寒武纪-U（688256）

首发报告

买入

2025 年 6 月 3 日

国产算力芯片中军，迎接智算时代历史机遇

- **国产芯片中军，技术立足，生态完整：**寒武纪是国内智算芯片领域的标杆企业，能提供完善的云、边、端系列化智能芯片和基础开发软件套件产品，具有完善统一的产品生态。公司与服务器厂商合作，将自研算力芯片应用于互联网、金融、交通、能源、电力和制造等领域。公司主要云端算力芯片思元 590 综合性能接近英伟达 A100，处于国内领先水平。下一代产品思元 690 预计性能翻倍，竞争力将更加突出。
- **人工智能高歌猛进，应用落地推升算力需求：**2022 年底，ChatGPT 横空出世，世界进入智算时代，全球算力芯片需求井喷。2025 开年，DeepSeek 发布 R1 推理模型，以其极低的训练成本和超强的推理能力震惊全球，带动大模型在各行各业广泛落地。国产芯片在推理应用阶段性价比和可用性大增，互联网、金融、运营商等行业客户国产芯片验证积极推进，订单持续落地。
- **中美竞争加速国产替代，成就千亿级市场：**大国科技竞争的宏大叙事下，算力国产化替代的紧迫性和重要性持续提升。预计 2024 年英伟达算卡在中国的市场规模超过 100 亿美元，占据 70-80%的市场份额。预计 2025 年国内算力需求高速增长，市场容量有望翻番。在美国限制芯片出口的背景下，国产算力芯片有望占据更多份额，成就千亿级市场。寒武纪凭借技术能力和先发优势，有望在新一轮国产算力投资周期中获得显著市场份额。
- **目标价 987 元，首次评级买入：**预计公司 2025-2027 年收入 84.2 亿/134.8 亿/216.0 亿元人民币，归母净利润 27.7 亿/48.1 亿/82.3 亿元人民币。首次评级买入，给与目标价 987 元，对应市值 4,114 亿元，对应 2027 年 50 倍 PE，较现价有 63.5%的上升空间，首次评级买入。
- **风险提示：**AI 芯片需求放缓，供应链波动，行业竞争加剧

黄晨

+852-2532 1954

chen.huang@firstshanghai.com.hk

主要资料

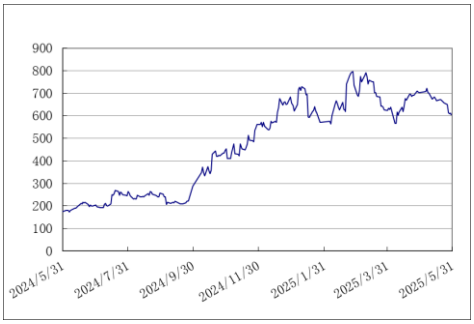
行业	半导体
股价	603.66 元人民币
目标价	987 元人民币
	+63.5%
股票代码	688256
已发行股本	4.17 亿股
总市值	2,520 亿元人民币
52 周高/低	818.87/168.64 元人民币
每股净资产	13.96 元人民币
主要股东	陈天石 28.63% 北京艾溪 7.34%

表：盈利摘要

12月31日	2023实际	2024实际	2025预测	2026预测	2027预测
营业额（百万人民币）	709	1,174	8,416	13,483	21,603
变动（%）	*(2.7)*	*65.6*	*616.6*	*60.2*	*60.2*
净利润（百万人民币）	(848)	(452)	2,766	4,808	8,228
每股盈余（人民币）	(2.0)	(1.1)	6.6	11.5	19.7
变动（%）	*167.5*	*46.7*	*711.5*	*73.8*	*71.1*
市盈率（倍）	NA	NA	91.1	52.4	30.6
市销率（倍）	355.2	214.6	29.9	18.7	11.7

资料来源：公司资料，第一上海预测

股价表现



资料来源：彭博

国产算力芯片中军，专注算力芯片

技术立足，深耕智算领域

国产算力龙头，产品对标英伟达主流产品

寒武纪是国内智算芯片领域的龙头企业，能提供完善的云边端系列化智能芯片和基础开发软件套件产品，具有云边端一体、软硬件协同、训练推理融合、具备统一生态的特点。公司与服务器厂商和产业公司合作，为互联网、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。公司的产品包括云端智能芯片、加速卡及训练整机，边缘智能芯片及加速卡，终端智能处理器 IP 等。公司核心产品为思元系列云端算力芯片，对标英伟达主流产品，为客户提供国产化芯片新选择。

董事长兼总经理陈天石从研究生开始便投入芯片研发工作，是全世界最早探索专用 AI 芯片的少数科研人员之一。自成立以来，公司一直深耕于专用 AI 芯片领域，在发展中进行了大量处理器指令集与微架构的创新，并对开发者套件和开发者生态进行了不断迭代完善，推动公司成长为国产智算芯片的核心供应商。

图表 1：公司发展历程

2014	董事长陈天石与其他研发人员合作论文获得国际顶级学术会议最佳论文奖，提出了一种专用 AI 芯片架构
2015	全球第一颗专用 AI 芯片流片成功，并被取名为“寒武纪”
2016	**寒武纪公司成立**
2017	终端智算芯片发布，其中 1A 型号被用于某国产手机品牌作为其 NPU 处理器
2018	**云端智算芯片思元 100 发布，是公司首颗高峰值算力计算芯片**
2019	第二代云端智算芯片思元 270 发布，同时推出边缘智能芯片思元 220，完成了产品云边端的布局
2020	**公司在上海科创板上市，代码 688256**
2021	**第三代云端智算芯片思元 370 发布，同时推出了基于思元 370 的加速卡产品。首次发布多芯互联技术 MLU-Link。**
2022	基于思元 370 推出了新款加速卡，以及训练整机产品玄思 1000 智能加速器
2023	**第四代云端智算芯片思元 590 发布，算力参数对标英伟达 A100**
2024	思元 590 通过互联网客户测试，开始正式出货

资料来源：公司资料，第一上海

核心团队科班出身，年富力强

核心团队专业知识扎实，产业经验丰富。多次股权激励计划提升团队稳定性和凝聚力

公司创始团队来自中科院计算所，核心研发人员多来自于著名高校或者科研院所。董事长兼总经理陈天石毕业于中科院计算所，是全球最早提出专用 AI 芯片概念的科研人员之一，其论文相继获得国际顶级学术会议 The ACM International Conference on Architectural Support for Programming Languages and Operating Systems（ASPLOS）以及 IEEE/ACM International Symposium on Microarchitecture 最佳论文奖，被《科学》杂志评价为深度学习处理器的“先驱”和“引领者”，相关研究成果被 Nvidia 以及 Google TPU 研发团队多次引用。副总经理、核心技术人员刘少礼、陈帅、刘毅、张尧等均科班出身，拥有扎实的专业知识积累和丰富的产业从业经验。

公司在 2020 年，2021 年，2023 年进行过三次限制性股票激励计划，绑定核心管理层和研发人员，提升了团队的稳定性和凝聚力，保证了公司的持续创新和技术领先，让公司在高科技领域能站稳脚跟，不断发展壮大。

图表 2：管理团队简介

姓名	职务	简介
陈天石	董事长兼总经理	生于 1985 年，中国科学技术大学计算机软件与理论专业博士学历。中国国籍。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，现任公司董事长、总经理、核心技术人员
王在	副总经理	生于 1984 年，中国科学技术大学计算机应用技术博士学历。中国国籍。2011 年至 2015 年就职于郑州商品交易所并任核心交易系统工程师，2015 年至 2016 年就职于中原银行并任信息科技部电子银行系统主管，2016 年至 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，现任公司董事、副总经理
刘少礼	副总经理、核心技术人员	生于 1987 年，中科院计算所计算机系统结构博士学历。中国国籍。2014 年至 2019 年就职于中科院计算所并任副研究员。2016 年作为公司创始团队成员加入公司，现任公司董事、副总经理、核心技术人员。
陈帅	副总经理、核心技术人员	生于 1986 年，中国科学院计算技术研究所计算机系统结构博士学历。中国国籍。2014 年至 2015 年，任中国科学院计算技术研究所工程师。2015 年至 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，现任公司副总经理、核心技术人员
刘毅	副总经理、核心技术人员	生于 1985 年，北京大学微电子与固体电子学硕士学历。中国国籍。2010 年至 2012 年，就职于龙芯中科技术股份有限公司，任工程师。2012 年至 2016 年，就职于上海英伟达半导体(科技)有限公司任高级工程师。2016 年至今，就职于中科寒武纪科技股份有限公司，现任公司副总经理、核心技术人员
张尧	副总经理、核心技术人员	生于 1986 年，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。中国国籍。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，现任公司副总经理、核心技术人员

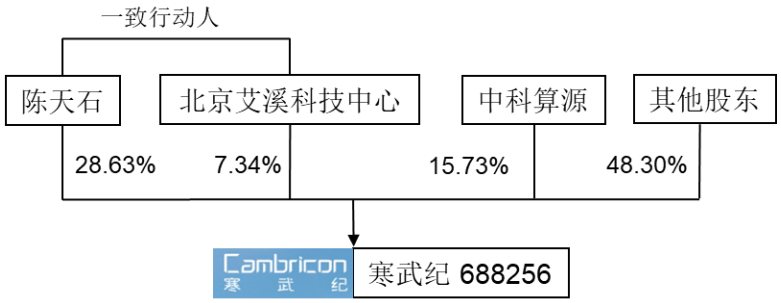
资料来源：公开信息，第一上海

股权结构

第二大股东中科算源为公司提供良好的股东背景，以及人才和产业资源

截至 2025 年一季报，公司总股本为 4.17 亿股，为全流通。公司董事长兼总经理陈天石先生直接持有公司 28.63%，通过北京艾溪间接持有公司 7.34%，合计持有 35.97%股权，是公司的实际控制人。第二大股东中科算源为中科院计算所旗下企业，持有公司 15.73%股权，其为公司发展提供了良好的股东背景，以及人才和产业资源，通过产学研协同使公司在 AI 芯片发展中充分受益。其余前十股东多为指数基金。

图表 3：公司股权结构



资料来源：公司资料，第一上海整理

云边端一体+软硬件协同构筑完整生态，

云端优先发力，冲破算力垄断

云边端一体化，战略聚焦云端算力。最新产品思元 590 取得商业化突破，取得互联网客户大批量订单。

公司围绕云、边、端三种场景，打造芯片产品矩阵。2016 年，公司首款商用端侧 AI 芯片——寒武纪 1A 发布，被集成在全球首款人工智能手机芯片华为麒麟 970 上，使华为旗舰手机 Mate 10 具备强大的本地智能处理能力。此后，基于寒武纪 1A，公司进一步推出了多款端侧 AI 产品，包括 1M，1H 等型号，用于图像识别，安防监控、智能驾驶、无人机、语音识别、自然语言处理等多个应用领域。

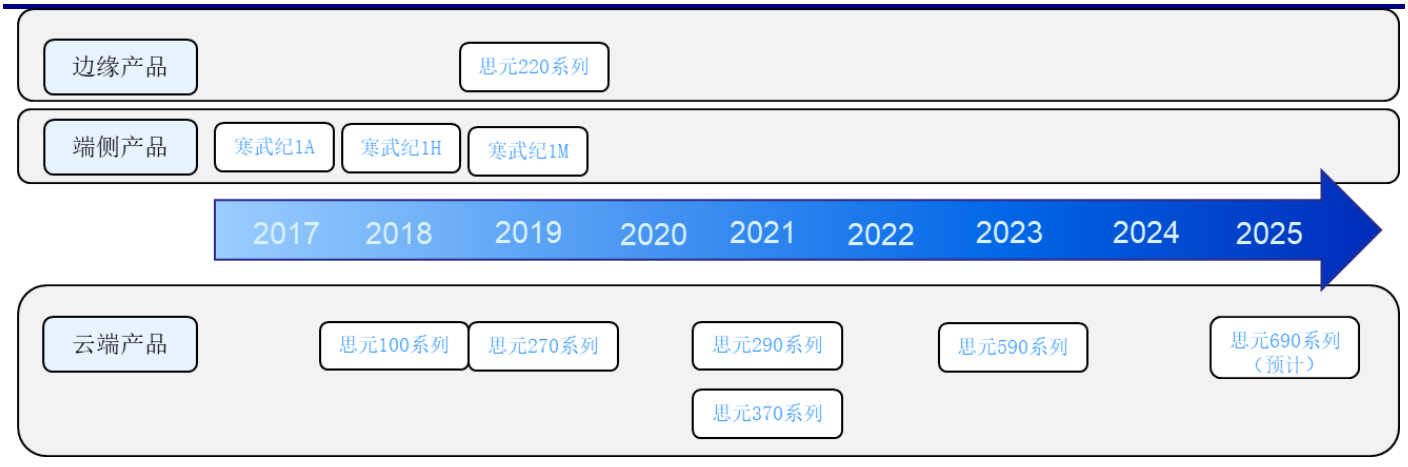
2017 年，公司发布了面向云端的高性能 AI 处理器产品线，以机器学习处理器（MLU）命名，确立了高算力芯片作为公司产品矩阵拼图的重要一环。2018 年，中国首颗高峰值算力云端 AI 处理器芯片思元 100 诞生。2019 年，第二代云端高算力产品——MLU200 系列芯片发布，包括思元 290、思元 270、思元 220 等型号，其中思元 290 主要面向云端高算力需求训练场景，思元 270 主要面向云端推理场景，思元 220 主要面向边缘计算场景。公司在产品端完善了公司云、边、端的布局。

2021 年，公司发布第三代云端高算力芯片——MLU300 系列芯片，主要产品为思元 370 芯片及相对应的系列加速卡和整机。公司同时推出了自有**卡间互联方案 MLU-Link**，对标英伟达 NVLink，实现高效的芯片间互联，充分发挥芯片的训练和推理效率。

2023 年，公司发布最新一代云端高算力芯片产品——思元 590 芯片。该芯片方便兼容主流 AI 大模型，综合性能对标英伟达 A100，实力处于国内领先地位。思元 590 的发布奠定了公司在国产高算力芯片领域的龙头地位。

全球进入智算时代，算力芯片需求出现井喷。公司将业务重心聚焦到云端算力芯片上，以更好满足下游客户对高性能国产芯片的需求。2023 年以来，公司与互联网、电信、金融、交通等领域客户密切配合，不断打磨产品，在近期取得阶段性成绩，有望获取互联网大客户的大批量订单。

图表 4：公司产品线梳理



资料来源：公司官网，第一上海

旗舰产品性能优异，处于国产领先水平

2021 年，公司发布思元 290 智能芯片及加速卡 MLU290-M5，该产品是公司首颗训练用芯片，采用台积电 7nm 先进制程工艺，集成寒武纪自研的 MLU-Link 多芯互联技术，可以高效执行多芯多卡训练和分布式推理任务。2022 年，公司发布基于思元 370 芯片的新款训练加速卡——MLU370-X8，该产品为双思元 370 芯片配置，集成 MLU-Link

多芯互联技术，主要面向训练任务。2023 年，公司发布最新一代芯片思元 590，性能相比思元 370 有翻倍以上的提升，综合性能对标英伟达 A100，处于国内领先水平。

图表 5：国内外算力芯片指标对比

公司	成立时间	产品系列	型号	场景	进展情况	显存容量 GB	算力（Tops） INT8	算力（Tops） FP16	显存带宽 GB/s	互联带宽 GB/s	制程
英伟达	1993	A100	PCIe	训推一体	已量产	40	624	312	2048	Nvlink:600,PCIe:64	7nm
		H100	PCIe	训推一体	已量产	80	3026	1513	2048	Nvlink:600,PCIe:128	4nm
华为	2004	昇腾910B	——	训推一体	已量产	64	640	320	——	400	——
		昇腾910C	——	训推一体	待量产	——	——	——	——	——	——
寒武纪	2016	思元370	MLU370 X8	推理为主	已量产	48	256	96	614.4	200	7nm
		思元590	——	训推一体	已量产	——	600	——	——	——	——
		思元690	——	训推一体	研发中	——	——	——	——	——	——
海光信息	2014	深算DCU	1号	训练	已量产	32	——	24.5	1024	184	7nm
			2号	训练	已量产	——	——	——	——	——	——
			3号	训练	测试中	——	——	——	——	——	——
昆仑芯（百度）	2011	昆仑芯	二代	训推一体	已量产	32	256	128	512	200	7nm
			三代	训推一体	待量产	——	——	——	——	——	7nm
沐曦	2020	曦云	C500	训推一体	已量产	64	480	240	1800	——	7nm
		曦思	N100	推理	已量产	——	160	——	——	——	——

资料来源：公开资料，第一上海

重视软件平台打造，软硬协同发挥芯片效能

Cambricon NeuWare 是公司的生态核心，帮助公司打造软硬件协同生态。

硬件端：

公司掌握智能处理器指令集、智能处理器微架构的设计和开发，目前已经迭代到第五代处理器微架构 MLUarch05，用于最新一代云端训练芯片思元 590。公司仍在迭代开发新一代智能处理器微架构和指令集，针对**语言大模型、图像视频大模型、推荐系统大模型等训练推理场景进行重点优化**，提升产品在性能、功耗上的技术指标，同时也将显著提升编程灵活性和产品易用性，以提升产品竞争力。

软件端：

公司也在智能芯片编程语言、智能芯片数学库有核心技术积累，推出 **Cambricon NeuWare** 作为云、边、端统一的智能处理器软件开发平台。Cambricon NeuWare 整合了训练和推理的全部底层软件栈，包括硬件驱动、AI 加速算子库（CNNL），通信库（CNCL），开发语言 BANG 等，同时将该软件平台与 Tensorflow、Pytorch 等 AI 框架深度融合，实现训推一体，让开发者可以非常方便地完成从云到端，从模型训练到推理部署的全部流程，提升 AI 算法的开发效率。

训练端，该平台支持丰富的图形图像、语音、推荐以及 NLP 训练任务，通过自有的底层算子库 CNNL 和通信库 CNCL，在实际训练任务中达到业界领先的硬件计算效率和通信效率。推理端，公司打造了 **MagicMind 推理加速引擎**，使得用户仅需投入极少的开发成本，就可以将推理业务部署到寒武纪全系列产品上，获得颇具竞争力的产品性能。

效仿 CUDA 生态，软硬一体构筑壁垒

CUDA 是英伟达在 2007 年推出的并行计算框架和编程模型，也是英伟达构筑起的坚实软硬件协同壁垒。英伟达通过 CUDA，给开发者提供了一整套开发工具，各种库（加速库，数学库，通信库等），以及各种调优方式，让开发者可以快速上手，完成项目开发，同时还方便项目在各种平台之间做迁移。公司的 Cambricon NeuWare 整合了训练和推理的全部底层软件栈、各种库和编程语言，让开发者可以快速实现项目开发，高效利用芯片算力，有效提升了公司产品的竞争力。

图表 6：寒武纪的“CUDA”——Cambricon NeuWare

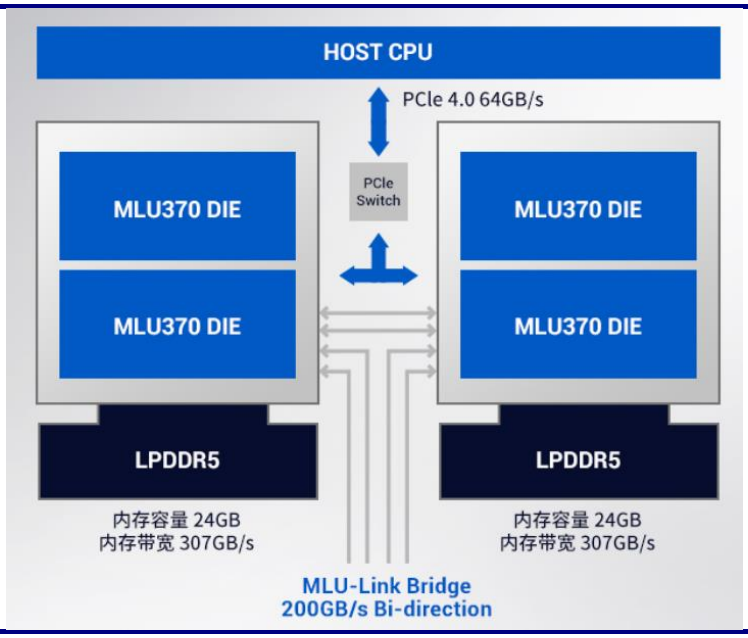
资料来源：公司官网，第一上海

多芯互联提升芯片组网效率

MLU-Link 提升卡间互联能力，打破集群性能瓶颈

2021 年，公司在发布思元 290 智能芯片时，首次推出自研的 MLU-Link 多芯互联技术，类似英伟达的 NVLink，帮助算力集群执行高效的多芯多卡训练和分布式推理任务。2022 年，公司发布 MLU370 X8 算卡，搭载了 MLU-Link，为每颗芯片提供 200GB/s 的额外跨芯片通讯能力，带宽是 PCIe 4.0 标准的 3 倍。公司为多卡系统专门设计了 MLU-Link 桥接器，可实现 4 张双芯 MLU370 X8 算卡的互联。

图表 7：寒武纪的“NVLink”——MLU-Link



资料来源：公司官网，第一上海

AI 应用遍地生根，国产算力迎历史机遇

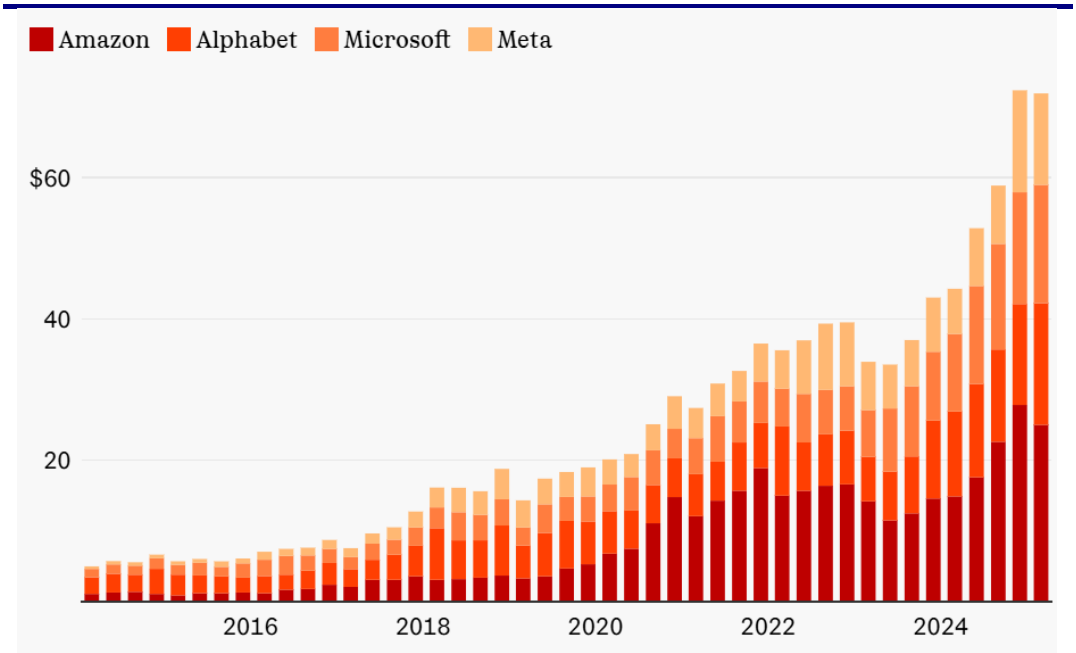
AI 技术新浪潮，全球算力需求高增长

2022 年底，OpenAI 发布 ChatGPT，世界迎来了生成式人工智能时代。科技巨头纷纷入局，参与大模型竞赛。OpenAI 对模型快速迭代，陆续推出 GPT4，GPT4o 等迭代大模型、Sora 视频生成大模型，以及 o3、o1 等推理大模型。Google 也不甘人后，推出了 Gemini，在多模态能力上表现出众。亚马逊则通过投资 Anthropic 入局，Anthropic 的 Claude 系列模型在 Agent 能力上领先全球。国内企业也纷纷入局，出现百模大战，华为，百度，科大讯飞，百川智能，Minimax，智谱，以及之后的阿里，字节，腾讯等也陆续发布了自己的大模型。

生成式人工智能需要巨量算力去做模型训练，全球对算力芯片的需求出现井喷。北美云厂商逐季上调资本开支，2024 年总资本开支达到 3000 多亿美元，其中约一半以上用于 AI 算力投资。2025 年北美云厂商算力开支进一步增长，其中谷歌预计 25 年资本开支 750 亿美元；Meta 预计 640-720 亿美元；亚马逊 25Q1 单季度 250 亿美元，兵预计之后季度仍将增加，全年预计超过 1000 亿美元；微软预计 FY26 资本开支 800 亿美元。

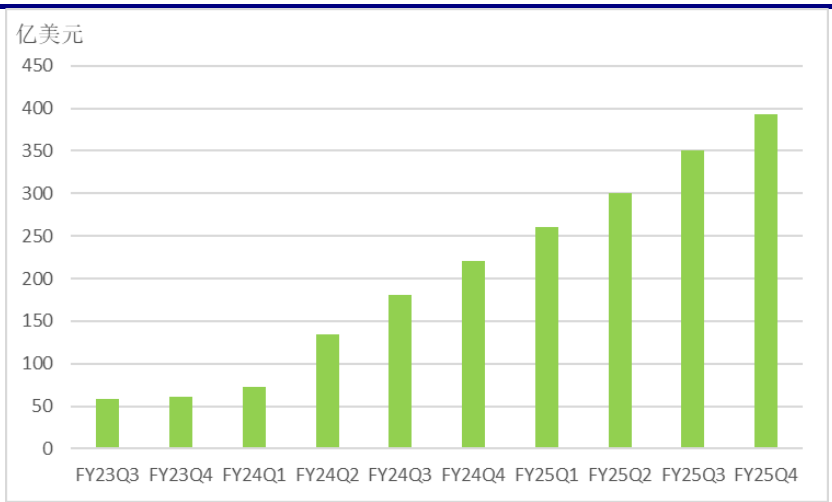
英伟达作为全球 AI 算力龙头，业绩逐季提升，单季度收入从 22Q3 的 59.3 亿美元，提升至 FY25Q4（日历年 25 年 1 月）的 393 亿美元。

图表 8：北美云厂商资本开支（14Q1-25Q1）



资料来源：Factset，Sherwood，第一上海

图表 9：英伟达收入逐季高增（FY23Q3-FY25Q4）



资料来源：彭博，第一上海

注：期间为日历年 2022 年 8 月至 2025 年 1 月

中国智算投资加速，AI 服务器市场将突破千亿美元

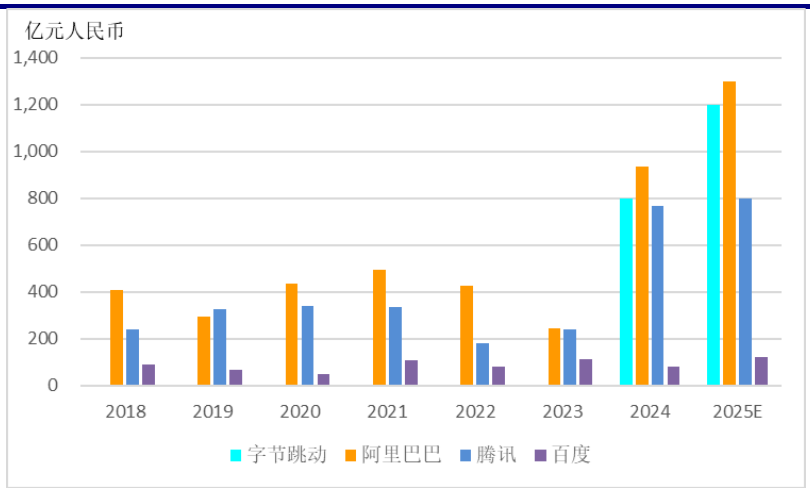
预计字节、阿里、百度、腾讯 2025 年资本开支突破 3,000 亿元，叠加运营商和行业需求，算力投资有望超过 5,000 亿元。

国内厂商以字节为代表，年度资本开支在 2024 年达到 800 亿元人民币，预计 2025 年进一步提升至超过 1,200 亿元，其中绝大部分将用于 AI 算力投资。阿里、腾讯、百度（BAT）三家的资本开支在 2024 年合计在 1,700 亿元左右，预计 2025 年三家资本开支将大幅增加，主要受 AI 投资的带动。

2024 年，英伟达在中国 AI 算卡市场出货量占比超过 70%。受美国出口管制影响，其市场份额将进一步缩小

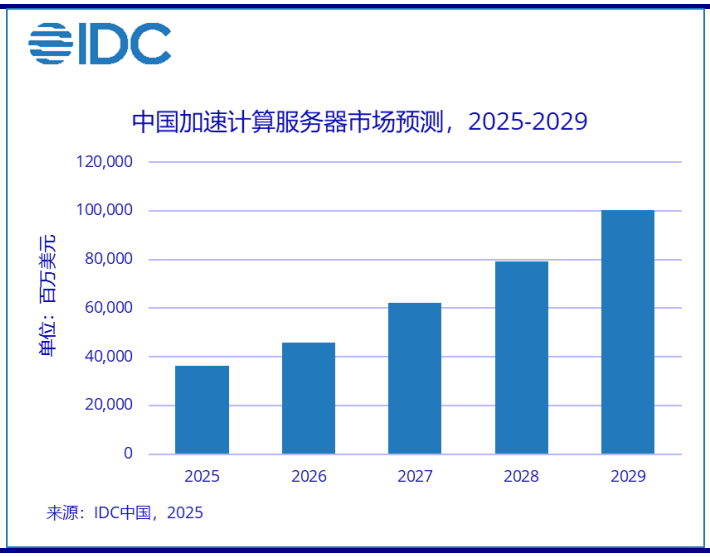
2025 年 3 月 31 日，IDC 发布的《中国半年度加速计算市场（2024 下半年）跟踪》报告，数据显示，2024 年中国加速服务器市场规模达到 221 亿美元，同比 2023 年增长 134%。IDC 预期 2029 年中国加速服务器市场规模将达到 1000 亿美元。2024 年，中国加速芯片的市场规模增长迅速，超过需求超过 270 万张，其中 GPU 卡占据 70%的市场份额，主要为 Nvidia 算卡。从品牌角度来看，中国本土人工智能芯片品牌的出货量已超过 82 万张。通过适配 DeepSeek，中国本土芯片在软件生态领域实现了突破，逐步完善软件生态。

图表 10：阿里、腾讯、字节资本开支（2028-2025E）



资料来源：公开资料，第一上海

图表 11：IDC 预测 2029 年中国加速服务器市场突破 1,000 亿美元



资料来源：IDC，第一上海

高性价比国产模型推动应用落地

AI 应用大批量落地中，推理需求将接力训练需求，带动算力需求增长

以 DeepSeek V3 为代表的“降本大模型”和 DeepSeek R1 为代表的推理模型，加速了 AI 应用的落地，带动推理需求的快速提升，推理算力需求接替训练算力成为算力需求主要来源。

2025 年 4 月底，阿里巴巴正式发布新一代 Qwen-3 系列大模型，涵盖 6 个稠密模型和 2 个 MoE 模型，参数量从 6 亿到 2,350 亿，满足从边缘端到服务器端不同场景下的推理需求。该系列模型在性能和推理效率上均进行了优化，使得其相较前代模型在准确率上提升 10-30%，推理速度上提高 20-40%，进一步降低了推理成本，增强了模型在端侧部署的实用性。据报道，2,350 亿参数版本 Qwen-3 的硬件投资成本相比满血版 DeepSeek R1 要降低 65-70%，日常推理场景成本是 GTP-4 Turbo 的 1/20。

业界预期 DeepSeek R2 有望在 2025 年内亮相，预计仍将采用 MoE 模式，但是训练参数量将达到 1.2 万亿，相比 R1 有接近翻倍的提升，而成本将进一步大幅下降。据传该大模型将由全国产算力训练完成，不依赖英伟达芯片。

阿里巴巴的 2025 年一季报业绩会中提到：公司看到两个趋势，一是大中型企业中 AI 应用从内部使用向用户场景渗透，二是积极使用 AI 的客户从大中型企业往中小企业延展。公司看到 AI 在互联网、智能汽车、金融、在线教育等新兴行业，以及传统制造业、养殖业等传统行业加速落地。

中美争夺 AI 高地话语权

中美竞争白热化，抢夺未来 AI 行业话语权

自 2018 年中美贸易摩擦以来，美国就开始对中国高科技，尤其是半导体产业，进行层层加码的限制。2022 年生成式人工智能横空出世，中美竞争进入 AI 制高点的争夺。美国挟其在半导体先进制程的优势不断对中国 AI 产业卡脖子。

在 2022 年，2023 年美国政府层层收紧对中国高端算力芯片供应之后，拜登政府在其执政最后阶段，连续出台了多项行政命令，进一步对中国高端芯片的获取能力进行限制。2025 年 1 月 13 日，拜登政府发布了《关于人工智能扩散的临时最终规则》，对芯片出口到其核心“朋友圈”以外的国家进一步收紧。1 月 15 日，拜登政府再签

署一道行政命令，收紧台积电、三星等晶圆代工厂所生产的 14nm 或 16nm 及以下先进制程的芯片流入中国大陆，同时将加强最终客户的审核。

2025 年 5 月，美国在 AI 先进算力端出现两个重要事件，1）美国商务部正式发文宣布废除拜登政府的人工智能扩散规则，同时宣布三项指引以加强对海外 AI 芯片的出口管制。2）英伟达 CEO 黄仁勋宣布，将向沙特企业 Humain 出售超过 1.8 万颗最新人工智能芯片。我们认为，美国在未来 AI 竞争中的立场和方向开始明确，即鼓励全球采用美国主导的 AI 芯片和软件，同时对中国等竞争国家采取压制措施，限制中国为主导的芯片和模型在全球的应用和发展。

我们认为 AI 芯片、AI 大模型将成为未来世界的基础设施，类似移动通信（4G、5G）、互联网等技术在全球科技产业中的地位。因此，掌握 AI 芯片和大模型的话语权将是大国之间科技竞争的核心战场。美国开启算力外交，期望复制其过去在全球供应链中的核心地位，抢抓全球 AI 发展主动权。

公司高算力芯片获得互联网及行业客户认可

公司产品获互联网客户测试通过，有望加速出货。公司也与行业客户深化合作，帮助行业大模型的落地。

自 2021 年思元 290 和思元 370 发布以来，公司高算力产品已经在包括阿里，百度，字节，腾讯等互联网大厂受到广泛测试和量产落地。同时，公司的产品也在运营商、金融、交通、能源等行业取得项目落地进展。

- ✓ **互联网领域**：在字节、阿里、腾讯、百度等互联网大厂测试，有望承接推理应用算力需求。
- ✓ **金融领域**：公司持续加深与银行、保险公司及基金公司的业务探索。除了为传统人工智能应用场景持续提供算力支持外，公司全面开展大模型的适配优化工作，帮助客户实现大模型在实际业务场景中的落地应用。
- ✓ **交通领域**：公司成功参与多地车路云一体化项目、智慧停车、智慧高速业务，助力交通数字信息化发展。
- ✓ **轨道行业**：公司在智慧货检、语音购票等方面与关键客户展开深入合作，推进铁路服务智能化升级。
- ✓ **其他垂直行业**：公司的智能芯片产品继续为传统产业智能化转型保驾护航，助力智慧矿山、智慧粮仓业务的落地。

公司新一代算力卡思元 590 对标英伟达 A100 芯片，作为性能更强，适配更优，开发者工具更完善的产品，有望在 25 年迎来加速出货。更新一代算力新品也在研发和测试中，据传综合性能对标英伟达 H100，有望引领国产算力替代。

先发优势推动公司业务实现正循环

Cambricon NeuWare——寒武纪的 CUDA

领先行业进行开发者生态迭代

CUDA 生态是保障英伟达业务持续高速发展的重要一环。英伟达于 2007 年正式发布 CUDA 计算架构，经过近 20 年的发展，CUDA 生态已经成为英伟达宽厚的护城河。任何芯片厂商想要冲击英伟达的芯片市场领导地位，都在 CUDA 构筑的完善生态体系下败下阵来。CUDA 不仅是一个工具，它是英伟达整个业务体系的神经中枢。一旦开发者在 CUDA 上投入，他们几乎无法轻易转向其他平台。英伟达通过不断优化 CUDA 与其硬件的配合，不仅可以发挥 GPU 最大性能，而且也加深了用户对英伟达生态的依赖。开发者可以利用 CUDA 高效地开发项目、迁移项目、优化项目，还能得到广泛的社区支持和培训体系。

Cambricon NeuWare——寒武纪的“CUDA”

Cambricon NeuWare 是公司智能处理器产品的软件开发平台，采用云边端一体、训推一体架构，可同时支持寒武纪云、边、端的全系列产品。寒武纪终端 IP、边缘端芯片、云端芯片共享同样的软件接口和完备生态，可以方便地进行智能应用的开发，迁移和调优。

训练端：平台支持丰富的**图形图像、语音、推荐以及 NLP 训练任务**。通过底层**算子库 CNNL 和通信库 CNCL，在实际训练业务中达到业界领先的硬件计算效率和通信效率**。同时提供**模型快速迁移方法**，帮助用户快速完成现有业务模型的迁移。

推理端：MagicMind 是寒武纪全新打造的**推理加速引擎**，也是业界首个基于 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寒武纪全系列产品上，并获得颇具竞争力的性能。

用户支持：公司提供成熟完善的开发者社区、开发者论坛、开发文档，以及参考案例，让用户高效快速的完成项目开发。

先发优势助力公司生态建设，大规模交付有望加速公司实现正循环

公司是国内最早一批提供高算力芯片的公司之一，率先完成了开发生态的搭建。前几年，公司产品在政府、垂直行业、运营商、互联网公司的应用过程中，领先国内同行完善了开发者生态和硬件产品迭代等工作。

我们认为，公司未来在互联网公司的大规模交付将进一步加速公司的开发者生态建设和产品迭代，帮助公司率先实现正循环。

财务亮点

存货预付增长明显：需求畅旺，供应保障

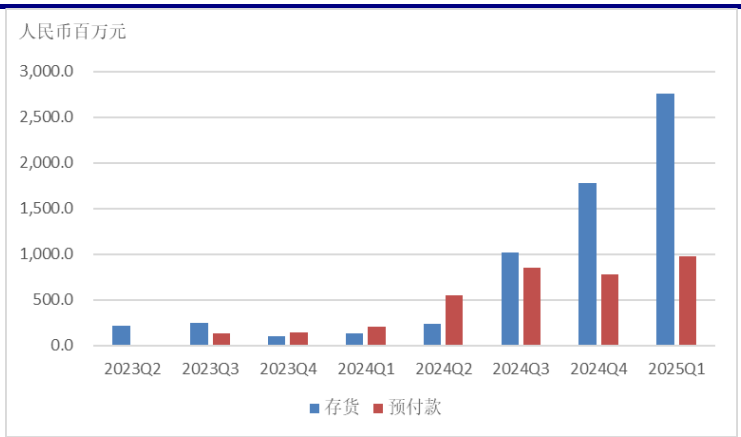
存货、预付款双双高增，印证需求畅旺，供应链有保障。

2025 年一季度，公司**存货 27.6 亿元**，相比 2024 年末的 17.7 亿元大幅增长 9.9 亿元，公司存货主要为委托加工物资（晶圆、零部件等）和库存商品，参考 2024 年年报，库存商品和委托加工物资合计占比 80.33%。公司存货的大幅增长为后续业绩增长提供有力支撑。

2025 年一季度，公司**预付款 9.7 亿元**，相比 2024 年末的 7.4 亿元增长 2.3 亿元，预付款主要为公司向上游锁定产能和重要零部件产品而提前支付的款项。我们认为预付款的持续增长，表明公司**供应链的稳健性**，为公司未来业绩持续增长提供保障。

按照公司 2024 年度 56.7%毛利率大致测算，公司 27.6 亿元的存货加上 9.7 亿元的预付款，可以支撑超过 70-80 亿元的收入。

图表 12：存货预付款双双高增



资料来源：公司财报，第一上海整理

高额研发投入打造产品竞争力

高额研发投入保障软硬件快速迭代，打造产品充足竞争力

近年来，公司持续投入高额研发开支，持续打造高性能的产品和易用的软件生态。2019 年以来，公司的研发费用率一直保持在 100%以上，2022/2023/2024 年分别达到收入的 209%/158%/104%。公司的研发开支强度以及绝对值，在国内上市芯片公司中均处于领先地位。

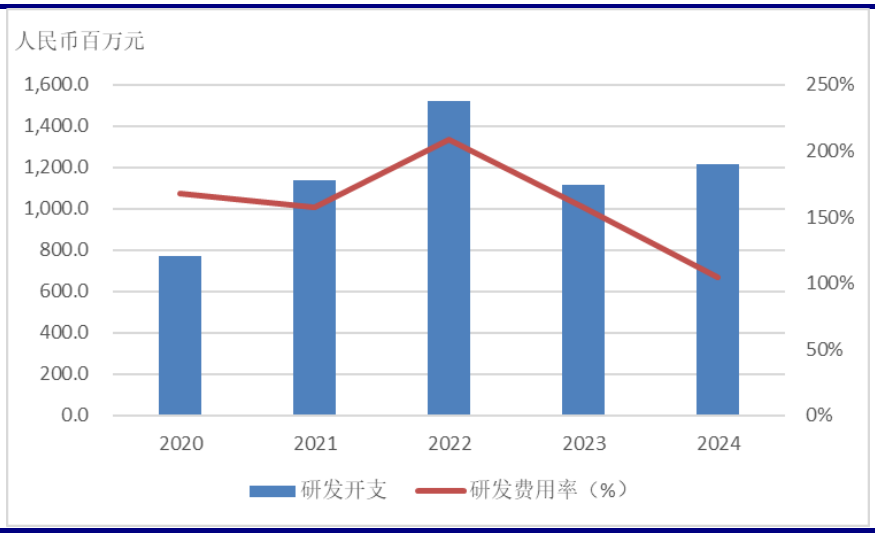
2024 年年报，公司提到研发的方向包括：

研发方向上，公司针对硬件进行场景适配，对软件进行易用性和稳定性提升，同时支持更多大模型。

硬件端：公司持续优化芯片架构设计，提升对**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的适配能力**，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。

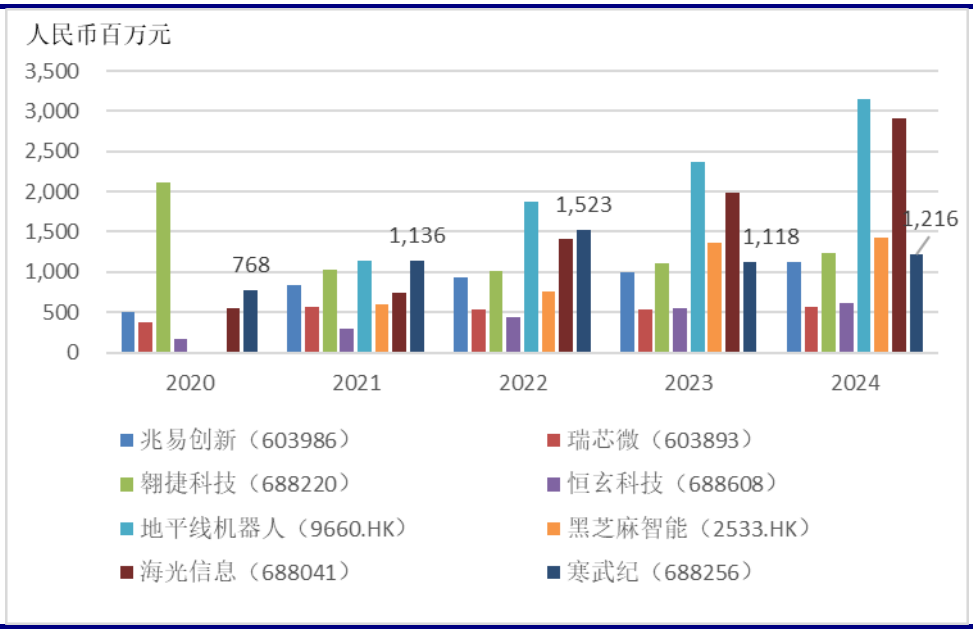
软件生态端：针对大模型训练和推理应用场景进行持续优化，完善基础系统软件平台功能，**提升公司软件平台的易用性和稳定性**，全面提升公司产品在训练和推理场景中的竞争优势。在生态方面，训练软件平台提供了 Pytorch2.1/2.3/2.4/2.5 的支持，实现了 Transformers、Accelerate、DeepSpeed 社区原生支持 MLU。在大模型方面，**训练软件平台增加了对 DeepSeek 系列、Llama 系列、Qwen 系列等主流模型训练的支持**。训练软件平台已支持并行训练功能，持续优化热点算子性能，通过优化融合算子、支持通算融合等优化策略，使得训练性能达到了业界主流水平，具备了更强的行业竞争力。

图表 13：研发开支和研发费用率持续高位



资料来源：公司财报，第一上海整理

图表 14：研发开支处于行业领先地位



资料来源：彭博，第一上海整理

定增增强现金储备，为公司高速发展保驾护航

如果定增项目顺利实施，将保障公司在国产芯片 10-100 放量阶段的产业地位，为公司高速发展保驾护航

2025 年 4 月 30 日，公司发布定增预案公告，计划向特定对象发行 A 股募集资金不超过 49.8 亿元人民币，扣除发行费用后，实际募集资金将用于大模型的芯片平台项目（拟投资 29 亿元人民币）、面向大模型的软件平台项目（拟投资 16 亿元人民币），以及补充流动资金（拟投资 4.8 亿元人民币）。

据公司公告，募投项目将：

1） **增强公司面向大模型的芯片技术和产品综合实力，提升公司在智能芯片产业领域的长期竞争力**。公司将开展面向大模型的智能处理器技术创新突破，研发覆盖不同类型大模型任务场景（训练用芯片、大语言推理用芯片、多模态推理用芯片、大模型通信用交换芯片）的系列化芯片产品，建设先进封装技术平台。

2） **构建面向大模型的软件平台，进一步提升公司软件生态的开放性和易用性**。基于公司智能芯片的硬件架构特点，拟研发面向大模型的软件平台，重点面向大模型技术开展相应的优化策略、软件算法以及软件工具的创新研究，构建面向大模型算法开发和应用部署的高效支撑与服务能力。

3） **满足公司营运资金需求，提升公司抗风险能力**。随着公司研发投入和业务规模的扩大，公司对营运资金的需求相应提高，因此需要有充足的流动资金支持公司经营，为公司进一步提升市场竞争力奠定良好基础。

我们认为，如果公司定增项目的顺利完成，将保障公司在国产算力芯片从 10 到 100 放量阶段把握主动权，使得公司不仅能投入更多资金做芯片研发迭代以更快速满足客户需求，而且能有更多资金保障供应链的安全性。

收入和盈利预测

收入预测

公司的收入主要来自进程电路产品的销售，按应用场景分，包括云端产品线，边缘产品线，IP 授权及软件业务，以及其他业务。

云端产品线：

云端产品线的主要收入来源就是向云计算厂商、数据中心、行业客户销售云端智能芯片及加速卡。该业务也是公司收入的主要来源，2024 年该业务占公司总体营收的 99.3%，预计未来该业务营收占比将继续维持高位。该板块主要产品包括思元 290、思元 370、以及思元 590 等系列芯片和加速卡产品。我们预期公司云端产品线将显著受益于互联网大厂、运营商、行业客户的 AI 算力建设，相关业务将在 2025 年开始放量，并在 2026、2027 年持续高速增长。

我们预计 2025-2027 年，公司云端产品线收入分别达到 84.1 亿元，134.7 亿元，215.9 亿元，同比增长达到 620.8%，60.3%，以及 60.3%，以反映公司云端算力产品受到客户认可，出货量从 2025 年开始爆发式增长。

边缘产品线：

边缘产品线的主要收入来源为销售边缘算力芯片及板卡，用于视觉、语音等边缘感知场景，包括电力、交通、金融、物流、医疗等行业。

IP 授权及软件业务：

该业务主要为向客户授权公司智能处理器 IP 在其产品中使用，以及向客户授权使用公司云边端一体的平台软件套件。

智能计算集群：

由于公司的业务模式以及与客户的合作方式变化，公司相关收入在 2024 年减少至零。我们预期该业务未来收入将主要体现在云端产品线收入中。

图表 15：公司收入预测（2024A-2027E）

收入（人民币百万）	2024A	2025E	2026E	2027E
云端产品线	1,166.3	8,407	13,473	21,593
增速（%）	*1187.8%*	*620.8%*	*60.3%*	*60.3%*
边缘产品线	6.5	7.2	7.9	8.7
增速（%）	*-39.6%*	*10.0%*	*10.0%*	*10.0%*
IP授权及软件	0.4	0.5	0.5	0.5
增速（%）	*76.1%*	*10.0%*	*10.0%*	*10.0%*
智能计算集群	0.0	0.0	0.0	0.0
增速（%）	*-100.0%*	——	——	——
其他业务	1.1	1.1	1.1	1.1
增速（%）	*-64.6%*	*0.0*	*0.0*	*0.0*
合计	1,174	8,416	13,483	21,603
增速（%）	*65.5%*	*616.6%*	*60.2%*	*60.2%*

资料来源：第一上海预测

盈利预测

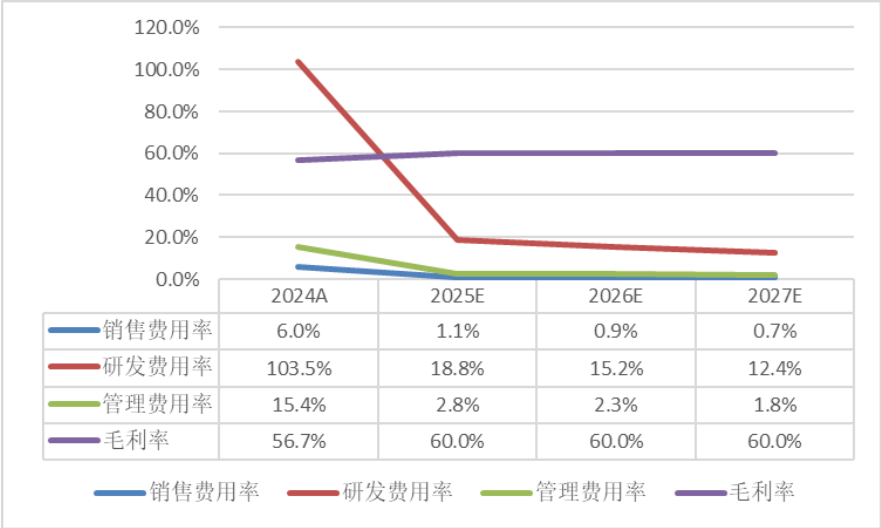
毛利率在业务放量后维持高位稳定

毛利率方面，我们预计公司云端产品线业务将在 2025 年开始放量，毛利率在 2025 年有所提升，并在之后几年维持稳定，预期未来三年毛利率水平为维持 60%。

三项费用率随着收入快速增长而降低。研发开支绝对额继续高速增长以维持公司的技术领先性

费用率方面，管理费用率和销售费用率将显著降低，主要得益于公司收入大幅上升带来的规模效应。管理费用和销售费用绝对值继续保持增长，以反映公司业务扩张带来的费用上升。研发费用率将显著降低，预期在公司业务放量后仍将维持 12-15%左右的研发费用率。研发费用绝对值预期仍将继续提升，以保障公司在所处领域的技术领先性。

图表 16：公司费用率显著降低，规模效应显现

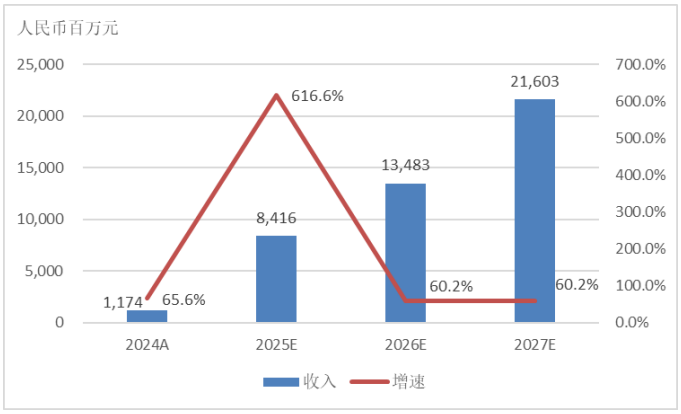


	2024A	2025E	2026E	2027E
销售费用率	6.0%	1.1%	0.9%	0.7%
研发费用率	103.5%	18.8%	15.2%	12.4%
管理费用率	15.4%	2.8%	2.3%	1.8%
毛利率	56.7%	60.0%	60.0%	60.0%

资料来源：公司资料，第一上海预测

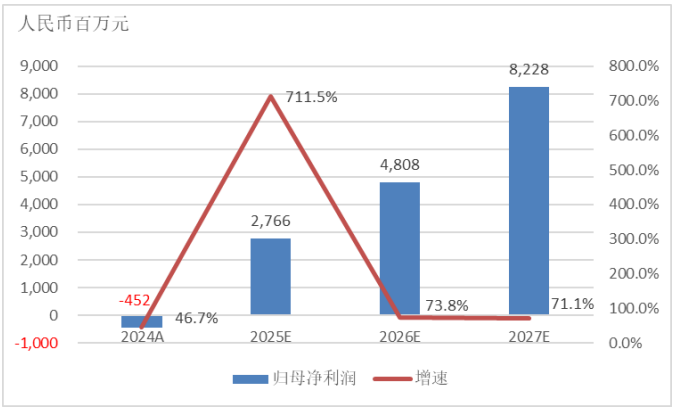
基于以上合理预测，我们预计公司 2025-2027 年度归属母公司股东净利润将达到 27.7 亿、48.1 亿和 82.3 亿元；实现每股收益分别为人民币 6.63 元、11.52 元和 19.71 元人民币。

图表 17：总收入及增速（2024A-2027E）



资料来源：公司资料，第一上海预测

图表 18：净利润及增速（2024A-2027E）



资料来源：公司资料，第一上海预测

公司估值比较

我们选取中国 A 股芯片设计公司，包括国产化 CPU 及 GPU 龙头企业海光信息、国产化存储及 MCU 龙头企业兆易创新，端侧芯片龙头企业恒玄科技、瑞芯微，基带及物联网芯片龙头翱捷科技，以及智能驾驶及机器人芯片龙头企业地平线机器人，黑芝麻智能等，作为公司的同类公司。

我们发现寒武纪在成长性上大幅领先于同行，在 PE 估值上处于同类公司的较低位置。我们认为公司作为国产算力芯片的龙头企业之一，拥有远高于上述同类公司的利润率和成长能力，应该享有估值溢价。

图表 19：同类公司估值比较（股价为截至 2025 年 6 月 2 日收市价）

公司	代码	现价	市值（亿）	PE（倍）			PS（倍）			收入CAGR 24-27
				2025E	2026E	2027E	2025E	2026E	2027E	
寒武纪	688256 CH	603.66	2,520.0	91.5	52.5	30.6	29.9	18.7	11.7	164%
海光信息	688041 CH	136.13	3,164.1	96.6	68.8	52.3	22.8	16.8	12.8	39%
兆易创新	603986 CH	112.30	745.7	46.7	35.6	33.7	7.9	6.6	5.7	21%
恒玄科技	688608 CH	391.66	470.2	56.7	41.5	28.1	10.1	7.8	6.1	33%
瑞芯微	603893 CH	144.30	604.5	70.9	53.6	37.5	14.1	11.1	8.8	30%
翱捷科技	688220 CH	75.30	315.0	(90.7)	3,273.9	149.4	7.0	5.3	4.2	30%
地平线机器人*	9660 HK	7.05	930.6	(41.6)	(95.4)	108.1	24.0	14.4	9.8	54%
黑芝麻智能*	2533 HK	18.08	114.2	(9.5)	(15.2)	(319.9)	12.4	7.5	5.1	63%

资料来源：彭博，第一上海预测，

*现价及市值为港币计价，港币人民币汇率按 0.92 换算

目标价 987 元，买入评级

公司是国产算力领军企业，在算力国产化替代浪潮中已经取得先发优势，包括 1）产品生态更完善，2）供应链保障更充分，3）资金实力更充足。在人工智能蓬勃发展的大趋势下，公司有望凭借其完善的软硬件协同能力，以及在客户中积累的先发优势，迎来业务的爆发式增长。

我们认为公司是国产算力的稀缺标的，有望同时受益于人工智能的蓬勃发展以及算力的国产化替代。给与公司 2027 年 50 倍市盈率，对应 4,114 亿市值。给与公司目标价 987 元人民币，仍有 63.5%上涨空间，首次评级买入。

风险因素

1. AI 芯片需求放缓：生成式人工智能的应用落地仍在初级阶段，AI 硬件的投入和产出可能出现不匹配，导致行业投资的减缓；
2. 供应链波动：公司是芯片设计公司，采用 Fabless 模式运营。中美竞争可能对公司的供应链稳定造成负面影响；
3. 行业竞争加剧；下游客户自研芯片、竞争对手加大投入、其他龙头厂商争抢份额、地域保护等均可能导致行业竞争加剧，对产业造成负面影响。

主要财务报表

损益表

百万元人民币，财务年度截至12月底

	2023年 实际	2024年 实际	2025年 预测	2026年 预测	2027年 预测
收入	709	1,174	8,416	13,483	21,603
毛利	491	666	5,050	8,090	12,962
销售费用	82	70	91	118	154
管理费用	154	181	235	305	397
研发费用	1,118	1,216	1,581	2,055	2,671
其他收益	144	220	220	220	220
财务费用	(45)	(19)	0	0	0
营业利润	(876)	(456)	3,254	5,656	9,680
营业外收支	1	(0)	0	0	0
税前盈利	(875)	(456)	3,254	5,656	9,680
所得税	3	1	488	848	1,452
少数股东应占利润	(30)	(5)	0	0	0
归母净利润	(848)	(452)	2,766	4,808	8,228
增长					
收入(%)	-2.7%	65.6%	616.6%	60.2%	60.2%
营业利润（%）	166.2%	48.0%	814.0%	73.8%	71.1%
归母净利润（%）	167.5%	46.7%	711.5%	73.8%	71.1%

财务分析

百万元人民币，财务年度截至12月底

	2023年 实际	2024年 实际	2025年 预测	2026年 预测	2027年 预测
盈利能力					
毛利率	69.2%	56.7%	60.0%	60.0%	60.0%
营业利润率	-123.5%	-38.8%	38.7%	42.0%	44.8%
归母净利率	-119.6%	-38.5%	32.9%	35.7%	38.1%
营运表现					
SG&A/收入(%)	33.2%	21.3%	3.9%	3.1%	2.5%
研发费用率	157.5%	103.5%	18.8%	15.2%	12.4%
实际税率（%）	——	——	15.0%	15.0%	15.0%
库存周转天数	163.5	1,256.1	203.7	203.7	165.3
应付账款天数	390.3	364.4	82.6	67.0	54.4
应收账款天数	326.8	93.4	101.7	101.7	101.7
财务状况					
总负债/总资产	0.11	0.19	0.15	0.11	0.08
收入/总资产	0.11	0.17	0.81	0.83	0.84

资产负债表

百万元人民币，财务年度截至12月底

	2023年 实际	2024年 实际	2025年 预测	2026年 预测	2027年 预测
现金	3,954	1,986	3,516	5,946	11,315
应收账款	644	305	2,377	3,809	6,103
存货	99	1,774	1,904	3,051	3,967
其他流动资产	950	1,735	1,429	2,242	2,903
总流动资产	5,648	5,800	9,227	15,049	24,289
固定资产	142	231	436	590	705
长期股权投资	230	247	247	247	247
其他固定资产	399	439	439	439	439
总资产	6,418	6,718	10,349	16,324	25,679
应付帐款	237	515	772	1,004	1,305
短期银行贷款	0	100	100	100	100
其他短期负债	226	203	203	203	203
总短期负债	463	818	1,075	1,307	1,608
长期银行贷款	0	0	0	0	0
其他负债	225	469	469	469	469
总负债	689	1,287	1,545	1,776	2,077
少数股东权益	80	8	8	8	8
股东权益	5,650	5,423	8,796	14,540	23,594

现金流量表

百万元人民币，财务年度截至12月底

	2023年 实际	2024年 实际	2025年 预测	2026年 预测	2027年 预测
净利润	(878)	(457)	2,766	4,808	8,228
折旧	98	88	95	147	185
摊销	232	122	0	0	0
营运资金变化	(190)	(1,443)	(1,946)	(2,347)	(2,908)
其他	(143)	(72)	(154)	(122)	(165)
营运现金流	(596)	(1,618)	1,069	2,730	5,669
资本开支	(100)	(366)	(300)	(300)	(300)
其他投资活动	525	(46)	760	0	0
投资活动现金流	425	(412)	460	(300)	(300)
负债变化	0	100	0	0	0
股本变化	1,804	56	0	0	0
股息	0	(0)	0	0	0
其他融资活动	(147)	(108)	0	0	0
融资活动现金流	1,657	48	0	0	0
现金变化	1,486	(1,982)	1,530	2,430	5,369
期初持有现金	2,467	3,954	1,972	3,502	5,931
期末持有现金	3,954	1,972	3,502	5,931	11,301

资料来源：公司资料，第一上海预测

第一上海证券有限公司
香港中环德辅道中 71 号
永安集团大厦 19 楼
电话: (852) 2522-2101
传真: (852) 2810-6789