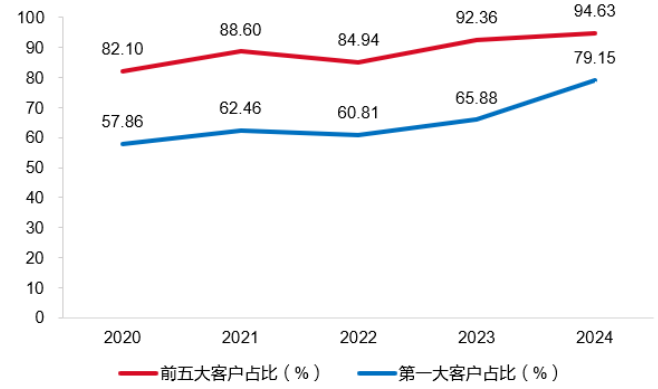
（4）公司主要采用 Fabless 业务模式，上游供应商包含晶圆代工及封测厂、芯片 IP 授权及 EDA 工具厂商、服务器厂商等，下游客户涵盖各行业，但集中程度较高。公司在完成芯片设计后，将最终的芯片版图交付给代工厂进行晶圆代工，然后委托日月光或 Amkor 等封测厂商完成芯片的封装测试，再由电路板厂商使用芯片生产出加速卡（即包含智能芯片的电路板），最后将加速卡销售给客户。公司自上市以来，前五大客户销售占比一直维持 80%以上高位，2024 年占比高达 94.63%，其中第一大客户占比均在 50%以上，2024 年第一大客户占比 79.15%，体现客户结构的高度集中性，若公司主要客户经营发生变动或者需求放缓，或新客户拓展情况不及预期，可能影响公司业绩。

图4 2020-2024 年公司前五大&第一大供应商占比



资料来源：公司公告，东海证券研究所

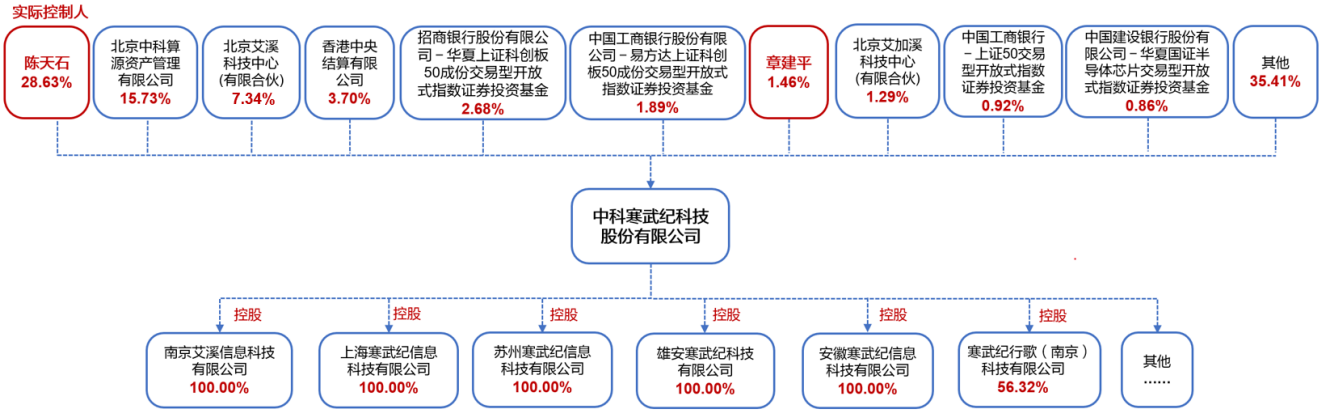
图5 2020-2024 年公司前五大&第一大客户占比



资料来源：公司公告，东海证券研究所

（5）公司控股股东、实际控制人为陈天石博士，股权结构较为稳定。自寒武纪上市以来，公司创始人、董事长、总经理陈天石博士一直为公司控股股东与实控人，截至 2025 年 3 月 31 日，陈天石博士直接持有公司 28.63%股份，同时，陈天石博士是北京艾溪科技中心（有限合伙）的执行事务合伙人及实控人，艾溪合伙持有公司 7.34%股份，因此陈天石博士合计拥有公司 35.97%的表决权，为公司的控股股东以及实际控制人。北京中科算源资产管理有限公司（由中国科学院计算技术研究所全资设立）为公司第二大股东，持有 15.73%股份。

图6 公司股权结构图（截至 2025.3.31）



资料来源：iFind，东海证券研究所

（6）公司高管多数拥有丰富的学历与技术背景及相关从业经验，技术团队成熟。公司拥有成熟的管理和研究团队，董事长陈天石博士在创立公司前曾任中科院计算技术研究所博士生导师，副总经理刘少礼先生、王在先生、陈帅先生也均拥有计算机相关博士学位并曾在中科院计算所工作，副总经理刘毅先生与张尧先生均拥有相关硕士学位和丰厚的产业从业经验，高管层整体年龄在 40 岁左右，较为年轻。

表1 公司高管信息

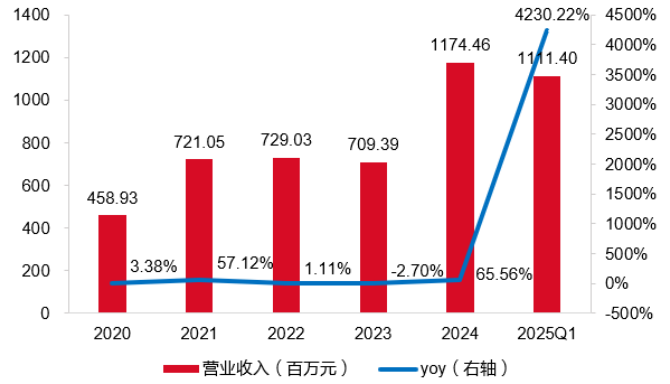
姓名	职位	年龄	履历
陈天石	董事长、总经理、核心技术人员	40	中国科学技术大学计算机软件与理论专业博士学历。2010 年 7 月至 2019 年 9 月就职于中国科学院计算技术研究所，历任助理研究员、副研究员及硕士生导师、研究员及博士生导师。2016 年 3 月创立公司。
刘少礼	董事、副总经理、核心技术人员	38	中科院计算所计算机系统结构博士学历。2014 至 2019 年任中科院计算所副研究员。2016 年作为公司创始团队成员加入公司。
王在	董事、副总经理	41	中国科学技术大学计算机应用技术博士学历。2011 至 2015 年任郑州商品交易所核心交易系统工程师，2015 至 2016 年任中原银行信息科技部电子银行系统主管，2016 至 2018 年就职于中科院计算所从事科研工作。2016 年作为公司创始团队成员加入公司。
叶澜尹	董事、副总经理、财务负责人、董事会秘书	37	北京大学西方经济学硕士学历。2012 至 2016 年就职于中国高新投资集团公司并任投资经理、高级投资经理，2016 至 2019 年就职于国投创业投资管理有限公司并任投资副总裁。2019 年加入公司。
陈帅	副总经理、核心技术人员	39	国科学院计算技术研究所计算机系统结构博士学历。2014 至 2015 年，任中国科学院计算技术研究所工程师。2015 至 2016 年，任多伦多大学电子和计算机工程系博士后。2016 年至今就职于中科寒武纪科技股份有限公司。
刘毅	副总经理、核心技术人员	40	北京大学微电子与固体电子学硕士学历。2010 至 2012 年，任龙芯中科技术股份有限公司工程师。2012 至 2016 年，任上海英伟达半导体（科技）有限公司高级工程师。2016 年至今就职于中科寒武纪科技股份有限公司。
张尧	副总经理、核心技术人员	39	中国科学院计算技术研究所计算机系统结构硕士研究生学历。2012 年至 2014 年任中国科学院计算技术研究所微处理器中心助理工程师。2014 年至 2015 年任龙芯中科技术股份有限公司高级工程师。2015 年至 2016 年任北京小米松果电子有限公司高级工程师。2016 年至今，就职于中科寒武纪科技股份有限公司。

资料来源：公司公告，东海证券研究所

1.2.云端产品线助力营收大幅增长，未来业绩兑现动能强劲

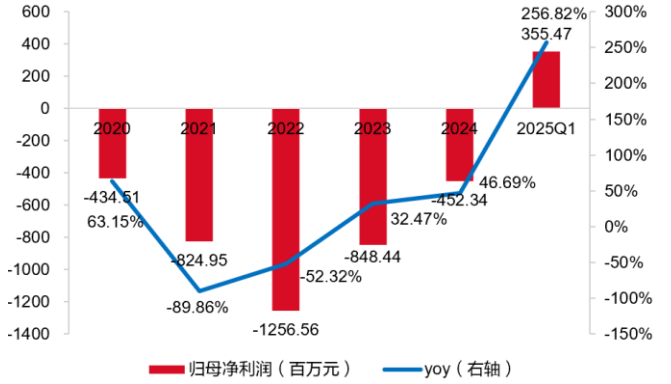
（1）2024 年与 2025 年一季度营收同比高增，归母净利润已扭亏为盈。2024 年以前，公司营收较为依赖智能计算集群系统，2021-2023 年营收均在 7 亿元左右，同时由于研发支出较高等原因，2024 年及以前公司归母净利润均处于亏损状态，但自 2023 年起，亏损逐年收窄，2024 年第四季度公司归母净利润首次扭亏为盈，为 2.72 亿元，2025 年第一季度延续盈利，归母净利润为 3.55 亿元，同比增长 256.82%。同时 2025 年一季度营收达历史新高，为 11.11 亿元，同比大增 4230.22%，主要系在当前政策背景下 AI 芯片国产替代需求强劲，以及公司云端产品国内性能领先、技术壁垒较高的前提下，公司云端产品线在客户处大幅放量。

图7 2020-2025 年第一季度公司营业收入与同比增速



资料来源：公司公告，东海证券研究所

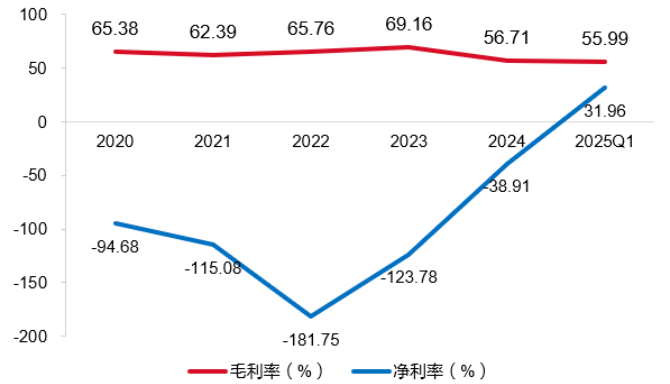
图8 2020-2025 年第一季度公司归母净利润与同比增速



资料来源：公司公告，东海证券研究所

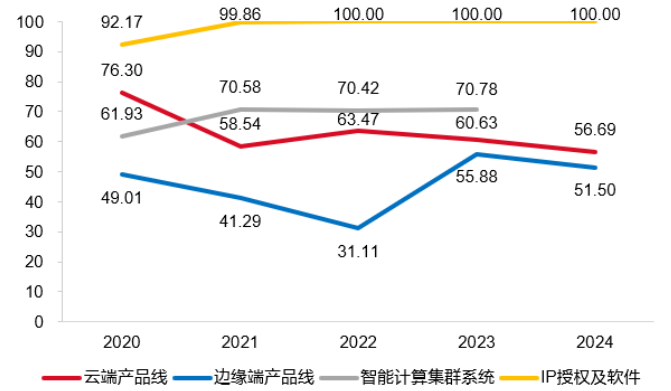
（2）公司主营业务毛利率均常年维持 50%以上，净利率大幅回升。2020 年至 2025 年一季度，公司毛利率均在 60%上下波动，净利率在 2025 年一季度回升至 31.96%，体现了经营情况的整体好转。分业务看，IP 授权及软件毛利率基本为 100%，智能计算集群系统毛利率在 70%左右水平，自 2023 年起，云端、边缘端产品线毛利率均在 50%以上。

图9 2020-2025 年第一季度公司毛利率与净利率



资料来源：公司公告，东海证券研究所

图10 2020-2024 年公司分业务毛利率（%）

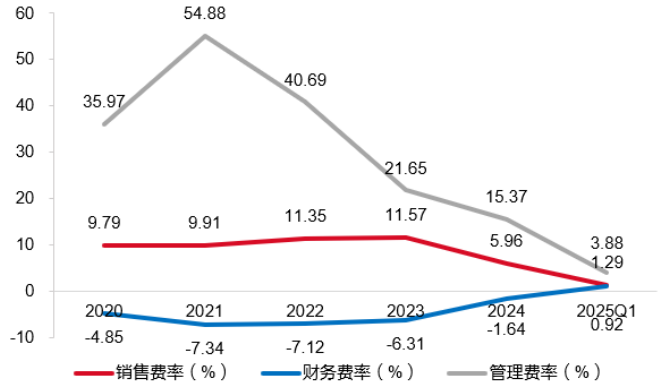


资料来源：公司公告，东海证券研究所

（3）期间费用率显著降低，高研发投入为业绩增长储备动能。公司管理费率自 2021 年的 54.88%起逐年降低至 2025 年一季度的 3.88%，销售费率也从 2023 年的 11.57%收窄至 2025 年一季度的 1.29%，体现出公司经营效率的不断提高。同时，从研发层面看，公司研发人员数量常年占据总员工数量的 75%以上，2024 年研发人员中 78.95%拥有硕士及以上学位，研发费率在 2022 年曾达到 208.92%的高位，主要来自于新产品流片费用增加，另外

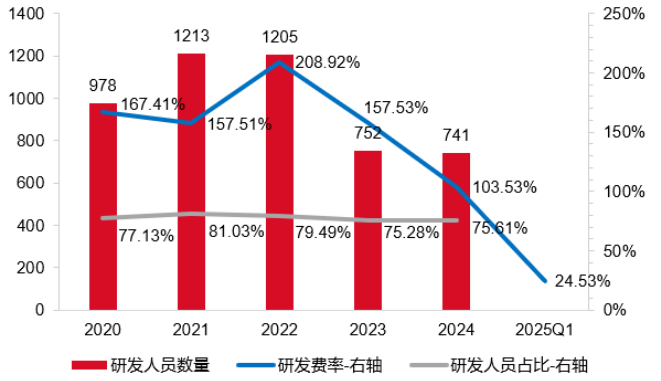
来自于公司提高研发人才薪酬水平和购置 IP、EDA 及相关研发设备等支出增多。2025 年一季度，公司研发费率降至 24.53%，主要系营收水平提升所致。

图11 2020-2025 年第一季度公司期间费用率情况



资料来源：公司公告，东海证券研究所

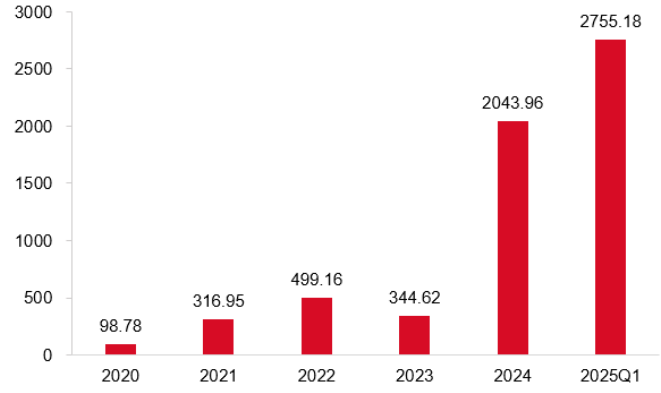
图12 2020-2025 年第一季度公司研发投入情况



资料来源：公司公告，东海证券研究所

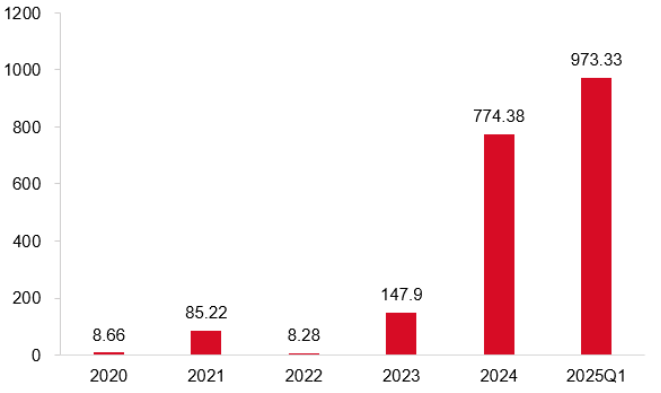
（4）公司存货与预付账款大幅增加，为未来业绩兑现储备动能。公司存货从 2023 年末的 3.45 亿元暴涨至 2024 年末的 20.44 亿元，2025 年一季度继续增至 27.55 亿元，主要系委托加工物资增加所致。同时，公司预付账款也实现了大幅增长，由 2023 年末的 1.48 亿元增至 2024 年末的 7.74 亿元，2025 年一季度上升至 9.73 亿元，主要系公司对供应商的预付款项增加所致。存货与预付账款大幅上升，反映了公司业务扩张需求以及对未来业绩的预期和信心，展望未来业绩有望进一步释放。

图13 2020-2025 年第一季度公司存货（百万元）



资料来源：公司公告，东海证券研究所

图14 2020-2025 年第一季度公司预付账款（百万元）



资料来源：公司公告，东海证券研究所

2.AI 寒武纪元将临，国产算力渗透加速

2.1.AI 芯片是人工智能产业链的核心器件

（1）AI 芯片也被称为 AI 加速器或计算卡，是专门用于处理人工智能应用中的大量计算任务的模块。目前，AI 芯片主要包括 GPU、ASIC、FPGA 等类型，其中 ASIC 芯片又可衍生出 TPU、NPU 等种类。GPU 能够进行大量并行数据处理和运算，通用性较强，擅长数学运算、图形渲染等任务，代表厂商为英伟达、AMD 等；ASIC 芯片为专门针对某一领域设计的芯片，因此专用性更强，但同时开发成本高、开发周期长，代表厂商如谷歌的 TPU，寒武纪、华为昇腾的 NPU 等；FPGA 为可现场多次编程的门电路阵列的硬件，其硬件可编程性使得其灵活性高，但同时设计难度和复杂性也较高，代表厂商如 Xlinx 等。AI 芯片可广泛应用于云计算、数据中心、智能驾驶、智慧家电等领域。

表2 各类 AI 芯片的原理、优劣势、代表厂商与产品和下游应用介绍

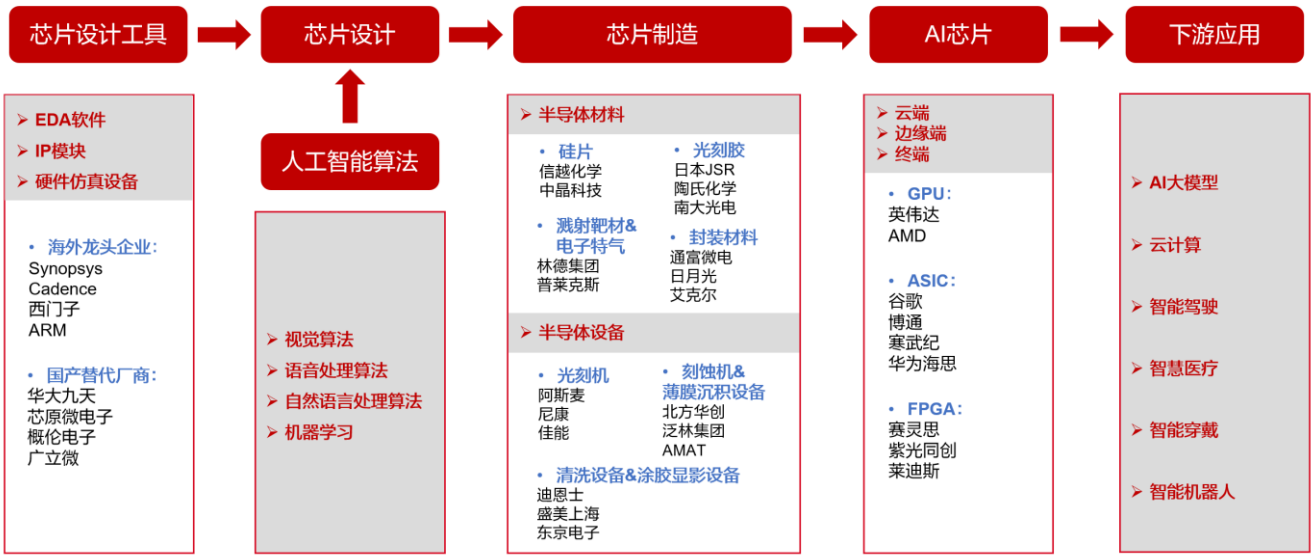
AI 芯片类型	技术原理与特点	优劣势	代表厂商与产品	下游应用
GPU	将极为繁重的数学进行任务拆解，以英伟达 GPU 为例，利用流式多处理器（SM）的机制，将大量的运算拆解为一个个简单的运算并行处理	可以进行大量并行数据处理和运算，强项是做数学运算和图形渲染，通用性极强	英伟达、AMD 等，产品如 B100、B200、AMD instinct MI325X 等	云计算，深度学习训练和数据中心
ASIC	专用集成电路，专门针对某一领域设计的芯片，所有接口模块都连接到一个矩阵式背板上，通过 ASIC 芯片到 ASIC 芯片的直接转发，可同时进行多个模块之间的通信，每个模块的缓存只处理本模块上的输入输出队列，因此对内存芯片性能的要求大大低于共享内存方式，访问效率高，适合同时进行多点访问，容易提供非常高的带宽，并且性能扩展方便	专用性强，性能更好；但研发成本高、周期长，不利于灵活多变的任务	谷歌、Marvell、寒武纪、华为昇腾等	智慧家电、智能驾驶、智慧工厂、国防军事等专用领域
TPU	张量处理器，ASIC 的一种，用以运行构建 AI 模型所需的独特矩阵和基于矢量的数学运算。TPU 的核心是 MXU（矩阵乘法单元），MXU 以脉动阵列为架构，使 TPU 能够以很高的吞吐量执行矩阵乘法和累加	在性能功耗比上优势明显，算力利用率高；但是算力较 GPU 落后一代；脉动式计算模式相对固定，不适合有大量控制流的计算	谷歌最早开发，2024 年发布第六代 TPU Trillium	智能驾驶域控制器芯片，工业机器人专用 TPU 芯片
NPU	神经网络处理器，专为深度学习与神经网络计算优化的处理器，ASIC 的一种，NPU 的计算模型基于数据流的并行执行和异步处理能力，允许大量神经网络操作同时进行	可同时在硬件与软件上实现高效的并行性，从而显著提高处理速度和吞吐量，性能高，但是软件生态目前不够成熟，兼容性较差	寒武纪 Cambricon-1A/1H/1M，苹果 M3 芯片，华为昇腾 910B 等	主要用于 AI 推理
FPGA	可现场多次编程的门电路阵列的硬件，通过利用 LUT（查找表）来实现灵活定义期望行为的方式来反复编程，从而支持不同的 AI 数据模型，更适合做需要低延迟的流式处理；所有模块都可以定制开发，计算逻辑灵活，保持数据的同构性	无指令，无需共享内存，延迟低；设计难度和复杂度高；处理重复度不高的任务时不如 GPU	Xlinx 的 Spartan，Artix 系列；深圳紫光同创的 Titan 和 logos-2 系列	如 LED 显示屏控制卡，覆盖航天航空、通信网络、信息安全、数据中心、工业物联网等多个行业

资料来源：CSDN，东海证券研究所

（2）AI 芯片产业链上游主要包括 AI 算法、设计工具以及制造中代工和封测环节涉及到的半导体材料与设备，AI 芯片可广泛应用于大模型、云计算以及各类终端智能场景。AI 算法主要包括视觉算法、语言处理算法、自然语言处理算法、机器学习等，芯片设计工具主要涵盖 EDA 软件、IP 模块与硬件仿真设备，芯片制造环节涉及的材料与设备类别则更为丰富，

AI 芯片按照应用场景的不同可分为云端、边缘端及终端 AI 芯片，下游应用包括但不限于 AI 大模型、云计算、智能驾驶、智慧医疗、智能穿戴、智能机器人等。

图15 AI 芯片产业链

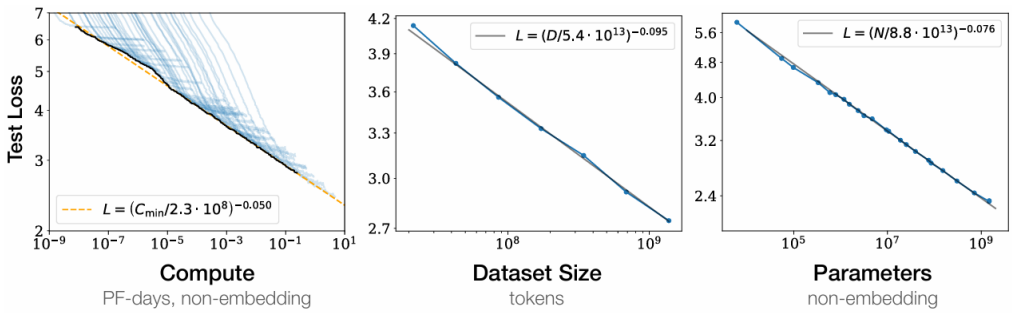


资料来源：公司公告，中商产业研究院，东海证券研究所

2.2.推理与训练需求驱动算力升级，云厂商加大 AI 基建投入

（1）随着当前 AI 大模型的迭代以及性能的提升，其参数量的指数级上升使得大模型训练所需算力同样迅速增长。根据大模型的预训练第一性原理“Scaling Law”，在机器学习领域，特别是对于大语言模型而言，模型性能（L，模型在测试集上的交叉熵损失）与模型的参数量大小（N）、训练模型的数据大小（D）以及训练模型使用的计算量（C）之间存在一种可预测的关系。这种关系通常表现为随着这些因素的增长，模型性能会按照一定的幂律进行改善，说明当在模型训练阶段提高算力投入，模型性能会显著增长。根据《Scaling Laws for Neural Language Models》，对于每个训练 Token、每个模型参数，约需要进行 6 次浮点运算。以 GPT 系列模型为例，GPT-2 参数规模为 15 亿，GPT-3 来到了 1750 亿，GPT-4 更是约为 1.8 万亿，随着模型迭代和性能提高，参数规模以指数级级别增长，以 GPT-3 大模型训练为例，模型参数量为 1750 亿，训练 Token 数量为 3000 亿，其需要的训练总算力为 175B × 300B × 6 = 3.15*10^23 FLOPs。

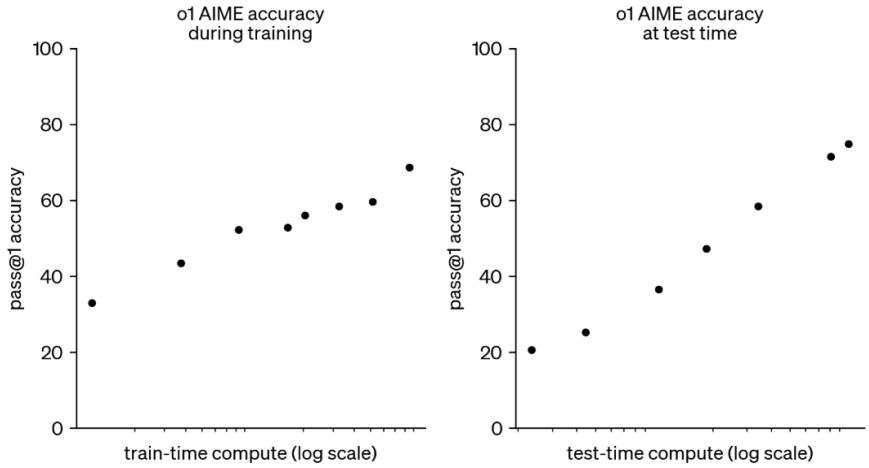
图16 模型性能与计算量、数据大小、参数量的关系



资料来源：《Scaling Laws for Neural Language Models》，Kaplan, McCandlish, Henighan, B. Brown, Chess, Child, &et al.(2020)，东海证券研究所

（2）目前推理侧算力资源地位愈发重要，AI 推理 token 的生成量在过去一年激增了 10 倍。随着 OpenAI o1 系列推理模型的发布，证明了推理侧的算力资源投入同样重要，“Scaling Law”在推理阶段或同样适用。o1 模型引入的思维链类似人类在回答困难问题之前的长时间思考，通过训练时的强化学习，o1 能够锻炼其思维链并改进其使用的策略，它还能够识别并改正错误，将棘手的问题拆分成更简单的步骤，如果目前的方式不奏效，o1 还会尝试不同的解决方式。上述思维链让 o1 的推理能力大幅增强。如下图所示，当推理侧的算力资源增加时，模型处理问题的准确度显著提升。根据英伟达 CEO 在 2026 财年第一财季业绩会时（2025 年 5 月 28 日）的发言，AI 推理 token 的生成量在过去一年激增了 10 倍，而随着 AI agents 成为主流，对 AI 算力的需求也会加速。

图17 OpenAI o1 在训练和推理阶段算力资源的投入与模型性能的关系

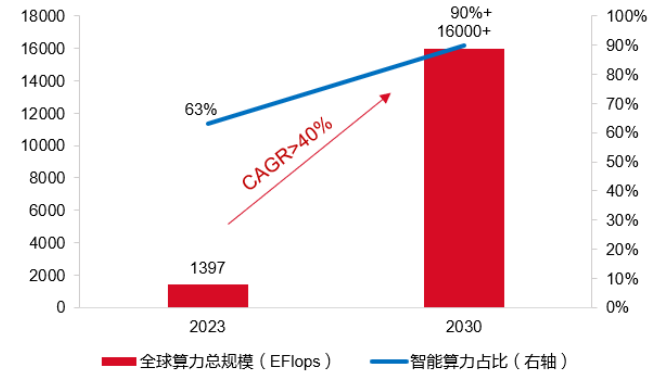


o1 performance smoothly improves with both train-time and test-time compute

资料来源：OpenAI 官网，东海证券研究所

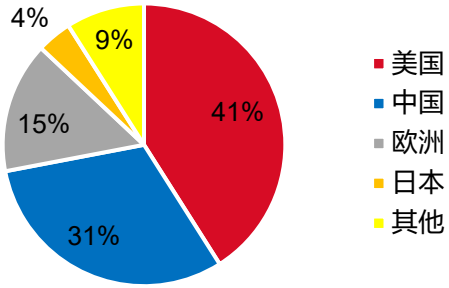
（3）自 2024 年起未来 5 年全球算力规模将以超过 50%的增速增长，2030 年智能算力占比有望达到 90%以上，2023 年中国算力规模全球份额位居第二，占比 31%。根据中国信通院《先进计算暨算力发展指数蓝皮书（2024 年）》，2023 年全球计算设备算力总规模为 1397EFlops，增速达 54%，预计未来 5 年全球算力规模仍将以超过 50%的速度增长，至 2030 年全球算力将超过 16ZFlops（ZFlops 为 EFlops 的一千倍），其中智能算力占比将超过 90%（按 AI 服务器算力总量估算）。随着我国通用数据中心、智能计算中心持续加快部署，2023 年我国基础设施算力规模达到 230EFlops，全国累计建成智算中心达 60 个，近 6 年累计出货超过 114 万台 AI 服务器，算力总规模达到 435EFlops，全球占比 31%，份额仅次于美国，增速达 44%，其中智能算力增速达 62%，占全国总算力比重 2/3。

图18 全球算力总规模及智能算力占比



资料来源：中国信息通信研究院，东海证券研究所

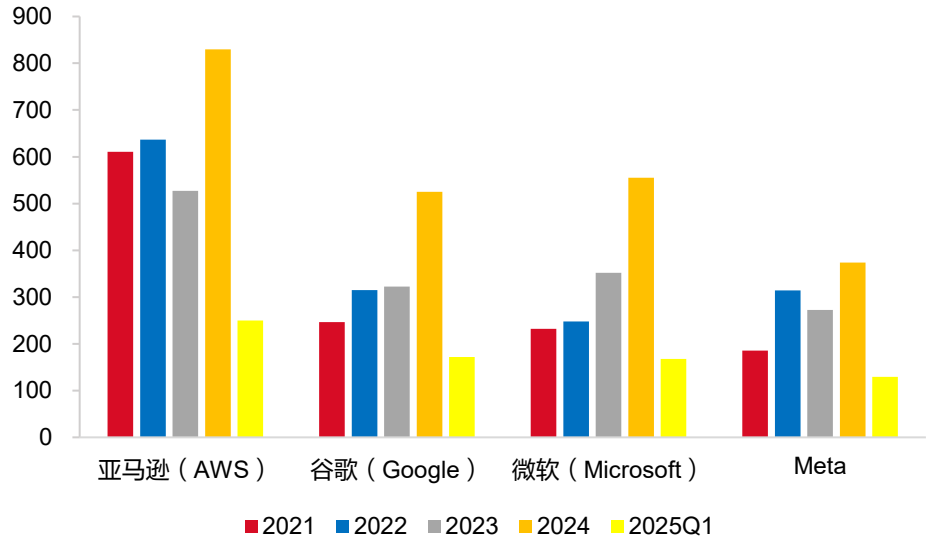
图19 2023 年全球算力规模分布情况



资料来源：中国信息通信研究院，IDC，Gartner，TOP500，东海证券研究所

（4）海内外云厂商加速 AI 基础建设投入，2024 年及 2025Q1 资本开支同比大幅增长。云厂商（云计算服务提供商）是通过互联网提供计算资源、存储、网络、软件等服务的厂商。全球头部云厂商包括亚马逊、谷歌、微软、Meta 等，根据 Canalys，2025 年一季度亚马逊云科技、微软 Azure 和谷歌云相关云支出合计占据全球云支出的 65%。亚马逊、谷歌、微软和 Meta 2024 年资本开支分别同比增长 57.41%、62.89%、57.81%和 37.16%，2025 年一季度资本开支分别同比增长 67.63%、43.17%、52.94%和 102.20%。AI 是云服务增长的核心引擎，随着 AI 从研究阶段迈向部署阶段，企业越来越关注推理阶段的成本效益，与一次性投入资源的模型不同，推理是一项持续的运营成本，对算力资源的需求也是源源不断的，因此云厂商不断加大资本投入建设 AI 基础设施。

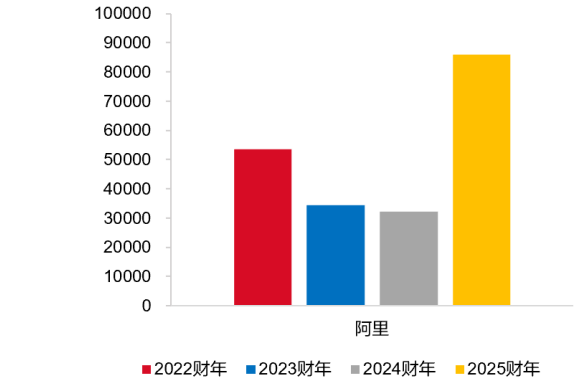
图20 海外头部云厂商 2021-2025 年一季度资本开支（亿美元）



资料来源：iFind，东海证券研究所

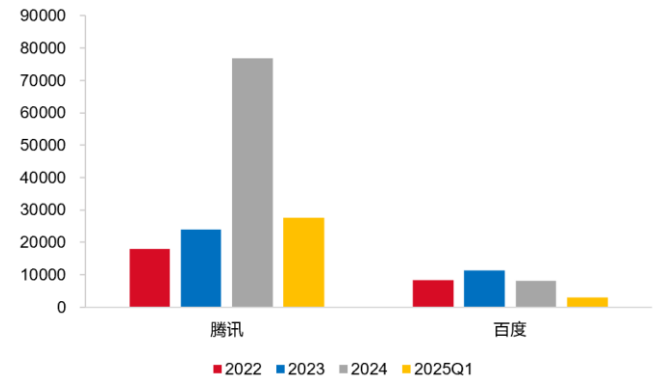
（5）以阿里、腾讯、百度为代表的国内互联网云厂商同样在加大 AI 投入，此外，中国移动、中国电信、中国联通等运营商也在加码算力基础设施投资。腾讯 2024 年资本开支同比增长 221.27%，预计 2025 年将进一步增加资本支出，占收入的低两位数百分比；阿里 2025 财年资本开支同比增长 167.93%，阿里计划未来三年将投入至少 3,800 亿元人民币，用于建设云计算和 AI 的基础设施，这一金额将超过阿里过去十年在云和 AI 基础设施上的投入总和； 2024 年，中国移动智算规模达到 29.2EFLOPS，净增 19.1EFLOPS，呼和浩特、哈尔滨两个万卡级超大规模智算中心上线提供服务。2025 年中国移动资本开支合计约为 1512 亿元，主要用于连接基础设施优化、算力基础设施升级、面向长远的基础设施布局等。其中，在算力领域的投资为 373 亿元，占资本开支的比例提升到 25%，计划智算规模超过 34EFLOPS；中国联通表示，算网数智业务已经成为中国联通第二增长曲线，2024 年中国联通算力投资同比上升 19%，2025 年预计算力投资同比增长 28%；中国电信 2025 算力方面资本开支预计同比增长 22%。

图21 2022-2025 财年阿里资本开支（百万元人民币）



资料来源：公司公告，东海证券研究所（注：阿里巴巴 2025 财年截至 2025 年 3 月 31 日）

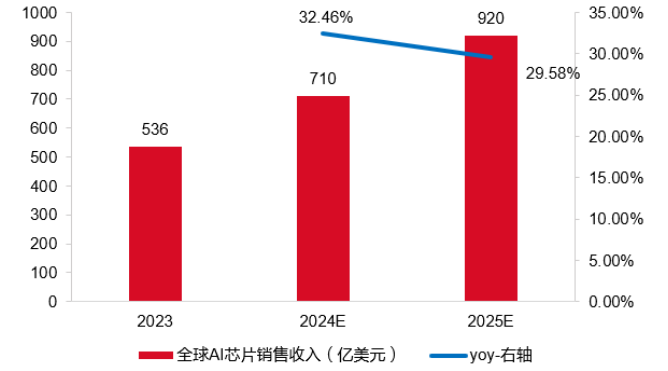
图22 2022-2025 年第一季度腾讯、百度资本开支（百万元人民币）



资料来源：公司公告，东海证券研究所

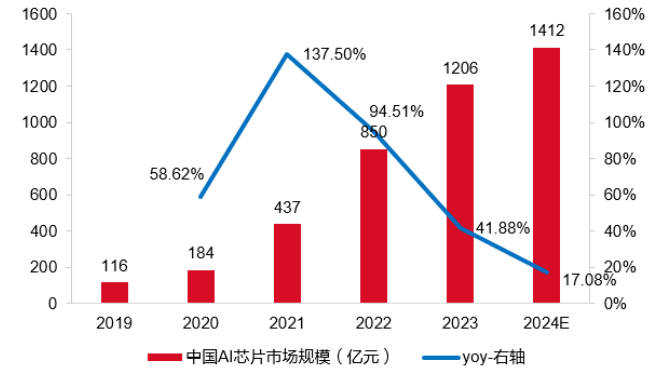
（6）算力需求增长以及 AI 基础设施建设采购需要拉动 AI 芯片市场规模扩大，2025 年全球 AI 芯片销售收入有望达到 920 亿美元，同比增长 29.58%，中国 AI 芯片市场规模 2024 年有望达到 1412 亿元，占比全球约 28%。AI 芯片是实现算力的核心硬件，芯片性能决定算力水平，AI 芯片通过不断优化制程工艺、架构设计等提升计算能力。随着全球算力需求不断上升，AI 芯片市场规模也在相应大增。全球 AI 芯片销售收入 2024 年有望达到 710 亿美元，同比增长 32.46%，2025 年有望继续增长 29.58%至 920 亿美元。中国作为算力需求的大国，2024 年 AI 芯片市场规模有望增至 1412 亿元，同比增长 17.08%，占比全球市场份额约为 27.92%。

图23 2023-2025E 年全球 AI 芯片销售收入及同比增速



资料来源：中国信息通信研究院，Gartner，东海证券研究所

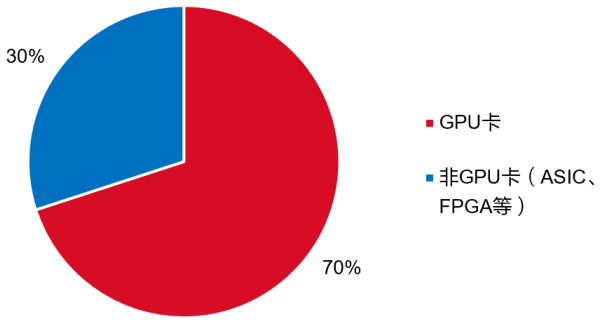
图24 2019-2024E 年中国 AI 芯片市场规模及同比增速



资料来源：中商产业研究院，东海证券研究所

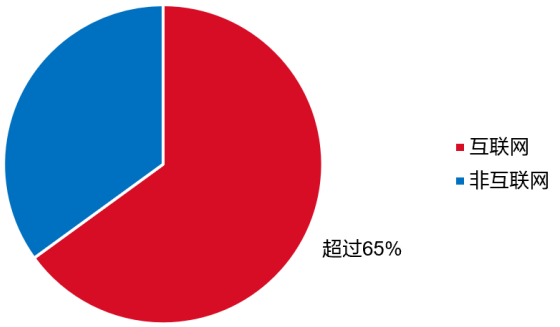
（7）从国内细分市场来看，2024 年我国 GPU 服务器仍占据 70%，但 ASIC、FPGA 服务器正加速增长，2029 年份额将接近 50%，从采购厂商看，互联网仍是最大的采购行业。根据 IDC，2024 年中国加速芯片市场中 GPU 卡占比达到 70%，占据主导地位，IDC 预测，到 2029 年中国加速服务器市场中非 GPU 服务器（ASIC 和 FPGA 等）市场规模将接近 50%。从行业角度看，互联网厂商采购了超过 65%的加速服务器，其余行业均有不同幅度的增长。

图25 2024 年我国各类加速芯片市场份额



资料来源：IDC，东海证券研究所

图26 2024 年我国服务器采购厂商份额

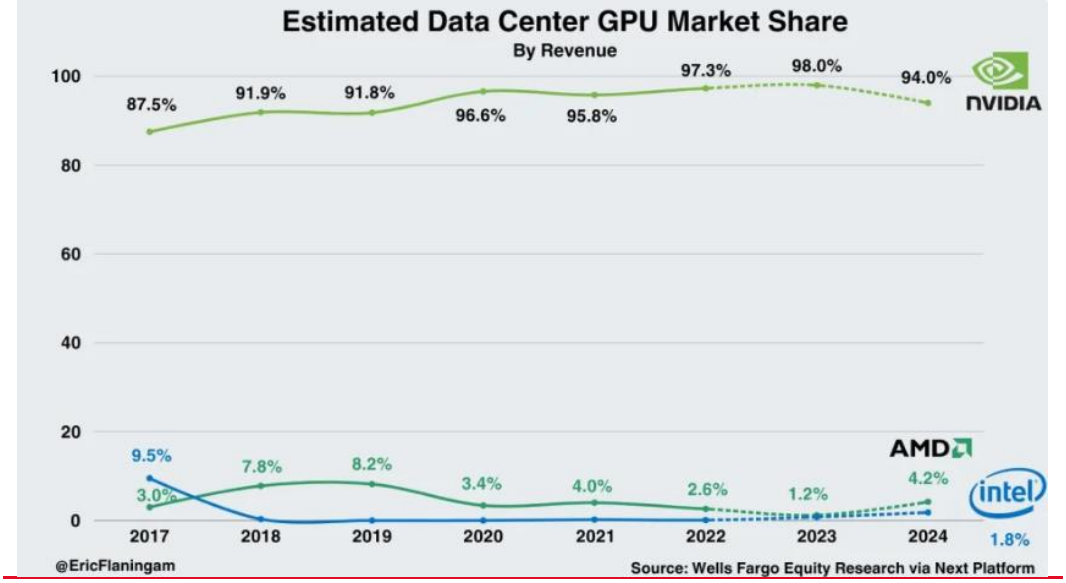


资料来源：IDC，东海证券研究所

2.3.海外芯片龙头仍主导市场，本土厂商加速布局提升份额

（1）英伟达仍在全球数据中心 GPU 市场占有绝对领先地位，市占率超过 90%。受益于生成式 AI 市场的爆发，英伟达（NVIDIA）在全球 AI 芯片市场遥遥领先，自 2017 年开始市占率从 87.5%持续上升至 2023 年的 98%，其余两位主要厂商为 AMD 和 Intel，但由于英伟达的 AI 芯片价格昂贵，且存在着供应不足的问题，因此一些客户希望选择其他厂商的替代产品，并且随着 AMD 和英特尔在 AI 芯片市场持续投入（比如推出 AMD Instinct MI300 系列、英特尔 Gaudi2/Gaudi3 等），其他 AI 芯片厂商的市场份额也将会逐步提升。2024 年英伟达全球数据中心 GPU 市场份额或略微回落至 94%，AMD 为 4.2%，Intel 为 1.8%。

图27 全球数据中心 GPU 市场份额

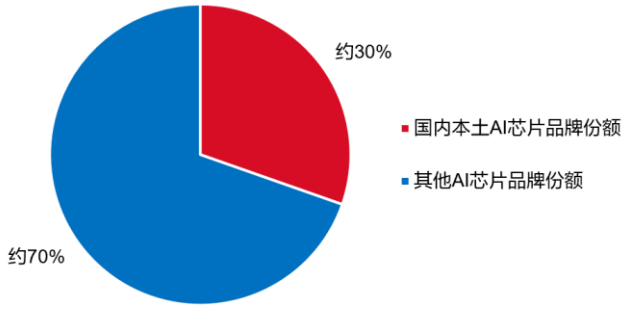


资料来源：富国银行，芯智讯，东海证券研究所

（2）国内 AI 芯片市场也主要被英伟达、AMD 等占据，但以华为、寒武纪、百度等厂商为代表的国产 AI 厂商正在加速布局，2024 年国内本土 AI 芯片品牌的出货份额已达 30%，预计 2025 年中国 AI Server 市场国内本土芯片供应商占比升至 40%。根据 IDC，随着我国加速芯片市场规模增长，本土 AI 芯片品牌的出货量也在不断上升，2024 年出货超过 82 万张，份额约为 30%，通过适配 DeepSeek 等国产 AI 大模型，中国本土芯片在软件生态领域实现了突破并逐步完善，本土芯片的市场竞争力也因此不断加强，同时也促进了本土芯片厂商的技术交流和资源共享，打破了国产芯片生态建设的僵局。根据 TrendForce，中国 AI Server 市场预计外购英伟达、AMD 等芯片比例会从 2024 年的约 63%下降至 2025 年的

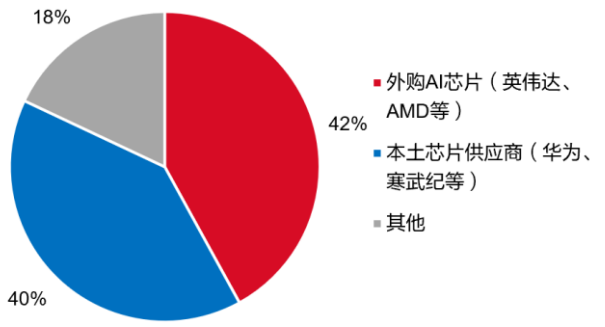
42%，而本土芯片供应商（如华为海思、寒武纪等）在国有 AI 芯片政策支持下，2025 年预计占比将提升至 40%。

图28 2024 年我国各类加速芯片市场份额



资料来源：IDC，东海证券研究所

图29 2025 年我国 AI 服务器芯片供应厂商份额



资料来源：IDC，东海证券研究所

（3）随着英伟达 GPU 架构迭代，其芯片算力、功耗表现也更为突出，拥有领先其他厂商 1-2 年的代际优势。英伟达成立于 1993 年，早期专注于图形芯片设计业务，1999 年英伟达挂牌上市，并发明和推出全球首款图形处理器 GeForce 256（GPU），极大推动了 PC 游戏市场的发展，重新定义了计算机图形技术。2006 年，英伟达发明并行计算平台和编程模型 CUDA，为后来的 AI 技术带来了重大影响。随着技术与业务的发展，目前已发展成为一家提供全栈计算的人工智能公司。自 2006 年的 Tesla 架构起，英伟达的 GPU 架构不断迭代，先后推出了 Volta（2017 年）、Turing（2018 年）、Ampere（2020 年）、Hopper（2022 年）、Blackwell（2024 年）等 GPU 架构，以及计划于 2026 年推出的 Rubin 架构，随着架构迭代，芯片制程越发先进，算力、功耗表现也更为突出。总体来说，英伟达的芯片拥有领先市面上其他竞争厂商 1-2 年的代际性能优势，同时 CUDA 平台让开发者形成路径依赖，“硬件+软件”共同构建其稳固的护城河。

图30 英伟达主要 AI 芯片性能

架构	芯片	算力	功耗	算力功耗比(TOPS/W)	制程	显存带宽
Volta	V100 PCIe	56 TOPS(INT8), 28 TFLOPS(FP16)	250W	0.22	12nm	900GB/s
	V100 SXM2	64 TOPS(INT8), 32 TFLOPS(FP16)	300W	0.21	12nm	900GB/s
	V100S PCIe	66 TOPS(INT8), 33 TFLOPS(FP16)	250W	0.26	12nm	1134GB/s
Ampere（A800为中国特供）	A100 SXM	1248 TOPS(INT8), 624 TFLOPS(FP16)	400W	3.12	7nm	
	A100 PCIe	624 TOPS(INT8), 312 TFLOPS(FP16)	300W	2.08	7nm	
	A800	874 TOPS(INT8), 437 TFLOPS(FP16)	240W	3.64	7nm	
Hopper（H800、H20为中国特供）	H100 SXM	3958 TOPS(INT8), 1979 TFLOPS(FP16)	最大700W	5.65	4nm	3.35TB/s
	H100 NVL	3341 TOPS(INT8), 1671 TFLOPS(FP16)	350-400W	8.35-9.55	4nm	3.9TB/s
	H800 SXM	3958 TOPS(INT8), 1979 TFLOPS(FP16)	700W	0.41	5nm	2.04TB/s
	H800 PCIe	3026 TOPS(INT8), 1513 TFLOPS(FP16)	300-350W	8.65-10.09	5nm	3.35TB/s
	H200 SXM	3958 TOPS(INT8), 1979 TFLOPS(FP16)	最大700W	5.65	4nm	4.8TB/s
	H200 NVL	3341 TOPS(INT8), 1671 TFLOPS(FP16)	最大600W	5.57	4nm	4.8TB/s
	HGX H20	286 TOPS(INT8), 148 TFLOPS(FP16)	400W	0.71	5nm	4.0TB/s
Blackwell（拟推出新的一款中国特供芯片）	HGX B200	72 POPS(INT8), 36 PFLOPS(FP16)	1000W	72	4nm	14.4TB/s
	HGX B300	2 POPS(INT8), 36 PFLOPS(FP16)			4nm	14.4TB/s

资料来源：英伟达官网，CSDN，东海证券研究所（注：SXM、NVL、PCIe 为不同的接口类型；HGX 为英伟达的服务器主板，通常包含 8 块对应芯片；INT8 为整数精度，FP16 为半精度）

（4）AI 芯片的设计和制造的复杂性使国内厂商难以在短期内比肩英伟达等国际巨头。国产 GPU 面临着技术门槛高、生态系统薄弱、研发投入巨大、市场进入壁垒高、制造工艺限制等问题，种种挑战使得中国本土 AI 芯片品牌还未能追赶上国际巨头最新世代产品的性能，市占率也亟待提高。

表3 国产 AI 芯片面临的挑战

挑战	具体内容
技术门槛高	现代 GPU 不仅要在图形处理上表现出色，还需要具备强大的计算能力以满足 AI 和大数据分析的需求。这意味着设计一个高性能的 GPU 涉及复杂的架构设计、精细的制造工艺以及高度优化的软件支持。国外巨头如英伟达和 AMD 在这些方面有着数十年的积累，而国产厂商尚处于追赶阶段。
生态系统薄弱	AI 芯片的成功不仅依赖于硬件本身，还需要一个完善的生态系统，包括驱动程序、开发工具和应用支持。国外厂商通过长期的市场占有，已经建立起了完善的生态系统，吸引了大量的开发者和用户。而国内在这一方面还处于起步阶段，生态系统的薄弱导致用户和开发者的接受度较低，进一步制约了市场份额的扩大。
研发投入巨大	AI 芯片研发需要巨大的资金和人力投入。从芯片设计、验证到流片，再到驱动和应用软件的开发，每个环节都需要大量的投入。国产厂商在资金和人才储备上相对有限，这使得他们难以在短时间内与国际巨头竞争。
市场进入壁垒高	首先是专利壁垒，国外厂商持有大量核心技术专利，国产厂商在设计过程中容易受到专利诉讼的威胁。其次是市场占有率，国际厂商通过长期的客户积累和品牌效应，已经占据了市场的主要份额，国产厂商在短时间内难以突破这些壁垒。
制造工艺限制	高端 AI 芯片的制造需要先进的制程工艺，目前全球只有少数几家公司具备这样的制造能力，如台积电和三星。国产厂商在制造工艺上还存在差距，这直接影响了本土芯片的性能和良品率。此外，受到国际形势和贸易政策的影响，国产厂商在获取先进制造设备和技术上也面临诸多限制。

资料来源：eefocus，东海证券研究所

（5）目前国内本土 AI 芯片品牌以寒武纪思元系列、华为昇腾、百度昆仑芯等为代表。

1）华为目前已有 NPU 芯片昇腾 310、910（昇腾 910B 可对标英伟达 A100）以及基于上述芯片的加速卡、AI 服务器、AI 集群等解决方案。华为拥有基于华为昇腾系列（HUAWEI Ascend）AI 处理器和基础软件构建的 Atlas 人工智能计算解决方案——昇腾计算，包括 Atlas 系列模块、板卡、小站、服务器、集群等丰富的产品形态，打造面向“端、边、云”的全场景 AI 基础设施方案，覆盖深度学习领域推理和训练全流程。目前华为发布了两款 AI 芯片（NPU）——昇腾 910 和昇腾 310，采用华为自主开发的达芬奇架构，昇腾 910 主要面向云端高性能计算，而昇腾 310 功耗较低，主要用在边缘计算等领域，基于昇腾芯片，华为开发了 AI 算力板卡、服务器、集群等一系列硬件产品。从性能上看，昇腾 910B 可对标英伟达 A100。

图31 华为昇腾 AI 芯片性能与相关加速卡性能

	算力	功耗	算力功耗比(TOPS/W)	制程工艺	带宽	用途
芯片						
昇腾310	16 TOPS(INT8), 8 TOPS (FP16)	8W		12nm		
昇腾910A	512 TOPS(INT8), 256 TFLOPS (FP16)	310W		台积电7nm增强版EUV工艺		
昇腾910B	640 TOPS(INT8), 320 TFLOPS(FP16)			中芯国际N+1工艺（等效7nm）		
昇腾910C （双die封装设计，将两颗910B封装在一起）				中芯国际N+2工艺（7nm）		
加速卡						
Atlas 300I Pro推理卡 （搭载1个Ascend 310P处理器）	140 TOPS(INT8), 70 TFLOPS(FP16)	最大72W	1.94		204.8GB/s	AI推理、目标检索
Atlas 300I DUO推理卡 （搭载2个Ascend 310P处理器）	280 TOPS(INT8), 140 TFLOPS(FP16)	150W	1.86		408GB/s	AI推理、视频分析
Atlas 300 V视频解析卡 （搭载1个Ascend 310P处理器）	100 TOPS(INT8), 50 TFLOPS(FP16)	72W	1.39		204.8GB/s	AI推理、 视频图片编解码
Atlas 300 V Pro视频解析卡 （搭载1个Ascend 310P处理器）	140 TOPS(INT8), 70 TFLOPS(FP16)	最大72W	1.94		204.8GB/s	AI推理、 视频图片编解码
Atlas 300T训练卡（型号9000） （搭载昇腾910 AI处理器）	440 TOPS(INT8), 220 TFLOPS(FP16)	最大300W	1.47			AI训练、 人机交互操作

资料来源：华为官网，CSDN，36Kr，东海证券研究所（注：2020 年华为被列入实体清单，自 910B 起代工厂转为中芯国际）

2）昆仑芯目前已有两代产品，可适用于云端训练、推理等场景。2018 年，百度在 2018 年百度 AI 开发者大会上宣布推出云端全功能 AI 芯片“百度昆仑”；2020 年昆仑芯 1 代系列产品大规模部署；2021 年 4 月，百度昆仑芯片业务完成独立融资，昆仑芯（北京）科技有限公司成立；2021 年 8 月，昆仑芯 2 代系列产品量产。昆仑芯 1 代 AI 芯片基于昆仑芯自研架构 XPU 设计，针对云端推理场景，支持通用 AI 算法，在计算机视觉、语音识别、自然语言处理和推荐的算法上性能指标表现高效且稳定。昆仑芯 2 代 AI 芯片基于新一代自研架构昆仑芯 XPU-R 而设计，聚焦通用性和易用性。相比 1 代产品，昆仑芯 2 代 AI 芯片的通用计算核心算力提升 2-3 倍，可为数据中心提供强劲 AI 算力。

表4 百度昆仑芯性能

芯片	算力	用途	制程	带宽
一代昆仑芯 818-100 （推理芯片）	128 TOPS(INT8), 32 TOPS (XFP16)	推理	14nm	256GB/s
一代昆仑芯 818-300 （训练芯片）	256 TOPS(INT8), 64 TOPS (XFP16)	训练	14nm	512GB/s
二代昆仑芯	128 TFLOPS(FP16)		7nm	512GB/s

资料来源：百度，中国日报网，东海证券研究所（注：XFP16/32 指 XPU FP16/32，是百度昆仑芯片自定义的数据格式，对软件提供标准的 FP16/32 接口，但能实现比标准 FP16/32 更高的计算精度）

（6）美国 AI 芯片管制政策趋严，国产 AI 芯片自主可控进程有望全面提速。从管制阈值看，美对华 AI 芯片出口管制从最初限制尖端算力芯片（如 A100/H100），逐步扩展至特供版中端芯片（H20）；从管制范围看，除硬件外，美国同步切断 EDA 工具更新、禁用华为昇腾芯片等，并威胁全球企业配合制裁，形成“硬件+软件+生态”三位一体封锁网。短期内，或导致国内算力缺口与技术断链，长期看，在国内厂商研发能力提升和技术升级的背景下，构建“设计-制造-生态”全自主产业链正在加速催化，AI 芯片国产替代有望全面提速。

表5 近期 AI 芯片相关管制政策

政策日期	具体内容
2024/12/2	美国商务部公布两份最新的出口管制文件总计 210 页，涉及具体的出口管制条例调整和实体清单明细更新，其中实体清单中中国芯片企业占 130 多家，名单就长达 58 页，月内生效。本轮出口管制条例调整有两大主题：限制中国获得尖端高算力的人工智能芯片，遏制中国的先进芯片制造能力。
2025/4/15	根据英伟达最新披露的 8-K 文件显示，英伟达面向中国市场“特供”的 H20 也已经被列入了出口管制，必须要有许可证才可出口。英伟达在文件中表示，美国政府已于当地时间 4 月 9 日通知英伟达，要求其向中国（含港澳）和 D:5 国家组，或总部或最终母公司位于该地区的企业，出口英伟达 H20 和实现 H20 存储带宽、互连带宽或其组合的任何集成电路都将需要获得许可证。
2025/5/13	美国商务部正式发布文件废除拜登政府的人工智能扩散规则，同时宣布采取三项额外政策以加强对全球 AI 芯片的出口管制，其中就包括认定在世界任何地方使用华为昇腾芯片均违反美国的出口管制规定，包括使用中国 3A090 集成电路比如华为昇腾 910B/C/D，可能会受到工业与安全局的执法行动，这些行动可能包括严重的刑事和行政处罚，直至包括监禁、罚款、丧失出口特权或其他限制。
2025/6/2	美国商务部工业和安全局通知全球三大 EDA 芯片设计厂商，要求停止对整个中国大陆地区的 EDA 服务与支持。Cadence、Synopsys、Simens 都在禁售之列，三者合计的全球份额高达 74%。

资料来源：观察者网，芯智讯，东海证券研究所

3.硬件筑基+软件赋能，造就国产 AI 芯片领跑者

3.1.云、边、端全栈布局，赋能训练与推理全场景

（1）从硬件层面看，公司在云、边、端三大场景都有芯片产品布局。AI 芯片按照应用场景的不同可分为云端、边缘端和终端三类，云端主要是指云计算数据中心等场景，边缘端主要指智能制造、智能家居、智慧交通、智能驾驶等场景，终端则是各类消费电子、IoT 产品等，上述场景对应硬件的算力和功耗需求都有所不同，总体来说云端 AI 芯片需求的算力和对应功耗较高，边缘端次之，终端对于算力和功耗的要求较低。公司面向云、边、端三大场景分别研发了三种类型的芯片产品，分别为云端智能芯片及加速卡、边缘智能芯片及加速卡和终端智能处理器 IP。

表6 云、边、端三种场景对于芯片的算力和功耗等特性要求

应用场景	芯片需求	典型算力	典型功耗	典型应用场景
云端	高性能、高计算密度、兼有推理和训练任务、单价高、硬件产品形态少	>30TOPS	>50W	云计算数据中心、企业私有云等
边缘端	对功耗、性能、尺寸的要求常介于终端与云端之间、推理任务为主、多用于插电设备、硬件产品形态相对较少	5TOPS-30TOPS	4-15W	智能制造、智能家居、智能零售、智慧交通、智慧金融、智慧医疗、智能驾驶等
终端	低功耗、高能效、推理任务为主、成本敏感、硬件产品形态众多	<8TOPS	<5W	各类消费类电子、物联网产品等

资料来源：公司公告，东海证券研究所（注：云、边、端应用场景尚无标准划分界限，该表为寒武纪基于自主研发技术体系划分）

（2）公司云端产品线包含云端智能芯片及加速卡、训练整机。云端智能芯片及加速卡需与服务器整机产品进行适配，通过服务器厂商、OEM 厂商针对其功能和性能的全方位严格认证再进入大规模商用阶段，公司除了要攻克智能芯片架构等一系列核心技术难关，还要跨越各服务器厂商的高准入门槛。训练整机主要提供计算集群中的单体训练服务器，由公司自研云端智能芯片及加速卡提供核心计算能力，且整机亦是公司自研的服务器产品。

1）云端 AI 芯片方面，公司已经迭代发布了四代产品以及其对应加速卡，思元 370 最大算力达 256TOPS(INT8)，集推理训练为一体，MLU370-X8 单卡性能与主流 350W RTX GPU 相当，思元 590 有望成为新的营收支撑。思元 100 芯片于 2018 年发布，是中国首款高峰值云端智能芯片。思元 270 在前一代基础上升级了指令集和芯片架构，是公司首款云端训练智能芯片，思元 290 芯片工艺为台积电 7nm 制程工艺，可高效支持分布式、定点化的人工智能训练任务，2021 年，公司发布了“推训一体”的思元 370，是公司首款采用 Chiplet（芯粒）技术的人工智能芯片（支持芯粒间的灵活组合，仅用单次流片就达成了多款智能加速卡产品的商用），芯片最大算力高达 256TOPS(INT8)，是思元 270 算力的 2 倍。同时，思元 370 芯片支持 LPDDR5 内存，内存带宽是思元 270 的 3 倍，可在板卡有限的功耗范围内给人工智能芯片分配更多的能源，输出更高的算力。通过在 Cambricon NeuWare SDK 上实测，在常见的 4 个深度学习网络模型上，MLU370-X8 单卡性能与主流 350W RTX GPU 相当。从客户层面看，公司已与互联网、金融、通信、交通等多个行业客户展开合作，与头部 AI 大模型进行适配，并在各行业垂直领域进行大模型应用探索与落地。目前新一代思元 590 芯片已进入国产供应链，实测训练性能较在售产品有了显著提升，它提供了更大的内存容量和更高的内存带宽，其 PCIe 接口也较上代实现了升级，有望成为公司新的营收支撑。

图32 公司云端 AI 芯片以及对应加速卡

芯片	推出时间	加速卡	算力	制程	内存容量	内存位宽	内存带宽	接口	功耗	应用场景
思元100 (MLU100)	2018		128 TOPS(INT8,稀疏), 32 TOPS(INT8,非稀疏), 64 TOPS(FP16,稀疏), 16 TOPS(FP16,非稀疏)	台积电 16nm					推理场景典型功耗小于75W	面向人工智能云端推理任务
思元270 (MLU100)	2019	MLU270-S4	256TOPS(INT4), 128TOPS(INT8), 64TOPS(INT16), 同时支持FP32，FP16计算精度	台积电 16nm	16GB DDR4, ECC	256 bit	102 GB/s	×16 PCIe Gen.3	最大热设计功耗70w、被动散热	在思元100基础上应用范畴拓展至人工智能训练，集成了丰富的视频图像编解码硬件单元
		MLU270-F4							最大热设计功耗(TDP)150W、最大整板功耗(TBP)160W、主动散热	
思元290 (MLU290)	2020	MLU290-M5	512 TOPS (INT8), 256 TOPS (INT16), 64 TOPS (CINT32)	台积电 7nm	HBM2高带宽内存, 32GB	4096 bit	1228 GB/s	×16 PCIe 4.0	训练场景典型功耗小于350W	面向复杂人工智能模型的云端训练任务
思元370 (MLU370)	2021	MLU370-S4/S8	192 TOPS (INT8), 96 TOPS (INT16), 72 TFLOPS (FP16), 72 TFLOPS (BF16), 18 TFLOPS (FP32),	台积电 7nm	LPDDR5，24GB/48GB		307.2 GB/s	x16 PCIe Gen4	最大热功耗75W，被动散热	可在服务器中实现高密度部署，在视频编解码方面具有较强竞争力
		MLU370-X4	256 TOPS (INT8), 128 TOPS (INT16), 96 TFLOPS (FP16), 96 TFLOPS (BF16), 24 TFLOPS (FP32)	台积电 7nm	LPDDR5, 24GB		307.2 GB/s	x16 PCIe Gen4	最大热功耗150W，被动散热	可充分满足推训一体AI任务需求
		MLU370-X8（采用双芯思元370配置）	256 TOPS (INT8), 128 TOPS (INT16), 96 TFLOPS (FP16), 96 TFLOPS (BF16), 24 TFLOPS (FP32)	台积电 7nm	LPDDR5, 48GB		614.4 GB/s	x16 PCIe Gen4	最大热功耗250W，被动散热	在常见的4个人工智能模型上，MLU370-X8单卡性能与主流350W RTX GPU相当，可高效执行多芯多卡训练和分布式推理任务

资料来源：公司公告，公司官网，东海证券研究所（注：算力数据为在 1GHz 主频下的理论峰值；非稀疏理论峰值性能代表处理非稀疏深度学习模型的理论最高性能，稀疏等效理论峰值性能代表处理稀疏深度学习模型的等效理论最高性能）

2）目前寒武纪发布了玄思 1000、1001 智能加速器整机，机箱内集成了多颗思元智能芯片。玄思 1000 智能加速器整机在 2U 机箱内集成了 4 颗思元 290 智能芯片，2 台玄思 1000 加速器与 CPU 服务器可组成一套包括 8 张加速卡的整机系统，可实现 AI 算力多向扩展，满足性能、扩展性、灵活性、鲁棒性的要求。2022 年发布的玄思 1001 智能加速器在 2U 机箱内集成 4 张 MLU370-M8 智能加速卡，MLU-Link 互联接口，实现智能算力在数据中心纵向扩展，可广泛支持 FP16、FP32 等不同数据精度的智能算力，提供大容量内存，支撑智能模型的分布式训练需求，是智能算力的高集成度平台，已在生物信息、医疗影像、语言模型等行业及科研场景广泛应用。

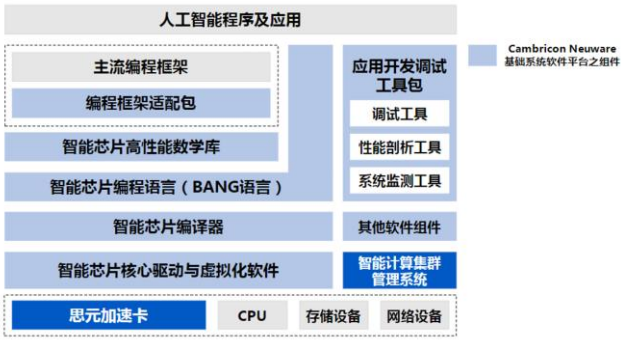
表7 公司训练整机性能

型号	智能加速卡支持	算力	内存容量	CPU 上联速率	典型功耗
玄思 1000 (MLU-X1000)	4× MLU290-M5	2.05 PetaOPS (INT8), 1 PetaOPS (INT16), 256 TOPS (CINT32), 同时支持 FP32, FP16	128GB	× 16 PCIe 4.0, 64GB/s Bi-direction	2300W
玄思 1001 (MLU-X1001)	4× MLU370-M8	-	-	-	-

资料来源：公司官网，公司公告，东海证券研究所

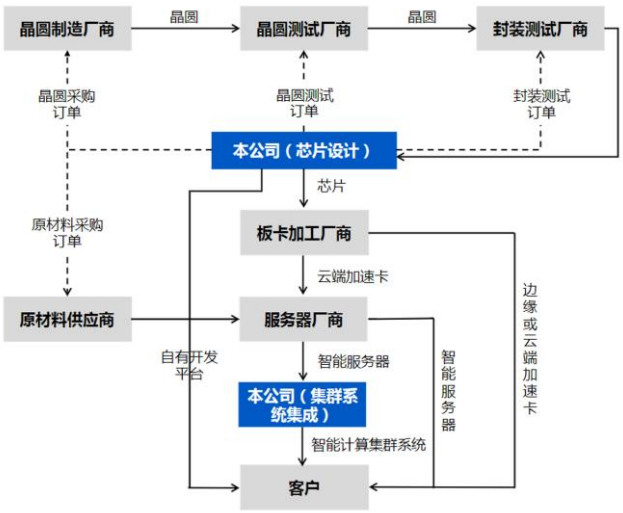
（3）公司的智能计算集群系统业务提供全集群搭建和管理服务，主要面向有一定技术基础的商业客户群体，国内市占率位于第一梯队。对于有人工智能计算能力建设的客户来说，部分客户选择单独采购云端智能芯片加速卡并将其自行集成至现有建设完毕的计算集群中，但部分客户则更希望公司能够提供定制化的软硬件整体解决方案，以科学地配置和管理集群的软硬件、提升运行效率。2021 年公司中标昆山智能计算中心等项目，公司已经陆续在西安沣东、珠海横琴、江苏南京、江苏昆山拓展了智能计算集群系统业务，国内市占率处在第一梯队；2022 年公司中标南京智能计算中心项目（二、三期）智能计算设备（二期）项目，以玄思 1001 智能加速器作为核心算力单元之一，集成配套软硬件，最终形成智能计算集群系统交付给客户；2023 年公司积极参与台州、沈阳两地的算力基础设施建设项目并交付相关智能计算集群系统。

图33 公司智能计算集群系统软硬件的总体架构



资料来源：公司公告，东海证券研究所

图34 公司智能计算集群系统的整体业务流程



资料来源：公司公告，东海证券研究所

（4）边缘端产品方面，公司围绕思元 220 芯片推出了相应加速卡及智能模组，面向 AI 边缘推理任务。边缘计算通过在终端和云端之间的设备上配备适度的计算能力，可有效弥补终端设备计算能力不足的劣势，同时能够缓解云计算场景下数据安全、隐私保护、带宽与延时等潜在问题。边缘计算范式和 AI 技术的结合将推动智能制造、智能零售、智能教育、智能家居、智能电网、智能交通等众多领域的高速发展。公司于 2019 年推出了边缘智能芯片思元 220 及相应的 M.2 加速卡，思元 220 基于台积电 16nm 工艺，在 1GHz 的主频下，理论峰值性能为 32TOPS(INT4)、16TOPS(INT8)、8TOPS(INT16)，芯片典型功耗小于 10W，支持视觉、语音、自然语言处理以及传统机器学习等多样化的 AI 应用。同时公司推出了 MLU220-SOM 智能模组，基于信用卡大小的模组上可以实现 16TOPS AI 算力的单系统解决方案，功耗仅为 15W。

表8 公司边缘端产品性能

型号	算力	CPU	内存	存储空间	接口	功耗	散热
MLU220-M.2	8TOPS (INT8)	-	LPDDR4x, 64 bits	-	M.2 2280, B+M key (PCIe3.0 x2)	8.25W	被动散热
MLU220-SOM	16TOPS (INT8)	ARM A55 ×4, 1.5GHz	8GB LPDDR4x, 64bits, 3733MHz	32GB eMMC 5.1 HS400 400MB/s	PCIe3.0 2x2 (RC); SDIO3.0×2	≤15W	裸板，客户根据自身情况进行散热设计

资料来源：公司官网，东海证券研究所

（5）终端产品方面，公司先后推出了寒武纪 1A、1H、1M 系列芯片，覆盖 0.5TOPS-8TOPS 内不同档位的 AI 算力需求，可集成于手机或 IoT 类 SoC 芯片中，从而快速获得在终端做 AI 本地处理的能力。公司终端智能处理器产品主要以 IP 授权形式于智能终端设备中，即将已完成逻辑设计或物理设计的芯片功能模块（如处理器、DRAM 接口等）以商业授权的形式交付给客户使用，允许客户将其集成在自己的芯片设计版图中，并通过流片形成最终芯片产品。公司收费模式包括固定费用（许可技术通过验收后，许可产品正式出货前，按照授权许可实施进度分阶段收取相应费用）和提成费用（被授权方量产芯片并销售许可产品后的每个季度末，按照许可产品的累计销售数量所在区间分标准收取相应费用），因此该项业务基本不产生对应成本。目前已有多家国内著名芯片设计公司获得了公司终端智能处理器的商业 IP 授权，迄今已集成于上亿台智能手机及其他智能终端设备中。

图35 公司终端智能处理器产品性能

型号	推出时间	版本	技术指标	产品特点
寒武纪1A	2016		1GHz主频下，非稀疏理论峰值性能0.5TOPS (FP16)，稀疏等效理论峰值性能2TOPS (FP16)	全球首款商用终端智能处理器IP产品，可支持视觉、语音和自然语言处理等消费电子领域的人工智能应用；搭载寒武纪1A的某旗舰手机芯片在AI应用上达到了4核CPU25倍以上的性能和50倍以上的能效，采用该手机芯片的旗舰手机产品每分钟可识别2005张图片
寒武纪1H	2017	Cambricon-1H8mini（轻量级版本）	使用256MAC 8位定点运算器。在1GHz主频下，峰值速度为0.5TOPS(INT8)	功耗和面积等指标较上一代产品有显著提升，支持双核模式，并增加了对8位定点(INT8)AI运算的支持。搭载寒武纪1H的某旗舰手机芯片，每分钟可识别4500张图片，是上一代产品的2.2倍
		Cambricon-1H8（中量级版本）	使用512MAC 8位定点运算器。在1GHz主频下，非稀疏理论峰值性能1TOPS (INT8)	
		Cambricon-1H16（高端版本）	使用256MAC 16位浮点运算器以及512MAC 8位定点运算器。1GHz主频下，非稀疏理论峰值性能0.5TOPS (FP16)或1TOPS (INT8)，稀疏等效理论峰值性能2TOPS (FP16)	
寒武纪1M	2018	Cambricon-1M-1K（轻量级版本）	使用了1024MAC 8位定点运算器。在1GHz主频下，进行8位定点AI运算的峰值速度为2TOPS，进行16位定点AI运算的峰值速度为1TOPS，进行32位定点AI运算的峰值速度为0.25TOPS	针对7nm等先进工艺作了专门优化，进一步提升了处理器性能和能效；提供不同性能档位的处理器配置，支持多核模式；在业界率先支持定点化训练，可在终端支持AI训练任务
		Cambricon-1M-2K（中量级版本）	使用了2048MAC 8位定点运算器。在1GHz主频下，进行8位定点AI运算的峰值速度为4TOPS，进行16位定点AI运算的峰值速度为2TOPS，进行32位定点AI运算的峰值速度为0.5TOPS	
		Cambricon-1M-4K（高端版本）	使用了4096MAC 8位定点运算器。在1GHz主频下，进行8位定点AI运算的峰值速度为8TOPS，进行16位定点AI运算的峰值速度为4TOPS，进行32位定点AI运算的峰值速度为1TOPS	

资料来源：公司公告，公司官网，东海证券研究所

3.2.基础系统软件平台&自研架构与指令集构建护城河

（1）在提供硬件的同时，公司也为云、边、端全系列智能芯片与处理器产品提供统一的平台级基础系统软件 Cambricon Neuware（包含软件开发工具链等）。Cambricon Neuware 打破了不同场景之间的软件开发壁垒，兼具高性能、灵活性和可扩展性的优势，无须繁琐的移植即可让同一 AI 应用程序便捷高效地运行在公司云边端系列化芯片与处理器产品之上。在 Cambricon Neuware 的支持下，程序员可实现跨云边端硬件平台的 AI 应用开发，大幅提升 AI 应用在不同硬件平台的开发效率和部署速度，同时也使云边端异构硬件资源的统一管理、调度和协同计算成为可能。

图36 Cambricon NeuWare 一览

资料来源：公司官网，东海证券研究所

（2）公司软件平台可分为训练软件平台和推理软件平台。（1）训练软件平台方面，公司拥抱开源生态，研发了兼具高性能和通用性的训练软件栈，原生支持业界的开源框架 Pytorch 和 Tensorflow，对两个框架都提供了完善的基础设施支持，包括原生 Profiler 和原生的分布式训练支持，用户基于开源框架的模型代码可以快速完成迁移。截至 2024 年底，公司持续投入在大规模分布式训练软件平台的研发，迭代更新了 Megatron、Transformer Engine 等主流分布式训练组件，使训练软件平台能够支撑主流的大模型分布式训练需求，降低新模型的适配周期，同时增加了对 DeepSeek 系列、Llama 系列、Qwen 系列等主流大模型训练的支持。（2）推理软件平台方面，公司于 2021 年发布全新推理加速引擎 MagicMind，是业界首个基于 MLIR 图编译技术达到商业化部署能力的推理引擎。借助 MagicMind，用户仅需投入极少的开发成本，即可将推理业务部署到公司全系列产品上。截至 2024 年底，在大模型适配方面，推理软件平台成功支持并优化了 DeepSeek 系列、Llama 系列、Qwen 系列等主流文生文模型，以及 Flux、hunyuanvideo、cogvideox 等多模态模型。

图37 公司训练软件平台

资料来源：公司官网，东海证券研究所

图38 MagicMind 架构

资料来源：公司官网，东海证券研究所

（3）公司云端、边缘端、终端的所有智能芯片和处理器 IP 产品以及基础系统软件均基于自研处理器架构，且均构建于自研的 MLU 指令集基础之上，有助于保持核心技术的自主可控。思元 590 将采用 MLUarch05 全新架构。通用型智能芯片及其基础系统软件的研发需要全面掌握核心芯片与系统软件的大量关键技术，技术难度大、涉及方向广，是一个极端复杂的系统工程，其中处理器微架构与指令集两大类技术属于最底层的核心技术。（1）智能处理器微架构方面，目前公司已自主研发了四代智能处理器微架构（MLUarch00、MLUarch01、MLUarch02 和 MLUarch03），其中思元 370 基于 MLUarch03 计算架构，思元 590 将采用 MLUarch05 全新架构。（2）指令集是处理器芯片生态的基石，公司是国际上最早开展智能处理器指令集研发的少数几家企业之一，自 2016 年来已自主研发了四代商用智能处理器指令集（MLUv00、MLUv01、MLUv02 和 MLUv03），同一套指令集能够同时支持 AI 训练和推理任务，适用于云端、边缘端、终端不同场景不同类型的智能芯片，支撑公司构建云边端一体化、训练推理融合的基础系统软件平台和具有公司特色的 AI 新生态。截至 2024 年末，公司新一代智能处理器微架构和指令集正在研发中，将对自然语言处理大模型、视频图像生成大模型以及垂直类大模型的训练推理等场景进行重点优化，将在编程灵活性、易用性、性能、功耗、面积等方面提升产品竞争力。

图39 公司智能处理器微架构迭代

资料来源：公司公众号，东海证券研究所

4.盈利预测

4.1.盈利预测假设与业务拆分

根据公司公告披露的业务拆分，我们将寒武纪的业务分为云端产品线、边缘端产品线、IP 授权及软件与其他业务并分别作盈利预测，其中：

（1）云端产品线：2024 年为公司主要营收来源，2024 年以前我们认为这部分营收大部分由智能计算集群系统业务体现，且 2024 年智能计算集群系统营收从 2023 年的占比 85.22%缩减为 0，因此后续我们暂且不将这部分业务单独列示，合并至云端产品线业务中。2024 年公司云端产品线营收 11.66 亿元，占比营收的 99.30%。考虑到公司思元 370、590 系列及后续高端云端芯片逐步落地，以及公司目前大额的存货及预付账款，叠加目前中美政策不稳定的局面下国产 AI 芯片自主可控程度亟待加速，公司云端芯片、加速卡以及相关服务器等营收有望受益于国内互联网、政务相关高算力需求及大额 AI 基础建设资本开支，后续将实现高速增长态势，我们预计公司云端产品线 2025-2027 年营收分别为 84.35、161.61、250.94 亿元，同比增长 623.26%、91.59%、55.27%。毛利率方面，尽管公司产品单价对比市场平均略低，但综合成本同样较低，考虑到后续批量订单落地，规模效应凸显且市场竞争力上升，毛利率有望上升至 60%以上。

（2）边缘端产品线：公司边缘端产品主要为思元 220 系列芯片及加速卡等产品，但近年来营收占比逐年缩减，2024 年仅占比 0.56%，或因市场竞争较为激烈以及公司业务重心主要在云端产品线方面，我们预计边缘端产品线 2025-2027 年营收分别为 623、715、822 万元，毛利率分别为 51.69%、51.47%、50.12%。

（3）IP 授权及软件：这部分主要包括公司智能终端处理器 IP 以及基础软件系统，其中基础系统软件方面公司尚未对其进行单独销售、主要配合云、边、终端产品线的推广和销售，智能处理器方面目前公司共有寒武纪 1A、1H、1M 系列芯片，这部分业务公司主要按照 IP 授权的方式进行收费，包含固定费用以及提成费用，且几乎不存在成本。2024 年公司 IP 授权及软件业务营收占比 0.04%，整体占比逐年下滑，我们预计 2025-2027 年公司 IP 授权及软件营收为 51、53、55 万元。

表9 2022-2027E 寒武纪分业务营收及毛利率预测（百万元）

	2022	2023	2024	2025E	2026E	2027E
总营收	**729.03**	**709.39**	**1174.46**	**8443.34**	**16170.54**	**25104.50**
总毛利率	**65.76%**	**69.16%**	**56.71%**	**60.57%**	**62.96%**	**63.01%**
云端产品线	**219.45**	**90.57**	**1166.28**	**8435.25**	**16161.45**	**25094.26**
- yoy	173.52%	-58.73%	1187.78%	623.26%	91.59%	55.27%
- 毛利率	63.47%	60.63%	56.69%	60.57%	62.96%	63.01%
边缘端产品线	**37.84**	**10.82**	**6.54**	**6.23**	**7.15**	**8.22**
- yoy	-78.40%	-71.39%	-39.56%	-4.77%	14.77%	14.97%
- 毛利率	31.11%	55.88%	51.50%	51.69%	51.47%	50.12%
智能计算集群系统	**458.51**	**604.53**				
- yoy	0.64%	31.85%				
- 毛利率	70.42%	70.78%				
IP 授权及软件	**1.14**	**0.23**	**0.41**	**0.51**	**0.53**	**0.55**
- yoy	-83.44%	-79.45%	76.38%	23.67%	3.92%	3.77%
- 毛利率	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

资料来源：公司公告，东海证券研究所

盈利预测结果：我们对公司 2025-2027 年各类费用等进行了预测，最终预计公司 2025-2027 年归母净利润分别为 15.95、38.60 和 69.13 亿元，同比增速分别为 452.69%、141.96%和 79.09%。

表10 2022-2027E 寒武纪盈利预测结果（百万元）

	2022	2023	2024	2025E	2026E	2027E
营业总收入	**729.03**	**709.39**	**1,174.46**	**8,443.34**	**16,170.54**	**25,104.50**
营业成本	249.62	218.80	508.41	3,329.39	5,989.93	9,285.58
税金及附加	3.22	4.35	2.97	25.33	50.13	80.33
销售费用	82.71	82.07	70.06	493.94	926.57	1,410.87
管理费用	296.63	153.56	180.53	336.04	684.01	878.66
研发费用	1,523.11	1,117.51	1,215.87	2,641.08	4,616.69	6,353.95
财务费用	-78.40%	-71.39%	-39.56%	-4.77%	14.77%	14.97%
营业利润	**-1,324.25**	**-875.81**	**-455.75**	**1,644.33**	**3,978.56**	**7,125.07**
营业外收支	1.52	1.07	-0.02	0.00	0.00	0.00
所得税	2.01	3.34	1.16	32.89	79.57	142.50
净利润	-1,324.74	-878.08	-456.93	1,611.45	3,898.99	6,982.57
归母净利润	**-1,256.35**	**-848.44**	**-452.34**	**1,595.33**	**3,860.00**	**6,912.75**

资料来源：公司公告，东海证券研究所

4.2.可比公司估值

公司主营业务为 AI 芯片，我们选取海光信息、龙芯中科、景嘉微和瑞芯微作为可比公司。

截至 6 月 27 日，上述可比公司的 2025-2027 年平均 PE 为 583、346、117 倍，考虑到公司为国内稀缺的云端 AI 芯片厂商，且目前已经扭亏为盈，净利润增速较快，受益于当前算力需求暴涨以及国产替代加速的时代背景，我们看好公司的长期发展，预计对应当前市值的 2025-2027 年 PE 分别是 153、63、35 倍。

此外，考虑到 AI 芯片企业前期需要较高的研发投入，部分公司存在具备一定销售规模但尚未盈利的状态，我们采用了 PS 估值，上述可比公司 2025-2027 年平均 PS 分别为 40、28、20 倍，考虑到寒武纪云端 AI 芯片放量较快，营收大幅增长，我们预计寒武纪对应当前市值的 2025-2027 年 PS 分别是 29、15、10 倍。

表11 可比公司 PE 估值

股票代码	公司简称	市值（亿元）	EPS（元/股）			PE（倍）		
			2025E	2026E	2027E	2025E	2026E	2027E
688041.SH	海光信息	3231	1.35	1.94	2.66	102.64	71.68	52.32
688047.SH	龙芯中科	557.4	-	0.16	0.61	-	887.77	227.06
300474.SZ	景嘉微	377.9	0.04	0.20	0.48	1574.61	368.81	150.65
603893.SH	瑞芯微	638.8	2.11	2.80	3.83	72.43	54.46	39.79
	可比公司均值		0.80	1.28	1.90	583.23	345.68	117.46
688256.SH	寒武纪	2444	3.82	9.25	16.56	153.21	63.32	35.36

资料来源：携宁，除寒武纪外均为同花顺一致预期，东海证券研究所（截止至 2025 年 6 月 27 日）

表12 可比公司 PS 估值

股票代码	公司简称	市值（亿元）	总营收（亿元）			PS（倍）		
			2025E	2026E	2027E	2025E	2026E	2027E
688041.SH	海光信息	3231	138.54	193.04	260.17	23.32	16.74	12.42
688047.SH	龙芯中科	557.4	7.32	10.34	14.39	76.15	53.91	38.74
300474.SZ	景嘉微	377.9	8.35	12.68	19.90	45.26	29.80	18.99
603893.SH	瑞芯微	638.8	42.26	53.85	67.87	15.12	11.86	9.41
	可比公司均值					39.96	28.08	19.89
688256.SH	寒武纪	2444	84.43	161.71	251.04	28.95	15.11	9.74

资料来源：携宁，除寒武纪外均为同花顺一致预期，东海证券研究所（截止至 2025 年 6 月 27 日）

4.3.投资建议

首次覆盖，给予买入评级。作为国内稀缺的云端 AI 芯片标的公司，公司云端产品线正大幅放量中，思元系列芯片产品受益于国内各产业算力需求的提升以及国产替代的趋势，有望带动公司整体营收高增以及归母净利润的持续盈利。我们预计公司 2025-2027 年营业收入分别为 84.43、161.71 和 251.04 亿元，同比增速分别为 618.91%、91.52%和 55.25%；归母净利润分别为 15.95、38.60 和 69.13 亿元，同比增速分别为 452.69%、141.96%和 79.09%。对应 2025-2027 年的 PE 分别为 153、63、35 倍，PS 分别为 29、15、10 倍。

5.风险提示

（1）产品研发及验证进度不及预期风险：公司有多款芯片产品正处于客户端放量阶段，且有新款芯片在研中，若进展不及预期，或将导致相关产品盈利贡献低于预期。

（2）地缘政治风险：目前中美关系正处于博弈阶段，半导体相关政策走向尚不明朗，若紧张局势进一步升级，或导致国内半导体供应链风险加剧，进一步影响公司业绩；

（3）宏观经济下行风险：若整体经济下行，或影响下游客户 AI 基建和算力需求，进而影响公司产品销售。

附录：三大报表预测值

利润表

单位：（百万元）	2023A	2024E	2025E	2026E
营业总收入	**1,174**	**8,443**	**16,171**	**25,104**
%同比增速	66%	619%	92%	55%
营业成本	508	3,329	5,990	9,286
毛利	666	5,114	10,181	15,819
%营业收入	57%	61%	63%	63%
税金及附加	3	25	50	80
%营业收入	0%	0%	0%	0%
销售费用	70	494	927	1,411
%营业收入	6%	6%	6%	6%
管理费用	181	336	684	879
%营业收入	15%	4%	4%	4%
研发费用	1,216	2,641	4,617	6,354
%营业收入	104%	31%	29%	25%
财务费用	-19	6	14	24
%营业收入	-2%	0%	0%	0%
资产减值损失	-53	-100	-120	-150
信用减值损失	137	-80	-100	-150
其他收益	220	169	243	251
投资收益	23	42	65	100
净敞口套期收益	0	0	0	0
公允价值变动收益	0	0	0	0
资产处置收益	1	2	2	3
营业利润	**-456**	**1,644**	**3,979**	**7,125**
%营业收入	-39%	19%	25%	28%
营业外收支	0	0	0	0
利润总额	**-456**	**1,644**	**3,979**	**7,125**
%营业收入	-39%	19%	25%	28%
所得税费用	1	33	80	143
净利润	-457	1,611	3,899	6,983
%营业收入	48%	453%	142%	79%
归属于母公司的净利润	**-452**	**1,595**	**3,860**	**6,913**
%同比增速	-39%	19%	24%	28%
少数股东损益	-5	16	39	70
EPS（元/股）	-1.08	3.82	9.25	16.56

资产负债表

单位：（百万元）	2023A	2024E	2025E	2026E
货币资金	1,986	1,564	2,286	5,419
交易性金融资产	760	460	260	160
应收账款及应收票据	314	1,398	2,650	4,063
存货	1,774	2,859	4,772	6,852
预付账款	774	1,332	2,396	3,714
其他流动资产	191	318	473	646
流动资产合计	5,800	7,931	12,836	20,855
长期股权投资	247	267	287	307
投资性房地产	0	0	0	0
固定资产合计	231	338	451	657
无形资产	183	199	325	574
商誉	0	0	0	0
递延所得税资产	0	0	0	0
其他非流动资产	257	293	311	355
资产总计	**6,718**	**9,027**	**14,210**	**22,747**
短期借款	100	300	600	1,000
应付票据及应付账款	515	740	1,331	2,063
预收账款	0	0	0	0
应付职工薪酬	156	333	539	743
应交税费	27	169	323	502
其他流动负债	20	48	80	120
流动负债合计	818	1,590	2,874	4,428
长期借款	0	0	0	0
应付债券	0	0	0	0
递延所得税负债	0	0	0	0
其他非流动负债	469	395	395	395
负债合计	**1,287**	**1,985**	**3,269**	**4,824**
归属母公司所有者权益	5,423	7,018	10,878	17,791
少数股东权益	8	24	63	133
股东权益	**5,430**	**7,042**	**10,941**	**17,923**
负债及股东权益	**6,718**	**9,027**	**14,210**	**22,747**

主要财务比率

	2023A	2024E	2025E	2026E
EPS	-1.08	3.82	9.25	16.56
BVPS	12.99	16.81	26.06	42.62
PE	—	153.21	63.32	35.36
PEG	—	0.34	0.45	0.45
PB	45.07	34.83	22.47	13.74
EV/EBITDA	-1,009.96	123.89	57.38	33.00
ROE	-8%	23%	35%	39%
ROIC	-9%	22%	34%	37%

现金流量表

单位：百万元	2023A	2024E	2025E	2026E
经营活动现金流净额	**-1,618**	**-391**	**685**	**3,195**
投资	-90	289	189	89
资本性支出	-366	-408	-488	-607
其他	44	-28	50	80
投资活动现金流净额	**-412**	**-147**	**-250**	**-438**
债权融资	100	194	300	400
股权融资	56	0	0	0
支付股利及利息	0	-6	-14	-24
其他	-108	-72	0	0
筹资活动现金流净额	**48**	**116**	**286**	**376**
现金净流量	**-1,982**	**-423**	**722**	**3,133**

资料来源：携宁，东海证券研究所，截至 2025 年 6 月 27 日

一、评级说明

	评级	说明
市场指数评级	看多	未来 6 个月内上证综指上升幅度达到或超过 20%
	看平	未来 6 个月内上证综指波动幅度在-20%—20%之间
	看空	未来 6 个月内上证综指下跌幅度达到或超过 20%
行业指数评级	超配	未来 6 个月内行业指数相对强于上证指数达到或超过 10%
	标配	未来 6 个月内行业指数相对上证指数在-10%—10%之间
	低配	未来 6 个月内行业指数相对弱于上证指数达到或超过 10%
公司股票评级	买入	未来 6 个月内股价相对强于上证指数达到或超过 15%
	增持	未来 6 个月内股价相对强于上证指数在 5%—15%之间
	中性	未来 6 个月内股价相对上证指数在-5%—5%之间
	减持	未来 6 个月内股价相对弱于上证指数 5%—15%之间
	卖出	未来 6 个月内股价相对弱于上证指数达到或超过 15%

二、分析师声明：

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，具备专业胜任能力，保证以专业严谨的研究方法和分析逻辑，采用合法合规的数据信息，审慎提出研究结论，独立、客观地出具本报告。

本报告中准确反映了署名分析师的个人研究观点和结论，不受任何第三方的授意或影响，其薪酬的任何组成部分无论是在过去、现在及将来，均与其在本报告中所表述的具体建议或观点无任何直接或间接的关系。

署名分析师本人及直系亲属与本报告中涉及的内容不存在任何利益关系。

三、免责声明：

本报告基于本公司研究所及研究人员认为合法合规的公开资料或实地调研的资料，但对这些信息的真实性、准确性和完整性不做任何保证。本报告仅反映研究人员个人出具本报告当时的分析和判断，并不代表东海证券股份有限公司，或任何其附属或联营公司的立场，本公司可能发表其他与本报告所载资料不一致及有不同结论的报告。本报告可能因时间等因素的变化而变化从而导致与事实不完全一致，敬请关注本公司就同一主题所出具的相关后续研究报告及评论文章。在法律允许的情况下，本公司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。

本报告仅供“东海证券股份有限公司”客户、员工及经本公司许可的机构与个人阅读和参考。在任何情况下，本报告中的信息和意见均不构成对任何机构和个人的投资建议，任何形式的保证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，本公司亦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。本公司客户如有任何疑问应当咨询独立财务顾问并独自进行投资判断。

四、资质声明：

东海证券股份有限公司是经中国证监会核准的合法证券经营机构，已经具备证券投资咨询业务资格。我们欢迎社会监督并提醒广大投资者，参与证券相关活动应当审慎选择具有相当资质的证券经营机构，注意防范非法证券活动。

上海 东海证券研究所
地址：上海市浦东新区东方路1928号 东海证券大厦
网址：Http://www.longone.com.cn
电话：（8621）20333619
传真：（8621）50585608
邮编：200215

北京 东海证券研究所
地址：北京市西三环北路87号国际财经中心D座15F
网址：Http://www.longone.com.cn
电话：（8610）59707105
传真：（8610）59707100
邮编：100089