

算法芯片化助力，全产业链发展打造推理 AI 龙头

2025 年 09 月 30 日

► **公司 AI 产品以消费级、企业级、行业级三方向发展。**公司以“算法芯片化”作为底层优势，在消费级、企业级、行业级三种场景部署产品和服务。企业级场景主要为 AI 推理芯片与相关产品、算力服务及 IP 授权服务等，其中公司自研的 DeepEdge 系列通过“算力积木”架构可提供 8-128T 算力，已经在智慧交通、机器人、家庭主机、推理服务器等领域与头部企业合作并实现商用，16.07 亿元的 AI 算力服务订单已经验收并正常回款。消费级场景主要包括噜咔博士 AI 拍学机、AI 眼镜等 AI 原生产品和 AI 耳机、AI 手表等可穿戴设备模组等 AI 赋能产品，并实现了快速增长。行业级注重项目质量，考核目标从销售额转向利润端。

► **Deepseek 开源模型，推动“算法-芯片协同进化”。**Deepseek R1 通过关键设计的奖励系统、训练模板、自进化现象，以及用于强化学习的核心算法 GRPO，用低算力的 GPU 训练出了媲美 OpenAI o1 的模型，证实了在 AI 行业中，软件端算法的创新可弥补硬件端差异，也可推动行业整体发展并产生经济效应。云天励飞持续强调的算法芯片化，通过自定义指令集、处理器架构及工具链的协同设计，通过软件端的改良提升芯片能力，从而拓展推理应用市场。

► **模型平权推动推理侧市场发展，公司相关硬件产品有望受益。**模型平权带动 AI 应用落地，应用的发展未来有望推动推理侧市场规模扩容。云天励飞推出 X6000 加速卡。X6000 具备 256T 算力、128GB 显存容量；采用 C2C Mesh 互联技术，可实现卡间高速互联，带宽达 64GB/s，最大可实现 64 张卡的互联；可应用于语言、视觉、多模态等各类大模型的推理加速。公司推出由 X6000 推理卡组成的 X6000 推理服务器，高效适配 Deepseek、通义千问等主流大模型。

► **重点布局推理市场，迎接推理市场起量机遇。**据公司招股说明书数据，2020-2024 年，中国 AI 推理芯片相关产品及服务行业市场规模预计从 113 亿元增长到 1626 亿元，CAGR 达 94.9%；预计 2029 年市场规模或达 13830 亿元，2024-2029 年 CAGR 达 53.4%。AI 推理芯片可分为云、边、端三种，不同算力范围支持不同功能特征。公司目前在销产品为算力积木 1.0 架构下的边缘和云的大算力推理芯片，已设计三代推理芯片迭代规划，可满足在推理场景下各类需求，公司重点布局迎接推理市场起量。

► **投资建议：**公司部署 AI 算力到应用全产业链，多项平台技术优势明显。此外，公司多方向开拓 AI 市场，通过企业级、消费级、行业级三方向渠道开拓，销售额增长显著。后续随着自研芯片及相关产品在云端推理市场，公司营收有望迎来快速增长。我们预计公司 2025-2027 年营业收入为 12.85、15.84、19.36 亿元，对应 PS 分别为 23X、19X、15X，首次覆盖，给予“推荐”评级。

► **风险提示：**核心芯片产品销售不及预期，技术研发不及预期，端侧设备产品销售不及预期风险。

盈利预测与财务指标

项目/年度	2024A	2025E	2026E	2027E
营业收入 (百万元)	917	1,285	1,584	1,936
增长率 (%)	81.3	40.0	23.3	22.2
归属母公司股东净利润 (百万元)	-579	-438	-390	-259
增长率 (%)	-51.1	24.3	11.0	33.7
每股收益 (元)	-1.61	-1.22	-1.09	-0.72
PE	/	/	/	/
PS	32	23	19	15

资料来源：Wind，民生证券研究院预测；（注：股价为 2025 年 9 月 29 日收盘价）

推荐

首次评级

当前价格：

82.77 元



分析师 吕伟

执业证书：S0100521110003

邮箱：lvwei_yj@glms.com.cn

目录

1 推理 AI 领军企业全产业链布局发展	3
1.1 “算法+芯片”支持，三种场景开拓销售 AI 产品和解决方案	3
1.2 合理股权架构支持公司发展	4
2 推理市场广阔，公司算法芯片化技术优势凸显	7
2.1 AI 技术迭代升级，AI 推理市场持续扩容	7
2.2 端云协同、算法芯片化助力公司拓展市场	9
3 凭借技术优势，消费级、企业级、行业级应用场景拓展	13
3.1 消费级应用场景：多 AI 端侧设备同步拓展	13
3.2 企业级应用场景：以推理芯片为核心，拓展企业云、边、端业务场景	17
3.3 行业级应用场景：智慧解决方案经验成熟，多行业拓展拓宽渠道	32
4 营收维持高增，利润端持续减亏	37
4.1 营收进入高速增长期，利润端大幅减亏	37
4.2 重研发投入，人均创收恢复高增速	39
5 盈利预测与投资建议	41
5.1 盈利预测与业务拆分	41
5.2 估值分析	42
5.3 投资建议	43
6 风险提示	44
插图目录	46
表格目录	47

1 推理 AI 领军企业全产业链布局发展

1.1 “算法+芯片”支持，三种场景开拓销售 AI 产品和解决方案

公司以“算法+芯片”作为底层优势，在三种场景开拓销售 AI 产品和解决方案。公司自 2014 年成立以来深耕人工智能领域，致力于面向场景实现算法和芯片技术的融合。公司拥有算法和芯片两大技术平台，凭借“算法芯片化”的核心能力和“端云协同”的技术路线不断推进推理 AI 在消费级、企业级和行业级场景的产品应用。

图1：公司产品出货形态及面向的终端场景



资料来源：公司公告，民生证券研究院

2014 年成立，AI 技术多领域取得成绩。公司于 2014 年成立，2015 年第一代深度学习处理器投入研发入选国务院侨办“重点华侨华人创业团队”，2016 年服务 G20 峰会核心安保圈，2017 年“深目”系统在深圳、北京、上海、杭州等 20 多个城市及东南亚多个国家落地协助破获“天图案”，2019 年人工智能芯片获国家发改委、科技部、工信部重大专项大满贯与深圳市政府办公厅战略合作。2020 年，创始人陈宁入选深圳特区 40 年 40 人，获习近平总书记接见，荣获第十届吴文俊人工智能专项奖芯片项目一等奖。2021 年，入选工信部人工智能重点任务揭榜优胜名单。

图2：公司发展历程

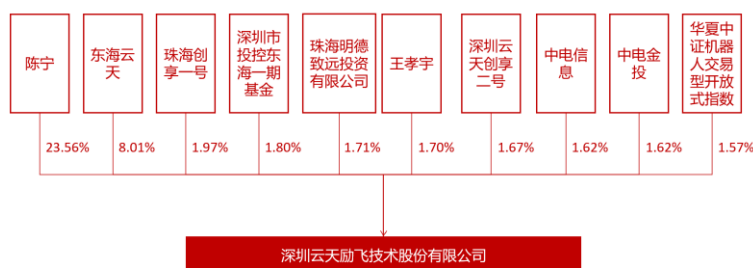


资料来源：公司官网，今日头条，公司公告，民生证券研究院

1.2 合理股权架构支持公司发展

公司股权架构稳定。截至 2025 年中报，公司董事长兼总经理陈宁博士直接持有公司 23.56%的股权，并通过珠海明德致远投资有限公司间接持有公司 1.71%的股份，合计持有公司 25.27%的股权，是公司第一大股东以及实际控制人。公司前十大股东中，珠海云天创享一号、珠海云天创享二号系公司员工持股平台；持股 8.01%的第二大股东是由深圳市特区建设发展集团有限公司控股的东海云天；持股 1.97%的第三大股东是公司持股平台珠海创享一号。

图3：云天励飞股权结构



资料来源：公司公告，民生证券研究院，截止 2025 年中报

公司的高级管理人员多数同为核心技术人员，团队研发与管理经验深厚。公司董事长兼总经理陈宁曾于 2009 年至 2014 年在中兴通讯担任 IC 技术总监，曾获得深圳市国家级领军人才。副总经理李爱军曾在中兴通讯先后担任芯片项目经理、大项目经理、手机芯片部长、手机芯片研发总经理等职；作为项目经理带领项目组完成国内首款 28nmLTE 多模手机 modem 芯片的研发和量产，具有资深的研发管理经验。公司副总经理程冰曾在深圳中兴力维技术有限公司历任程序员、开发经理、产品总监职务；自其参加工作以来作为项目负责人参与过多个省市级重大课题科研项目研究。

表1：公司管理层结构
管理层结构

陈宁	董事长、 总经理、 核心技术 人员	2006年毕业于美国佐治亚理工学院，获得博士学位；2006年5月至2009年5月，在美国飞思卡尔半导体公司担任系统架构师；2009年5月至2014年7月，在中兴通讯担任IC技术总监；2014年8月至今，在公司担任执行董事/董事长兼总经理，也是公司的核心技术人员。陈宁曾获得深圳市国家级领军人才、深圳市海外高层次人才计划（“孔雀计划”）A类人才、第十六届广东省青年“五四奖章”、深圳经济特区建立40周年创新创业人物和先进模范人物等荣誉称号。
邓浩然	董事、财 务总监、 董秘	2005年5月至2007年10月，在安永华明会计师事务所担任审计员、税务咨询顾问；2008年3月至2015年10月，在华为技术有限公司先后担任区域税务经理、国家子公司财务经理、中国税务管理部部长、运营商BG售前财经五级专家等职务；2015年10月至2018年1月，在领益科技（深圳）股份有限公司担任财务副总监、监事；2018年2月加入公司，现任公司董事、财务总监兼董事会秘书。
李爱军	首席技术 官 (CTO)	研究员（科技型企业）职称。2001年4月至2017年2月，在中兴通讯先后担任芯片项目经理、大项目经理、手机芯片部长、手机芯片研发总经理等职；作为项目经理带领项目组完成国内首款28nmLTE多模手机modem芯片的研发和量产，具有资深的超大规模SoC芯片研发功底和丰富的芯片大团队运作的研发管理经验；2017年2月加入公司，现任公司副总经理。2019年1月获得深圳市人民政府颁发的“深圳市龙岗区优秀专家”荣誉称号。
程冰	副总经 理、核心 技术人员	高级工程师职称。2004年毕业于中国民用航空学院，获得学士学位。2006年12月至2015年1月，在深圳中兴力维技术有限公司历任程序员、开发经理、产品总监职务。自参加工作以来，长期从事公共安全领域的产品研发和产品规划工作，作为项目负责人，参与了多个省市级重大课题科研项目研究。2015年11月加入公司，现任公司副总经理。

资料来源：公司公告，民生证券研究院

公司推行股权激励，充分调动人才积极性与凝聚力。公司推出限制性股票激励计划，计划向董事、高级管理人员、核心技术人员、核心骨干等共计233人授予1106.16万股限制性股票。首次授予部分考核目标为以2022年为基数的营收增长率，考核年度为2024-2025两个会计年度。公司通过股权激励进一步提升员工积极性。

表2：2023年公司股权激励情况

授予数量		激励人数	考核年度	考核目标：营业收入增长率（A）	
				目标值（Am）	触发值（An）
1106.16 万股（预留 167.41 万股，占比 15%）	233 人	2024 年	以 2022 年营业收入为基数，2024 年营业收入增长率不低于 60%。	以 2022 年营业收入为基数，2025 年营业收入增长率不低于 42%	
		2025 年	以 2022 年营业收入为基数，2025 年营业收入增长率不低于 90%。	以 2022 年营业收入为基数，2025 年营业收入增长率不低于 63%	
考核目标：营业收入增长率（A）				归属权益数量占授予权益总量的比例	
A≥Am				X=100%	
An≤A < Am				X=A/Am	
A < An				X=0%	

资料来源：公司公告，民生证券研究院

公司于2023年募集资金上市，充足现金补充支撑公司穿越创业研发投入期。

公司在 2023 年 4 月 4 日于科创板上市，公司拟投入募集资金 30 亿元，实际募资 38.99 亿元，扣除发行费后，募集资金净额为 35.84 亿元。公司募集的资金分别用于城市 AI 计算中枢及智慧应用研发项目、面向场景的下一代 AI 技术研发项目、基于神经网络处理器的视觉计算 AI 芯片项目、补充流动资金项目。由于 AI 行业处于发展前期，研发仍需大量投入，资金充足有助于公司穿越研发投入期。

表3：募集资金项目

序号	项目名称	总投资规模 (万元)	拟投入募集资金 (万元)
1	城市 AI 计算中枢及智慧应用研发项目	80,064.72	80,000.00
2	面向场景的下一代 AI 技术研发项目	30,010.00	30,000.00
3	基于神经网络处理器的视觉计算 AI 芯片项目	50,088.60	50,000.00
4	补充流动资金项目	140,000.00	140,000.00
合计		300,163.32	300,000.00

资料来源：公司公告，民生证券研究院

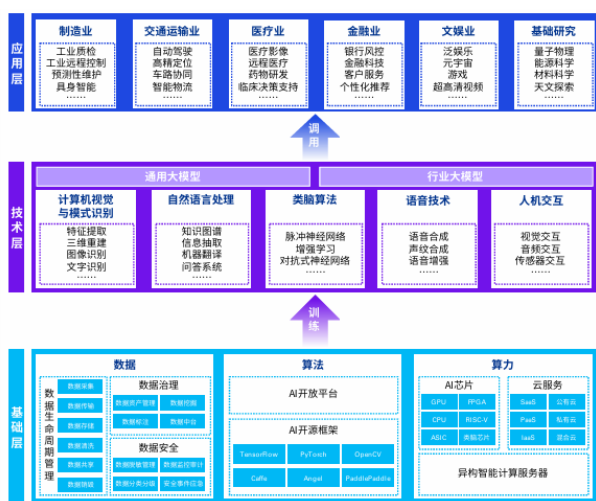
2 推理市场广阔，公司算法芯片化技术优势凸显

2.1 AI 技术迭代升级，AI 推理市场持续扩容

大模型技术迭代升级，AI 推理市场规模持续扩容。现阶段，AI 模型技术持续迭代升级，自 Chat-GPT 出世带动了全世界对 AI 产业关注度提升，AI 产业的投资额及市场规模逐步提升。AI 产业中的大模型随着模型训练技术的提升，模型对于多模态数据处理能力以及行业应用能力快速改善。2025 年 2 月，Deepseek R1 横空出世，该模型采用大规模强化学习，采用少量数据训练出与 ChatGPT o1 同水平的推理模型，此模型开源推动了 AI 平权，任何公司可使用此款低成本高性能的模型，因此进一步加快了 AI 应用落地。

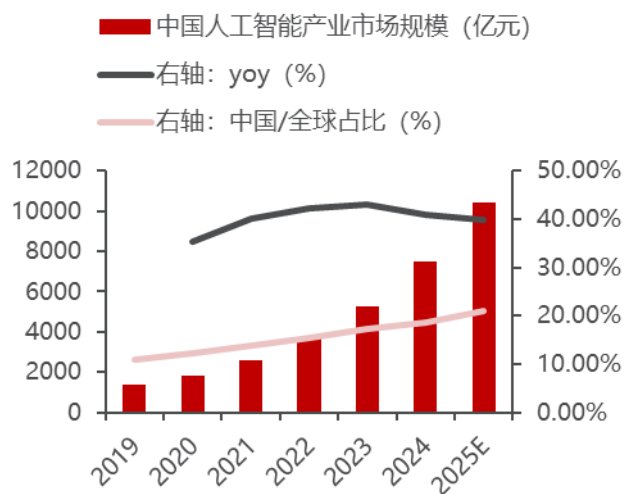
AI 产业链可分为基础层、技术层、应用层。基础层中分为数据、算法、算力及异构智能计算服务器。技术层分为通用大模型、行业大模型，如计算机视觉与模式识别、自然语言处理、类脑算法、语音技术及人机交互。应用层为 AI 模型对于具体行业的落地，可涉及制造业、交通运输行业、医疗业、金融业、文娱业及基础研究等。据艾媒咨询，2019 年-2025 年，中国人工智能产业市场规模有望从 1372 亿元增长至 10457 亿元，CAGR 达 40.28%，中国占全球市场占比有望从 11.00% 提升至 20.90%。随着整个 AI 产业链技术的迭代升级及产业落地速度加快，AI 市场规模持续扩容，有望迎来行业加速扩容期。

图4：AI 产业链



资料来源：毕马威、中关村产业研究院：《2023 年人工智能全域变革图景展望》，民生证券研究院

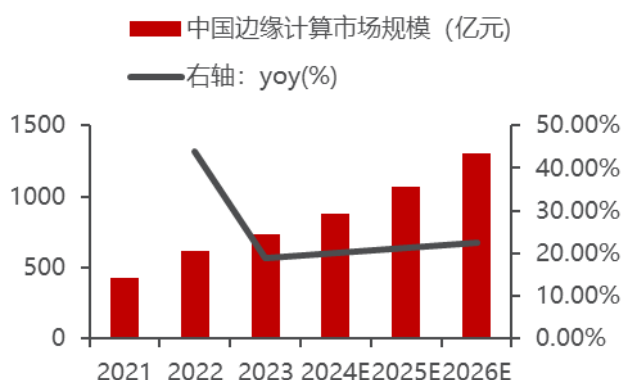
图5：中国人工智能产业市场规模（亿元）&右轴：yoy（%）&右轴：中国/全球占比（%）



资料来源：艾媒咨询，民生证券研究院

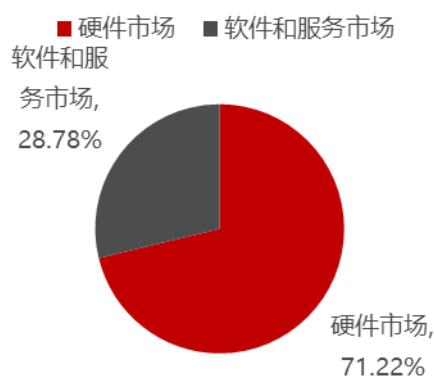
边缘计算市场增速明显，市场潜力广阔。边缘计算通过在网络边缘部署数据处理功能，靠近数据源，有效降低了响应时间并提升了数据处理的即时性。随着数据流量激增，边缘计算凭借低延迟、高带宽的特性，有望为物联网、自动驾驶、智能制造等领域提供实时和高效的数据处理。边缘计算产品结构分为基础设施（边缘硬件）、系统软件（边缘软件平台）和应用与服务（集成与开发服务、网络与通信服务）三大类。市场结构上，其中硬件市场占比达 71.22%，软件和服务市场占比达 28.78%。据赛迪顾问，2021-2026 年，中国边缘计算市场规模有望从 427 亿元增长至 1305.2 亿元，CAGR 达 25.04%。边缘计算市场增速明显，市场潜力广阔。

图6：中国边缘计算市场规模（亿元）&yoy（%）



资料来源：赛迪顾问，民生证券研究院

图7：2023 年中国边缘计算市场结构



资料来源：中商产业研究院，民生证券研究院

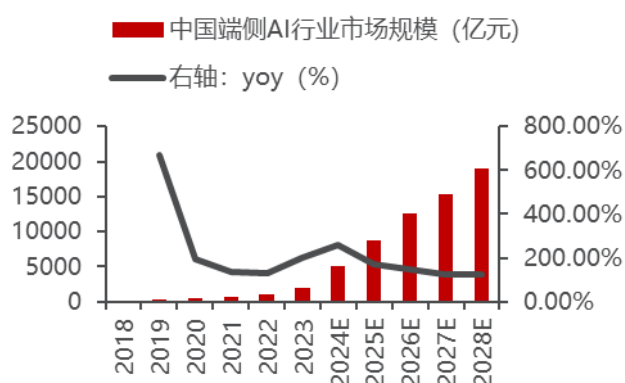
表4：边缘计算行业产品结构

解决方案类别	组件类型	详细分类及描述	应用领域
基础设施	边缘硬件	包括专用于边缘计算环境的物理设备，如服务器、网关、传感器等。	工业自动化、视频分析处理
		边缘服务器：高性能计算设备	智能家居中控、车载信息处理系统
		网关设备：连接和数据转发设备	
系统软件	边缘软件	传感器/执行器：数据收集与执行操作	
		包括操作系统、管理工具、安全软件等，用于边缘设备的运行和管理。	设备管理平台、远程监控系统
		边缘计算 OS：边缘设备操作系统	加密软件、入侵检测系统
应用与服务	集成与开发服务	管理软件：设备监控与维护	
		安全软件：数据保护与访问控制	
		提供云集成、应用开发支持和行业解决方案等服务。	混合云管理、数据同步服务
	网络与通信服务	云服务集成：边缘与云服务结合	智能医疗系统、智慧城市管理平台
		应用开发平台：支持应用创新	
		行业解决方案：定制化服务	
		包括网络基础设施建设、通信技术支持等。	5G 基站部署、企业级路由器管理

资料来源：中商产业研究院，民生证券研究院

端侧 AI 市场维持高增速，多应用场景增加市场潜力。端侧 AI 将 AI 能力下沉到设备端，减少云服务依赖，从而实现数据隐私保护、更低网络需求和更快响应速度。端侧 AI 强调本地化运行，应用场景涉及多类端侧：手机、自动驾驶汽车、AR、PC、电视、耳机、音响、安防设备、智能手表和机器人等。多应用场景提升了端侧 AI 市场的增长潜力。据头豹产业研究院数据，2018-2023 年，中国端侧 AI 市场从 41 亿元增长至 1939 亿元，CAGR 达 116.25%；经预测，2024-2028 年，市场从 5001 亿元增长至 19071 亿元，CAGR 达 57.96%，端侧 AI 市场呈现较高增速。

图8：中国端侧 AI 行业市场规模（亿元）&yoy（%）



资料来源：头豹产业研究院，民生证券研究院

图9：端侧 AI 应用场景分类



资料来源：智研资讯，民生证券研究院

2.2 端云协同、算法芯片化助力公司拓展市场

2.2.1 发展战略：双核心技术优势拓展 AI 行业级、企业级、消费级应用

云天励飞推理 AI 战略围绕一主一辅两条发展线进行布局。公司主线以 NPU 推理芯片为核心，涵盖边缘推理芯片与云端推理芯片，构建起具备更高性价比的计算平台，并以此为基础推进大模型一体机、软硬一体推理服务器等产品，以及相关服务及生态。辅线包含自有品牌“噜咔博士”和子公司岍丞技术的消费侧业务，为面向消费者的产品及 AI Agent 硬件服务，提供更可靠的 AI 落地应用载体，如 AI 眼镜、拍学机、AI 玩具等。

2.2.2 算法芯片化增强公司竞争力

场景算法芯片化拓展推理 AI 市场。公司搭建了算法分析-指令集定义-芯片架构设计-工具链设计的 AI 芯片研发设计流程，通过自定义指令集、处理器架构及工具链的协同设计，实现算法技术芯片化。公司 AI 芯片主要面向嵌入式前端和边缘计算应用，具备高性能、低功耗、低成本的优点。在 2022 年流片的 Edge10 针对 Transformer 的计算需求进行了优化，凭借自研芯片 DeepEdge10 创新的

D2Dchiplet 架构打造的 X2000、X6000 系列推理卡,可应用于语言、视觉、多模态等各类大模型的推理加速,目前已适配云天书、通义千问、百川智能、Llama2/3 等近 10 个主流大模型。

算法芯片化技术平台特点:此平台为公司基础技术研究平台,基于此平台重点研究算法与芯片协同设计技术,包括大模型算法技术、视觉语义类算法技术、AI 处理器设计仿真验证技术、芯片工具链技术等。

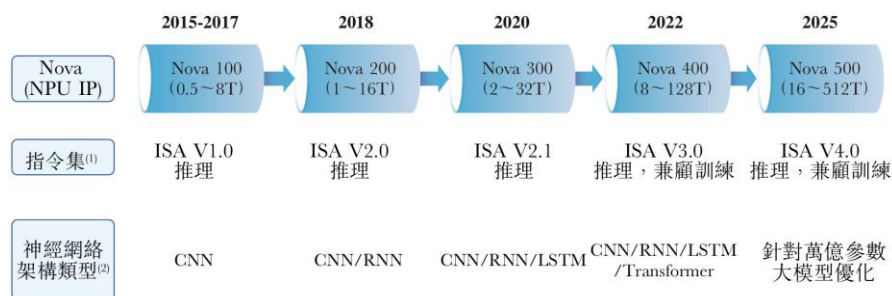
图10: 算法芯片化基础技术



资料来源:公司公告,民生证券研究院

公司持续加强神经网络处理器技术研发,持续迭代逐步实现自主可控。公司定义了针对深度学习神经网络算法优化的 ISA。公司 NPU 产品 Nova,通过差异化设计战略,可同时应对终端、边端与云端推理场景。公司 NPU 指令集由 ISA V1.0 升级到 ISA V4.0,架构大幅演进,增强了对更广泛模型和运算能力的支援。同时通过高并发流水线并行,该架构能够对复杂 AI 工作负荷进行高效处理,大幅提升了计算吞吐量和跨应用场景部署效率。

图11: 公司指令集进步路线



资料来源:公司公告,民生证券研究院

Deepseek 开源模型,推动“算法-芯片协同进化”。Deepseek 开源推理模型 R1 引起产业变革,低成本高效能的模型开源,推动了行业的整体发展。由于硬件

GPU 使用受限，Deepseek 采用软件端技术创新提升训练效率，整体训练成本约为 550 万美元，仅为 OpenAI o1 开发费用的 3%。**Deepseek R1 通过关键设计的奖励系统、训练模板、自进化现象，以及用于强化学习的核心算法 GRPO，用低算力的 GPU 训练出了媲美 OpenAI o1 的模型，证实了在 AI 行业中，软件端算法的创新可弥补硬件端差异，也可推动行业整体发展并产生经济效应。**云天励飞持续强调的算法芯片化，通过自定义指令集、处理器架构及工具链的协同设计，通过软件端的改良提升芯片能力，从而拓展推理应用市场。

表5：DeepseekR1 技术报告总结

技术创新点	主要内容	细则
关键设计	奖励系统	准确性奖励（Rule-based）：通过规则验证数学答案正确性或代码编译结果； 格式奖励：强制模型在<think>和<answer>标签内输出推理过程与答案。
	训练模板	为了训练 DeepSeek-R1-Zero，设计了一个简单的模板。首先产生一个推理过程，然后是最后的答案。
	自进化现象	随着 RL 步数增加，模型逐步涌现出反思（Re-evaluation）、**长链推理（Long CoT）**等行为； 在 AIME 2024 上，Pass@1 从 15.6%提升至 71.0%，多数投票（cons@64）可达 86.7%。
	核心算法	GRPO 通过分组策略优化降低训练成本，省去传统 PPO 中的 Critic（评论家）模型。 这个目标函数的核心思想是通过比较新旧策略在给定问题上的表现，同时确保策略更新不会过于剧烈（通过剪切操作），以及通过调整参考策略来保持一定的探索性，从而实现对策略模型的有效优化。这种方法特别适用于强化学习场景下的大规模数据处理和复杂推理能力的发展。 优势：相比传统 PPO，GRPO 在 32B 模型训练中节省约 40% 计算资源，同时保持策略稳定性。
蒸馏模型的效率突破	这里的蒸馏就是通过将 DeepSeek-R1 的推理模式蒸馏迁移至小模型，大部分得分也很高	技术意义：验证了知识蒸馏在推理任务中的高效性，为边缘设备部署提供可能。

资料来源：CSDN，民生证券研究院

表6：DeepSeek-R1：冷启动与多阶段优化

阶段	描述	数据量	目标
冷启动	人工标注长链推理数据微调基座模型	数千条	提升初始推理质量
RL 阶段 1	基于 GRPO 强化数学/代码推理	大规模	优化准确性
SFT 阶段	拒绝采样生成数据 + 通用能力数据混合训练	80 万条	平衡推理与通用能力
RL 阶段 2	全场景对齐（推理+安全性）	多样化提示	最终对齐

资料来源：CSDN，民生证券研究院

2.2.3 自研多模态大模型的技术研发及商业化落地带动公司销售

“云天天书”系列大模型支持推理设备落地，多场景应用落地带动公司销售额

增长。2024 年，“云天天书”大模型在多模态信息融合理解能力、超长文本生成能力进一步提升，获得 MMbench 多模态评测榜单第一名；在万物识别能力方面，研发高效推理多模态大模型，万物识别能力行业领先，推理效率提升 3 倍以上；结合算法芯片化能力，“云天天书”大模型实现在多款自研芯片以及华为昇腾的无缝适配；目前“云天天书”大模型已覆盖了城市治理、智慧教育、智慧政务、智慧交通、商业企业等多个行业。

3 凭借技术优势，消费级、企业级、行业级应用场景拓展

公司凭借底层软硬件技术优势，根据部署地点和算力范围不同，在端、边、云三个层级，企业级、消费级、行业级三个应用场景，进行相关硬件与软件产品销售。

图12：公司多场景多层次销售产品



资料来源：公司公告，民生证券研究院

3.1 消费级应用场景：多 AI 端侧设备同步拓展

消费级应用场景主要为多种 AI 端侧设备。在消费级业务上，公司销售产品为推理 AI 场景的各类硬件设备，包括可穿戴设备类产品、AI 交互类产品等。2024 年，公司成立子品牌“噜咔博士”，推出大模型原生教育硬件产品，和合作伙伴闪极科技推出联名智能眼镜产品，以及子公司研丞技术为多个知名品牌提供智能耳机和智能手表的芯片模组、操作系统及产品设计服务。**依托公司自研的大模型或系统、软件应用等为消费级产品提供技术基础，以硬件形式进行交付。**

消费级业务未来规划：公司将推出多款消费级产品，构建“硬件+ 内容+服务”生态闭环，建立“线上+线下”全渠道营销体系；启动“噜咔博士”品牌 IP 化运营，通过内容共创强化用户心智和品牌传播力；继探索“硬件+订阅制内容”商业模式；拓展海外市场，开发面向海外市场的多语言版本系列产品。

图13：公司消费级业务

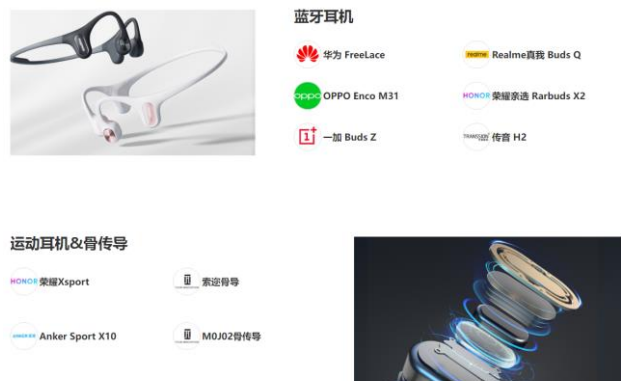

资料来源：公司公告，民生证券研究院

收购可穿戴设备拓展端侧应用落地，子公司并表提升公司营收。公司 2024 年 3 月，公司收购智能穿戴方案设计子公司研丞技术，现已并表。根据天职国际出具的《审计报告》，研丞技术在 2022 年度、2023 年 1-10 月的营业收入分别为 24,145.01 万元、23,394.57 万元，净利润分别为 915.81 万元、149.07 万元。研丞技术在《股权收购协议》中承诺公司在 2024 年度、2025 年度、2026 年度实现的主营业务收入分别为 35,000 万元、42,000 万元、51,000 万元，净利润分别为 1,200 万元、1,440 万元、1,728 万元。

深圳市研丞技术有限公司于 2018 年 08 月 29 日成立，专注于为客户提供无线蓝牙音频解决方案，集合蓝牙模块、方案研发、生产供应和销售以及客户服务为一体的公司，也是上海恒玄科技的第三方合作伙伴。现有员工 450 人左右，其中研发人员 113 人，产线工人 300 人左右。目前公司服务的客户：HUAWEI、HONOR、OPPO、vivo、小米、传音、一加、Realme、酷狗等。公司汇聚了一批在半导体硬件集成电路，底层蓝牙协议等拥有多年经验的资深人士，经过多年积累，已经拥有多项产品认证、实用新型专利以及软件著作权。**公司在细分行业中领先优势明显，2024 年公司中国 IDH 模型智能耳机模组出货量接近 3000 万套，市占率接近 23%，市占率为相关细分市场第一。**

公司硬件设计有以下优势：1) PCB 设计能力，优化底噪能力；2) 射频、天线链路设计能力；抗静电设计能力；3) 抗静电设计能力。**软件定制化有以下优势：**1) OTA，自研加密算法；2) 智能交互，实现耳机与手机端的智能交互；3) 主动降噪，独有的降噪算法经验补强；4) 霍尔回连，独有的机制实现超低功耗。

图14： 研丞技术销售支持产品



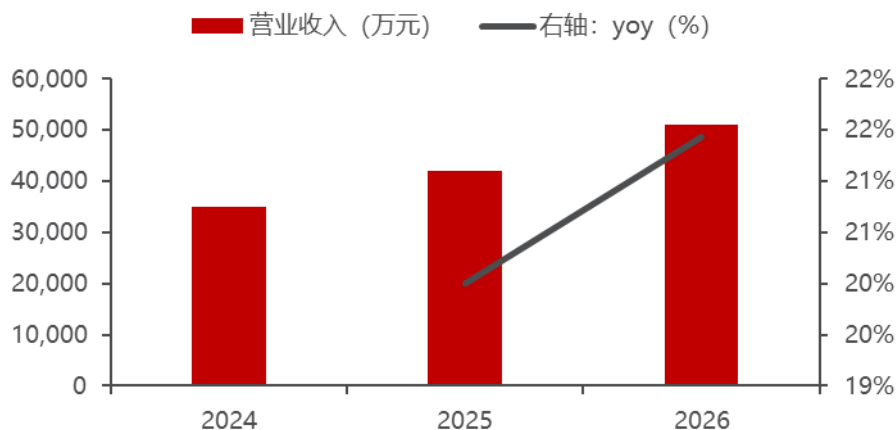
资料来源：子公司研丞技术官网，民生证券研究院

图15： 研丞技术硬件设计能力



资料来源：子公司研丞技术官网，民生证券研究院

图16： 研丞技术收购协议营收目标（万元）&yoy（%）



资料来源：公司公告，民生证券研究院

公司凭借模型与芯片优势开拓端侧设备市场。公司的核心优势为芯片+模型算法的技术优势，公司凭借此技术优势，积极拓展 AI 上下游产业链，公司首先凭借模型优势开拓端侧设备市场。公司推出第一款自主研发品牌“噜咔博士”拍学机，设备内置公司自研大模型，基于百万级万物标签体系与百亿级跨模态语料库，秒级精准识别超 4800 个门类、230 万种物品。AI 拍学机在天猫、京东、抖音等渠道开启销售，产品按照内存区分分为两款，分别为 32G 和 128G，价格分别为 649 元和 799 元，配色分别为荔枝白、石榴粉、柠檬黄、橄榄绿四种。AI 拍学机在 2025 年“618”销售活动中，累计 GMV3000 万元，累计销量 50000 台。

图17：噜卡博士拍学机



资料来源：噜卡博士微信公众号，民生证券研究院

图18：AI 拍学机 618 销售情况



资料来源：公司资料，民生证券研究院

战略投资闪极科技，公司视觉语义模型 Agent 端侧赋能。公司战略投资闪极科技，助力闪极推动 AI 眼镜市场拓展。2024 年 12 月，闪极科技发布国产 AI 眼镜拍拍镜，1499 元零售价、999 元共创版体现性价比。闪极发布会所展示可接入大模型为：文心一言、星火、豆包火山、统义千问、混元、Kimi、智谱、云天励飞、商汤、MiniMax，可接入多款大模型进行赋能。云天作为 Agent 赋能眼镜万物识别能力与语义搜索能力，可实现物品识别科普、拯救“脸盲症”、提示展品背景信息、语义回忆物品与场景。公司凭借模型优势赋能端侧设备逻辑有望逐步兑现。

图19：大模型赋能闪极 AI 眼镜



资料来源：品玩，民生证券研究院

3.2 企业级应用场景：以推理芯片为核心，拓展企业云、边、端业务场景

企业级业务方面，以推理芯片为核心，拓展系列产品；凭借算力储备提供算力支持与运维服务。公司自研 AI 芯片，打造了覆盖 NPUIP、SOC 芯片、SOC 模组、边缘计算盒子、AI 推理服务器及智算集群的系列产品，广泛应用在摄像头终端、车路云、机器人、智算中心等场景。该项业务包括 IP 授权、芯片、服务器整机和智算中心等，未来计划在云边缘推理加速计算等场景提供高能效、低功率和高实时性要求的算力，并寻求在使用量级上持续突破，以 AI 推理加速卡、服务器整机、智算集群提供服务等形式，基于原有生态上持续扩大在 AI 公司、互联网厂商等的国产化 AI 算力机会。

图20：公司企业级产品及服务



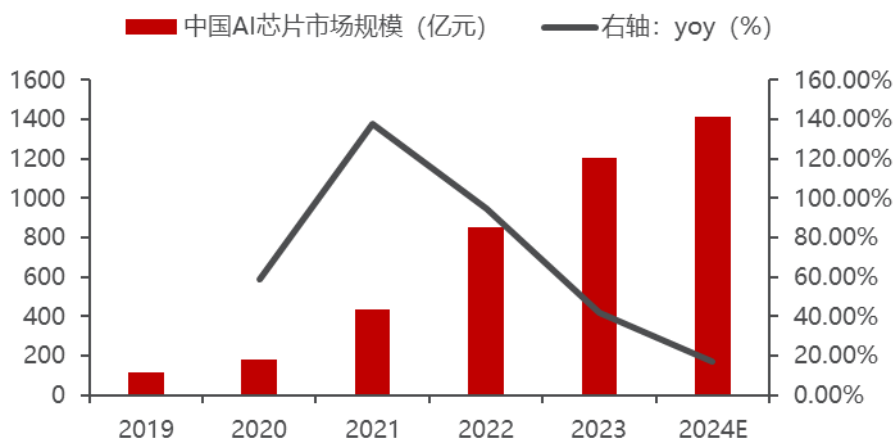
资料来源：公司公告，民生证券研究院

3.2.1 全自研芯片布局推理场景

AI 芯片市场持续增长。随着 AI 市场规模提升带动，主要的基础硬件设施 AI

芯片规模有望迎来高速增长。据中商产业研究院发布的《2024-2029 年中国人工智能芯片行业市场发展监测及投资潜力预测报告》数据，2019-2024 年，中国 AI 芯片市场规模从 116 亿元增长至 1412 亿元，CAGR 达 64.84%，AI 芯片市场规模持续增长。

图21：中国 AI 芯片市场规模（亿元）&yoy（%）



资料来源：中商产业研究院，民生证券研究院

不同场景下 AI 芯片需求产生差异。根据承担任务不同，AI 芯片可分为训练芯片、推理芯片。训练芯片用于构建神经网络模型，需要高计算性能、低功耗、注重计算能力。推理芯片利用训练好的神经网络模型进行推理预测，比较关注低延时、低功耗，对计算能力要求较低。

表7：算力芯片应用场景分类

序号	应用场景	芯片需求	典型应用领域	功耗
1	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	云计算数据中心、企业私有云等	高
2	边端	对功耗/性能/尺寸要求介于终端与云端之间、推理任务为主、多用于插电设备、硬件产品形态较少	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	中
3	终端	低功耗、高效能、推理任务为主、成本敏感、硬件产品形态众多	各类消费类电子物联网产品等	低

资料来源：寒武纪招股说明书，民生证券研究院

AI 芯片目前以 GPU 为主，多品类各有特点，其他品类或逐步占据市场。AI 芯片是针对 AI 算法做了特殊加速设计的芯片。AI 芯片主要包括 GPU、FPGA，以及以 TPU、DPU、NPU 为代表的 ASIC 芯片。目前以 GPU 数量占据主导，未来随着推理需求的提升，以 TPU、DPU、NPU 为代表的 ASIC 芯片有望逐步占据主流市场。

表8：AI 芯片分类

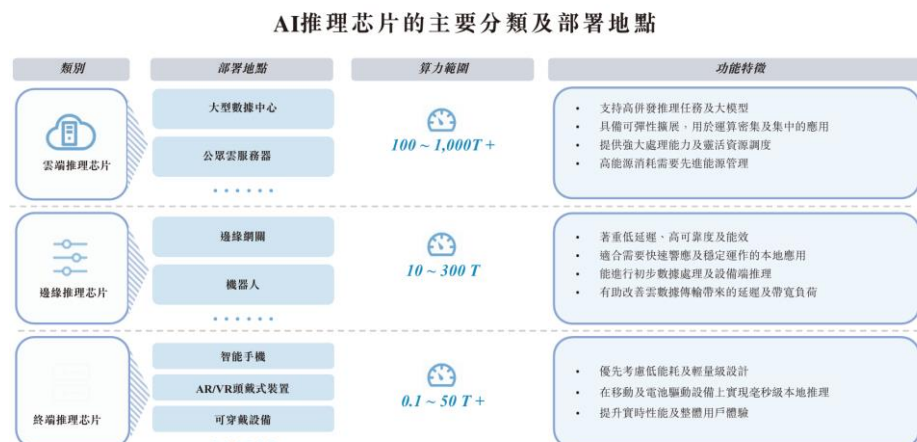
种类	简介
图形 GPU	通用性较强且适合大规模并行运算；设计和制造工艺成熟，但其图形渲染功能无法在推理端完全发挥计算能力。主要应用于高级复杂算法和通用性人工智能平台。
算力 GPU	相较传统的 GPU，缩减了图形图像显示、渲染等功能实现，具有更优的计算能效比，但其算力性能及设计难度极高。主要应用于互联网、安防、金融等通用计算领域和人工智能平台。
FPGA	可通过编程灵活配置芯片架构适应算法迭代，平均性能较高；功耗较低；开发时间短。但是，量产单价高；峰值计算能力较低；硬件编程困难。适用于各种具体的行业
ASIC	通过算法固化实现极致的性能和能效、平均性很强；功耗很低；体积小；量产后成本最低。但是前期投入成本高；研发时间长；技术风险大。

资料来源：中商产业研究院，民生证券研究院

AI 推理芯片可分为云、边、端三种类型，不同算力范围支持不同功能特征。

AI 推理芯片从类别上可分为云端、边缘端、终端。云端推理芯片主要部署在大型数据中心、公众云服务器，算力范围通常在 100~1000T，支持高并发推理任务及大模型。边缘推理芯片主要部署在边缘网关、机器人，算力范围 10~300T，适合需要快速响应及稳定运作的本地应用。终端推理芯片通常部署在智能手机、AR/VR 头戴式装置、可穿戴设备，算力范围 0.1~50T，在移动及电池驱动设备上实现毫秒级本地推理。

图22：AI 推理芯片主要分类及部署地点

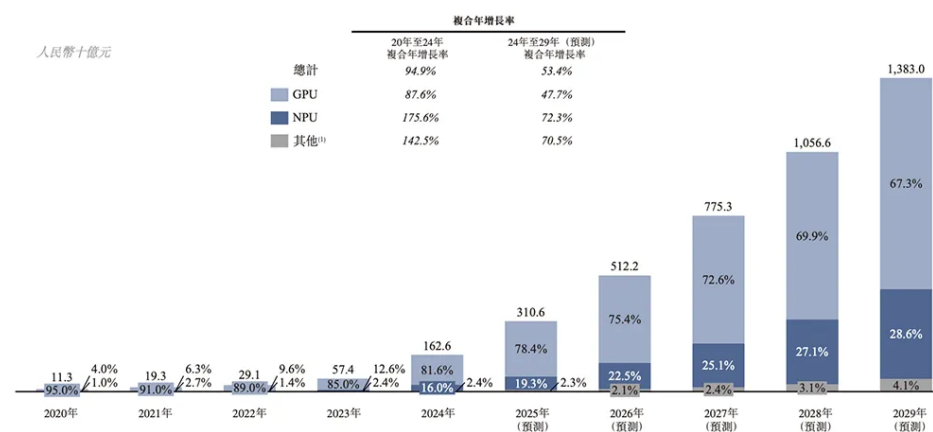


资料来源：IDC，灼识咨询，公司招股说明书，民生证券研究院

推理芯片市场规模未来高速增长，NPU 在细分品类中增速最高。根据公司招股说明书，中国 AI 推理芯片相关产品及服务行业快速增长，市场规模预计从 2020 年的 113 亿元增长到 2024 年的 1626 亿元，CAGR 达 94.9%；预计从 2024 年至 2029 年 CAGR 达 53.4%，2029 年市场规模或达 13830 亿元。从细分品类增速上看，2024-2029 年 GPU、NPU、其他品类 CAGR 分别为 47.7%、72.3%、70.5%，NPU 在 AI 推理芯片中增速最高。

图23：中国 AI 推理芯片相关产品及服务行业市场规模（十亿元）&占比（%）

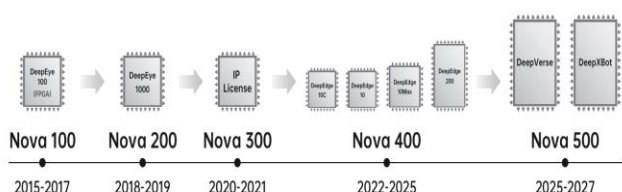
中國按收入劃分的AI推理芯片相關產品及服務行業的市場規模，
2020年至2029年（預期）



资料来源：IDC，灼识咨询，民生证券研究院

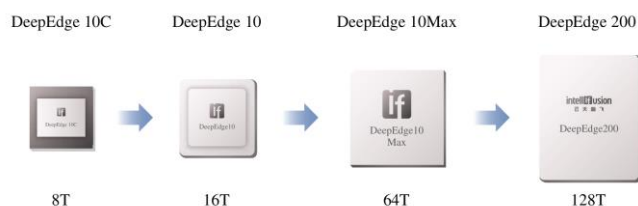
多代 AI 推理芯片持续发展，算力水平持续提升。公司推理芯片目前已更新至第四代，DeepEdge 系列对应 Nova 400，采用算力积木设计，融合 D2D chiplet 和 C2C mesh tour 互连技术。DeepEdge10 系列可支持单个芯片 8T 至 128T 的算力需求。

图24：公司 AI 推理芯片发展规划



资料来源：公司招股说明书，民生证券研究院

图25：DeepEdge 推理芯片系列

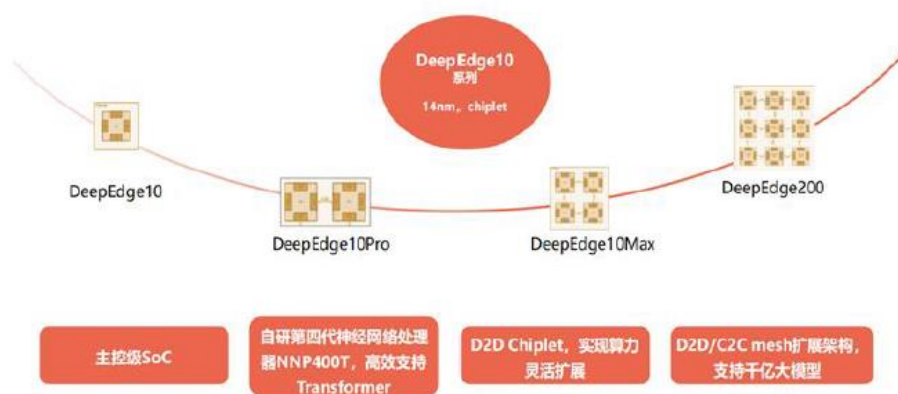


资料来源：公司招股说明书，民生证券研究院

公司支持多芯粒扩展 Chiplet 技术，全自研 DeepEdge10 芯片满足市场多样化需求。 DeepEdge10 基于公司自研的神经网络处理器（NNP400）架构，采用国内先进工艺、支持多芯粒扩展的 Chiplet 技术，DeepEdge10E 可提供 16TOPS（INT8）整型计算和 2TFLOPS FP16 浮点计算的深度学习推理计算算力，满足市场对处理芯片在算法的多样性、准确性、算力密度及效能方面的要求。

针对各类应用场景，云天励飞已开发出 Edge10C、Edge10 标准版、Edge10E 和 Edge10Max 系列芯片。Edge10 系列主要应用于边缘大模型推理领域，能高效支持 Transformer 模型中的矩阵乘法运算。DeepEdge10 芯片主要应用包括摄像头、边缘计算设备、机器人、汽车智能座舱等行业的客户。依托 DeepEdge10 创新的 D2Dchiplet 架构打造的 X6000 系列推理卡，目前已适配多个主流大模型。

图26：云天励飞 DeepEdge10 系列



资料来源：公司公告，民生证券研究院

Edge10Series 芯片平台支持大模型。据云天励飞在 2024 年 AI 芯片峰会的宣传，公司芯片平台支持 Transformer 模型 (SwinTransformerV1、SwinTransformerV2, DETR、ViT、BERT)、BEV 模型 (FastBEV, PETRv1, PETRv2) CV 大模型 (GroundingDino, SAM, CLIP)、LLM 大模型 (LLaMa1/2/3、天书、百川、通义千问、Atom、Mistral、StableLM、Yi)。生成速率上，Edge10 单芯片为 1.8B@14.5 tokens/s；Edge10-Chiplet 为 1.8B@55tokens/s 或 7B@30tokens/s；Edge10-Mesh 为 70B@60tokens/s。

图27: Edge10Series 芯片平台支持大模型



资料来源: AI 芯片峰会, 民生证券研究院

面对不同场景需求, 推出不同系列产品覆盖全场景需求。DeepEdge10 系列使用公司自研神经网络处理器 NNP400T, 提供自 8T-128T 的算力, 其中 DeepEdge10C、DeepEdge10、DeepEdge10max、DeepEdge200 为面向边缘视频智能计算和边缘大模型推理推出的高集成度、高性能、低功耗轻量级 SOC 主控芯片 DeepEdge200 为面向云端大模型推理的高性能 SOC 芯片, 采用 D2D 的 chiplet 片上互联技术, 实现单芯片算力的灵活可扩, 满足人工智能大模型云端推理对算力的需求。

表9: 云天励飞系列芯片

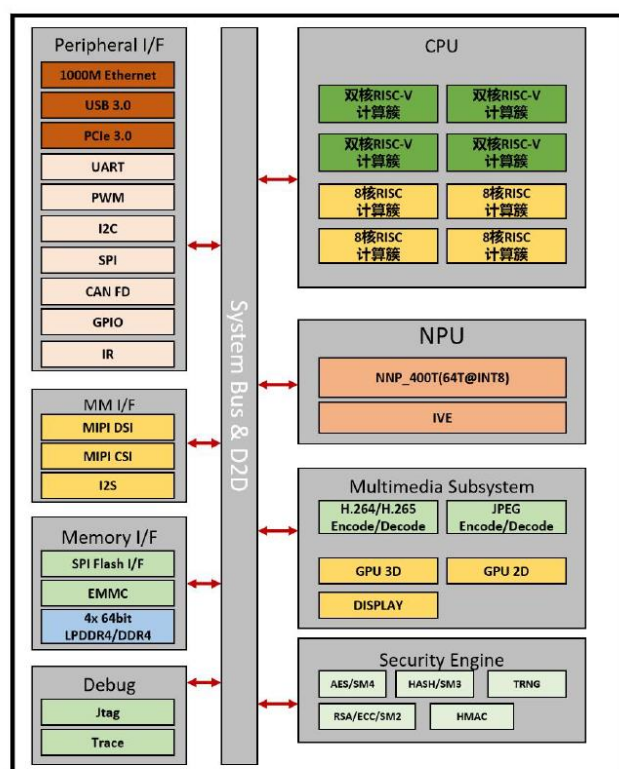
DeepEye 1000	DeepEdge10C	DeepEdge10	DeepEdge10Max
面向计算机视觉的深度学习神经网络处理器芯片	面向边缘视频智能计算推出的一颗高集成度、高性能、低功耗轻量级 SOC 主控芯片	面向边缘视频智能计算推出的一颗高集成度、低功耗、高性能 SoC 芯片	一款面向边缘大模型推理的高性能 SOC 芯片
内置自研 CNN 网络加速引擎, 可以实现高性能低功耗的 CNN 网络模型的加速; 提供强大的可编程运算能力, 满足 CNN 算法实时性处理的运算要求; 自定义指令集和编程框架, 支持主流的 CNN 算法移植, 如人脸识别等, 以及其他计算机视觉 CNN 算法的移植和应用; 可广泛应用于智能摄像机、工业检测、机器人、无人机等领域。	广泛应用于数字城市的安防和泛安防领域的各类视频物联边缘计算场景, 如智能社区、智慧园区、智慧楼宇、智慧工厂的视频分析和机器视觉, 以及各类环境感知、决策控制等机器人和复杂工业控制等领域场景。	采用了云天励飞第四代自研神经网络处理器 NNP400T; 可应用于数字城市的安防和泛安防领域的各类视频物联边缘计算场景, 如智能社区、智慧园区、智能楼宇、智慧工厂的视频分析和机器视觉, 以及各类环境感知、决策控制等机器人和复杂工业控制等领域场景。	采用 D2D 的 chiplet 片上互联技术, 实现单芯片算力的灵活可扩, 满足人工智能大模型边缘推理对算力的需求; 可广泛用于边缘 CV 大模型、NLP 大模型的私有化部署场景。
神经网络处理器: 自研四核 NNP 神经网络处理器, 主频最高 800Mhz;	神经网络处理器: 自研神经网络处理器 NNP400T;	神经网络处理器: 自研神经网络处理器 NNP400T;	通用处理器内核: 自研神经网络处理器 NNP400T;

提供 Total 2TopsAI 算力； 支持 INT16/INT12/INT8 数据类型。	提供 8 TOPS(INT8)/4 TOPS(INT16)/1.5TFLOPS(FP16)的算力； 支持常用深度学习网络，如：CNN，RNN，Transformer，GNN 等。	提供 16 TOPS(INT8)/8 TOPS(INT16)/2TFLOPS(FP16)的算力； 支持常用深度学习网络，如：CNN，RNN，Transformer，GNN 等；	提供 64 TOPS(INT8)/32 TOPS(INT16)/8 TFLOPS(FP16)的算力； 支持常用深度学习网络，如：CNN，RNN，Transformer，GNN 等；
---	---	--	---

资料来源：公司官网，民生证券研究院

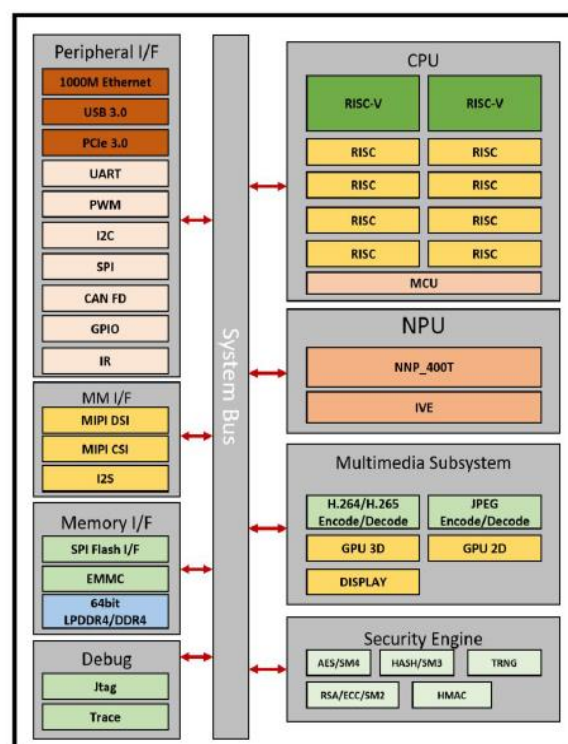
公司自研芯片结构助力多场景应用落地。 DeepEdge10Max 搭载了 4 个八核 64 位 RISC 通用处理器计算簇，4 个双核 64 位 RISC-V 处理器计算簇，支持 RV64GCV 指令集，支持 256-bit 数据位宽。DeepEdge10 搭载了八核 64 位 RISC 通用处理器，双核 64 位 RISC-V 处理器，支持 RV64GCV 指令集，支持 64-bit 数据位宽。

图28: DeepEdge10Max 架构图



资料来源：公司官网芯片架构图，民生证券研究院

图29: DeepEdge10 架构图



资料来源：公司官网芯片架构图，民生证券研究院

先进封装架构和互联扩展技术，高性能互联满足千亿级大模型的部署。

DeepEdge10 选择基于国内自主可控的 14nm 工艺，采用了基于 D2D (Die to Die) 互联的 Chiplet 先进封装架构，通过 D2D Chiplet 先进封装技术，在一颗封装里面装入更多的 DeepEdge10 的 Die，就能够实现性能的成倍提升。DeepEdge10 还支持 C2C (Chip to Chip) Mesh Torus 互联扩展，可以实现高达 512GB 的统一内存、高达 192GB 的统一内存带宽、实现各计算节点的最短传

输延迟、Mesh 可重构特性可以支持 $2 \times 2/3 \times 3/4 \times 4$ Mesh，以满足千亿级大模型的部署需求。

公司打造了一套基于 Chiplet 国产工艺算力积木的架构，包含从底层 IP 到联合生产及封装厂家生产封装以及良率提升，再到顶层软件站调度平台。该架构可把大模型拆解成多个计算模块，在多个 Die 上进行并行计算。公司已实行单 Die、两个 Die、四个 Die、到八个 Die，单 Die16T 算力，可实现单芯片 128T，一张加速卡 256T 大算力推理，全部基于纯国产工艺。

图30: C2C Mesh Torus 互联扩展



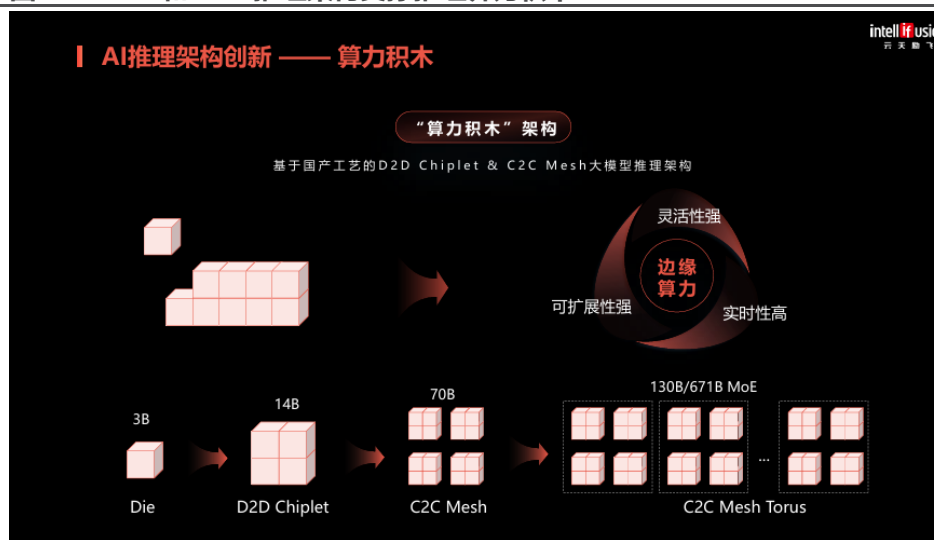
资料来源：芯智讯，民生证券研究院

图31: D2D Chiplet 先进封装架构



资料来源：芯智讯，民生证券研究院

图32: D2D 和 C2C 推理架构支撑推理算力积木

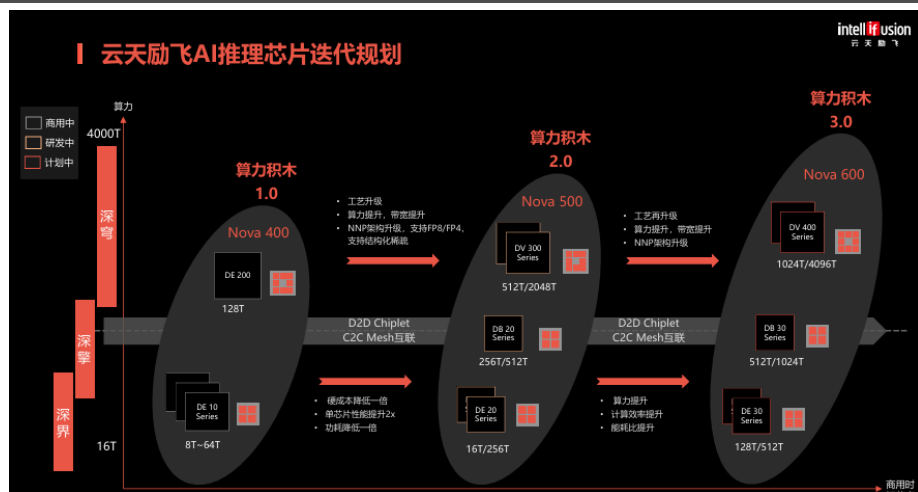


资料来源：公司资料，民生证券研究院

算力积木结构支撑，设计大算力推理芯片迎接推理市场扩容。公司目前在销产品为算力积木 1.0 架构支撑下的边缘和云的大算力推理芯片。在研即将发布的下一代产品为算力积木 2.0 架构支撑，从边缘到具身智能到云端的大算力推理芯片。算力积木 2.0 融合了三项架构创新，Nova500 最新神经网络处理器指令微架构，可适配各种模型，提供高性价比解决方案；国产工艺的 Chiplet；引入 3D memory 堆叠来替代对于 HBM 的需求，从而更好对应在推理场景下的各类需求。同时，公司计划未来升级算力积木体系至 3.0，对应 Nova600，从而支撑更大规模云端集

群类推理。

图33：云天励飞 AI 推理芯片迭代规划



资料来源：公司资料，民生证券研究院

DeepEdge10 已完成 DeepSeek R1 系列模型适配。公司芯片团队完成 DeepEdge10 “算力积木” 芯片平台与 DeepSeek-R1-Distill-Qwen-1.5B、DeepSeek-R1-Distill-Qwen-7B、DeepSeek-R1-Distill-Llama-8B 大模型的适配，可以交付客户使用。DeepSeek-R1-Distill-Qwen-32B、DeepSeek-R1-Distill-Llama-70B 大模型、DeepSeek V3/R1 671B MoE 大模型也在有序适配中。适配完成后，DeepEdge10 芯片平台将在端、边、云全面支持 DeepSeek 全系列模型。

DeepEdge10 芯片已完成国产鸿蒙操作系统的适配。目前已构建适配鸿蒙的芯片编译平台框架，完成图形等子系统的适配。基于 DeepEdge10 的鸿蒙操作系统桌面环境已经可以成功启动，系统原装应用可顺利运行。

图34：DeepEdge10 已完成 DeepSeek R1 系列模型适配

端	边	云
DeepEdge10 DeepSeek-R1-Distill-Qwen-1.5B DeepEdge10 Max DeepSeek-R1-Distill-Qwen-7B	X5000加速卡 DeepSeek-R1-Distill-Qwen-14B	X6000加速卡 DeepSeek-R1-Distill-Qwen-32B X6000 Mesh一体机 DeepSeek V3/R1(671B MoE)



DeepEdge10



DeepEdge10 Max



X6000 加速卡

资料来源：公司公众号，民生证券研究院

Deepseek 技术开源日开放训练技术，加速训练与推理领域拓展。 Deepseek 在连续一周的开源日中，开放了多个提升训练与推理的软件项目，从而帮助客户提升训练与推理效率。其中首日开放的 FlashMLA 技术，通过 MLA 解码优化与分页 KV 缓存技术，显著提升硬件利用率，加速大语言模型解码过程，有效提高响应速度与吞吐量，适用于各种实时对话生成场景。

公司芯片团队在 FlashMLA 开源后，迅速完成了 DeepEdge10 平台与 FlashMLA 的适配工作。在适配过程中，云天励飞采用了 Op fusion tiling、Online softmax、Double buffer、细粒度存算并行等先进技术，并基于自研的 Triton-like 编程语言快速开发验证了高效的 FlashMLA 算子。通过一系列优化，不仅显著提升了计算效率，还大幅降低了显存占用。同时，公司将关代码提交至开源平台 Gitee，贡献开源技术。

表10：Deepseek 开源日开放项目

项目名称	发布时间	简介与性能	应用领域
FlashMLA	DAY1	针对 Hopper GPU 优化的高效 MLA 解码内核，专为处理可变长度序列而设计。使用 CUDA 12.6，H800 SXM5 在内存受限配置下可达 3000 GB/s 带宽，在计算受限配置下可达 580 TFLOPS 算力。	训练和推理
DeepEP	DAY2	DeepEP 是一款专为混合专家 (MoE) 和专家并行 (EP) 设计的通信库，提供了高吞吐量和低延迟的 all-to-all GPU 内核，常用于 MoE 派发和合并操作。该库还支持低精度计算，包括 FP8。	训练和推理
DeepGEMM	DAY3	DeepSeek 发布的是支持稠密和混合专家模型 (MoE) 的 FP8 矩阵乘法 (GEMM) 库，用以驱动 V3 / R1 模型的训练和推理。核心代码仅约 300 行，在 Hopper GPU 上可实现高达 1350+ FP8 TFLOPS 性能。	训练和推理
DualPipe	DAY4	双管道 (DualPipe) 是在《深度搜索-V3 技术报告》中引入的一种创新的双向流水线并行算法。它实现了正向和反向计算-通信阶段的完全重叠，同时也减少了流水线气泡。	训练
EPLB	DAY4	在使用专家并行 (EP) 时，不同的专家被分配到不同的 GPU。由于不同专家的负载可能因当前工作负载而异，因此保持不同 GPU 的负载平衡非常重要。正如在 DeepSeek-V3 论文中所述，我们采用冗余专家策略，复制高负载的专家。然后，我们通过启发式方法将复制的专家分配到 GPU 上，以确保不同 GPU 之间的负载平衡。	训练和推理
3FS	DAY5	3FS (Fire-Flyer File System)，这是一个专为充分利用现代 SSD 和 RDMA 网络带宽而设计的并行文件系统，具备惊人的数据访问性能，为深度学习等数据密集型应用提供了强大的支持。	训练
Smallpond	DAY5	是一款构建于 DuckDB 和 3FS 之上的轻量级数据处理框架。它拥有高性能数据处理能力，可扩展至 PB 级数据集，并且操作简便，无需长期运行的服务。	训练

资料来源：智能物联技术，民生证券研究院

国产 AI 芯片与海外芯片相比仍有差距，部分场景逐步实现国产替代。 将国产 AI 芯片与海外芯片做对比，海外芯片在用于训练的浮点运算以及用于推理的整数运算上，算力水平平均高于国产芯片，部分国产芯片存在相对功耗比优势。由于推理场景的算力需求相对有限，部分场景可逐步实现国产替代。

表11：芯片对比

厂商	产品	芯片类型	制程	算力	内存类型	内存容量	内存带宽	功耗
NVIDIA	H20	GPU	4nm	FP32 44TFLOPS FP16 148TFLOPS	HBM3	96GB	4TB/s	400W
高通	Cloud AI100 ultra	ASIC	7nm	INT8 : 870 TOPS FP16 : 288 TFLOPS	LPDDR4 X-4200	128GB	548 GB/s	150W
Cerebras	WSE-3	ASIC	5nm	125 PetaFLOPS	SRAM	44GB	21PB/s	20kW
Groq	GropChip	LPU	14nm	INT8 750 TOPS FP16 188 TFLOPS	SRAM	230 MB	80TB/s	185W
华为	昇腾 910B	ASIC	7nm	640TOPS(INT8) 320 TFLOPS (FP16)	HBM	64GB	400 GB/s	310W
	昇腾 910C	ASIC	7nm	FP16 800 TFLOPS	HBM2E	128GB	3.2 TB/s	310W
寒武纪	MLU370-X8	ASIC	7nm	256 TOPS (INT8) 128 TOPS (INT16) 96TFLOPS(FP16) 96TFLOPS(BF16) 24TFLOPS(FP32)	LPDDR5	48GB	614.4GB/s	150W
隧原	云遂 i20 (邃思 2.5)	ASIC	12nm	256 TOPS (INT8) 128 TFLOPS (FP16) 128 TFLOPS(BF16) 32TFLOPS(FP32)	HBM2E	16GB	819GB/S	
壁仞	BR100	GPU	7nm	2048TOPS(INT8) 1024TFLOPS(BF16) 512TFLOPS (TF32+) 256TFLOPS (FP32)	HBM2E	64GB	2.3TB/s	550W
沐曦	曦思 N100	GPGPU	7nm	160TOPS(INT8) 80 TFLOPS(FP16)	HBM2E	16GB	460GB/s	70W
摩尔线程	MTT S4000	GPU	7nm	25 TFLOPS (FP32) 200TOPS(INT8)	GDDR6	48GB	768 GB/S	450W
云天励飞	DeepEdge 10	ASIC	14nm	64 TOPS (INT8) 32 TOPS (INT16) 8 TFLOPS (FP16)	LPDDR4	32GB	120GB/s	40W

资料来源：CSDN，China-icer 公众号，AI 芯天下，博客园，芯东西等，民生证券研究院整理

多渠道拓展芯片销售模式，未来有望迎来高速增长期。随着未来 AI 技术的进步，AI 在各行各业的普及率有望提升，与之对应的硬件市场有望迎来快速增长期。公司所销售的芯片为公司技术核心代表产品，为未来公司发展重点。公司凭借 AI 全产业链布局优势，多渠道拓展芯片销售模式。AI 芯片的使用场景主要为云侧、边缘侧、端侧。

云侧方面，公司通过智算运营业务，为有算力需求的企业提供算力支持。未来待大模型技术成熟后，此前训练大模型的企业对推理需求或大幅提升，公司的推理芯片有望凭借此前智算运营所积累的渠道优势销售公司推理芯片。边缘侧上，公司已在多场景实现 AI 智慧解决方案销售，交通领域的智慧盒子搭载公司自研芯片销

售，公司可凭借 AI 应用智慧解决方案的销售推动边缘侧推理芯片销售。端侧上，公司收购子公司岼丞技术，主要为可穿戴设备提供 PCBA 板与无线蓝牙音频解决方案，为 AI 端侧设备的重要供应商，公司可凭借子公司在端侧的渠道优势，联合设计推理硬件产品，进行推理芯片的进一步销售。

3.2.2 以芯片为核心，发展相关硬件设备拓展市场

智能硬件产品推动公司软硬一体销售。基于公司自研“云天天书”大模型和神经网络处理器，打造面向各类边缘 AI 场景的各类硬件设备并对外销售，包括 AI 芯片、天舟系列云边服务器、边缘计算盒子、“深目”AI 模盒、X6000 推理加速卡，以及基于 X6000 加速卡打造的推理服务器等。

图35：云天励飞支撑性平台



资料来源：公司公告，民生证券研究院

云天励飞计划推出 X6000 加速卡。X6000 具备 256T 算力、128GB 显存容量、486GB/s 显存带宽；采用 C2C Mesh 互联技术，可实现卡间高速互联，带宽达 64GB/s，最大可实现 64 张卡的互联；可应用于语言、视觉、多模态等各类大模型的推理加速，目前已适配云天天书、通义千问、百川智能、Llama2/3 等近 10 个主流大模型。

X6000 加速卡内置全国产算力芯片 DeepEdge200，该芯片采用 D2D Chiplet 技术，基于 DeepEdge10 芯片平台打造。X6000 单卡可实现 130B 参数量大模型推理，在执行 70B 参数量大模型推理时性能达 11 tokens/s，有望在大模型推理领域逐步实现国产替代。

图36：云天励飞 X6000 推理卡



资料来源：公司公众号，民生证券研究院

国产推理卡对比，云天励飞新款 X6000 推理卡处于第一梯队。将云天励飞、寒武纪、华为推理卡进行对比，在用于推理场景的 INT8 整数运算下，云天励飞 X6000、寒武纪 MLU290-M5、华为 Atlas 300I Duo 算力水平分别为 256Tops、512Tops、280Tops，算力水平处于第一梯队。

表12：国产推理卡对比

	云天励飞	寒武纪	寒武纪	华为	华为
推理卡名称	IPU-X6000 加速卡	MLU370-X8 智能加速卡	MLU290-M5 智能加速卡	Atlas 300I Pro 推理卡	Atlas 300I Duo
算力水平	256T	256 TOPS (INT8), 128 TOPS (INT16), 96 TFLOPS (FP16), 96 TFLOPS (BF16), 24 TFLOPS (FP32)	512 TOPS (INT8), 256 TOPS (INT16), 64 TOPS (CINT32)	140 TOPS(INT8) 70 TFLOPS(FP16)	280 TOPS(INT8) 140 TFLOPS(FP16)
显存容量	128GB	48GB	32GB	24GB	48GB/96GB
显存带宽	486GB/S	614.4 GB/s	1228 GB/s	204.8 GB/s	408GB/s
带宽	64GB/s	200GB/s	600GB/s	-	-
功耗	200W	最大热功耗 250W	最大热功耗 350W	最大 72W	150W
内存类型		LPDDR5	HBM2 高带宽内存	LPDDR4X 24 GB	LPDDR4X
制程	14nm	7nm	7nm	-	-

资料来源：各公司官网，民生证券研究院整理

一体机迎接私有化部署市场。随着 Deepseek 的开源，多数注重数据隐私性的企业为结合公司业务与 AI 技术融合，开启 AI 私有化部署，一体机作为私有化部署的主力产品，市场规模有望迎来快速发展。为迎接一体机市场的发展，公司计划推出由 X6000 推理卡组成的 X6000 mesh 推理一体机，该一体机适配 DeepSeekV3/R1 671B MoE 模型，为迎接未来企业推理需求。

图37: X6000 mesh 一体机



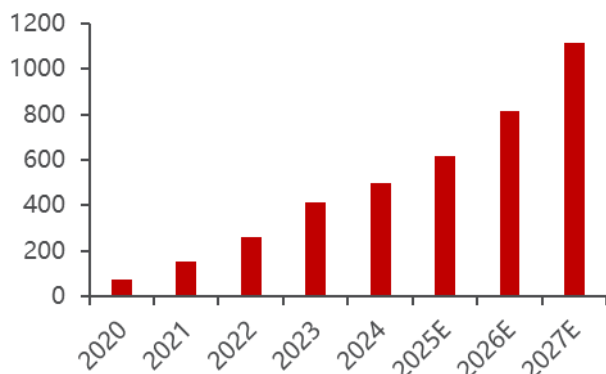
资料来源: 公司公众号, 民生证券研究院

3.2.3 依托算力储备提供算力服务与运维

中国智能算力与通用算力规模提升。基于《IDC 中国加速计算服务器半年度市场跟踪报告》及智能加速卡半精度 (FP16) 运算能力数据, 2022 年中国智能算力规模为 259.9EFLOPS, 预计到 2027 年将达到 1117.4EFLOPS, CAGR 达 33.9%。

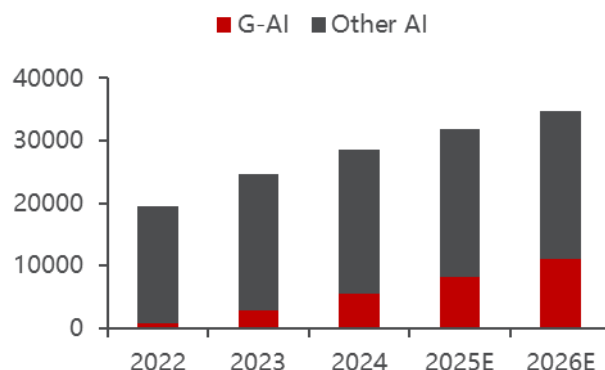
计算所需的服务器设施市场规模持续提升。IDC 预计, 全球人工智能硬件市场 (服务器), 将从 2022 年的 195 亿美元增长到 2026 年的 347 亿美元, CAGR 达 15.46%; 其中, 运行生成式人工智能的服务器市场规模在整体人工智能服务器市场的占比将从 2023 年的 11.9% 增长至 2026 年的 31.7%。随着生成式人工智能的发展, 市场对于高性能的人工智能服务器需求或将持续增长。

图38: 中国智能算力规模及预测 (EFLOPS)



资料来源: IDC, 民生证券研究院

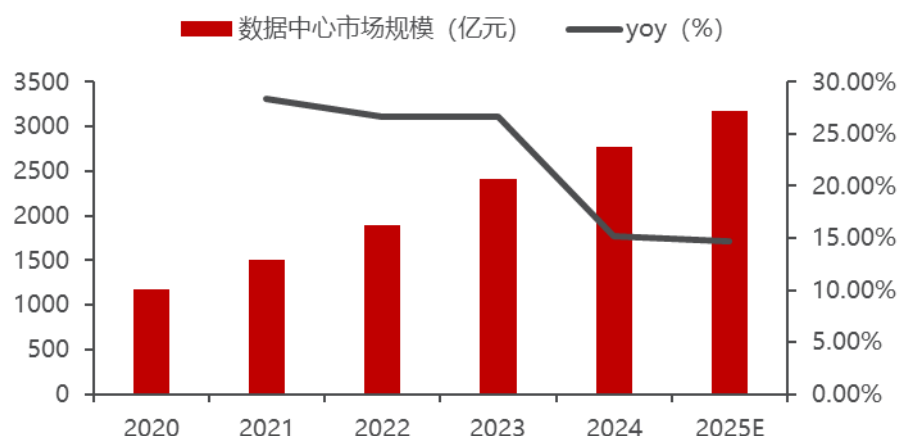
图39: 全球人工智能服务器市场规模预测 (\$M)



资料来源: IDC, 民生证券研究院

数据中心市场规模持续提升。受新基建、数字化转型等国家政策促进，我国数据中心市场规模或维持高速增长。根据中商产业研究院数据，2024 年中国数据中心市场规模约为 2773 亿元，同比增长 15.21%，预测 2025 年中国数据中心市场规模将达 3180 亿元。

图40：数据中心市场规模（亿元）&yoy（%）



资料来源：中商产业研究院，民生证券研究院

公司智算运营项目分为重资产模式与轻资产模式，稳定补充现金流，未来拓展推理芯片销售。公司于 2023 年入局 AI 基础设施运营创新业务，通过落地超大规模异构高性能算力集群，为客户提供 AI 训练集推理算力服务及对应交付物，包括智能算力调度及 AI 大模型开发配套服务。**公司智算运营项目分为重资产模式与轻资产模式，重资产模式为公司持有算力卡及相关算力基础设施，为客户提供算力服务；轻资产模式为客户持有算力卡，公司提供日常算力运维服务。**公司在同行业中资源储备较为齐全，已逐渐开拓了相关业务客户资源，形成了部分已落地项目并取得了相关收入或在手订单。**公司推理芯片为未来发展重点，公司有望凭借智算运营积累的渠道满足不同客户需求，从而凭借渠道优势在未来推理需求提升的背景下拓宽公司推理芯片渠道。**

与德元签订合同项目，大额订单补充现金流。2024 年，公司与德元方惠就德元向公司购买 AI 训练及推理算力服务达成框架安排。在正式提供服务后三年内，公司预计将获取金额约为人民币 16.07 亿元的 AI 算力服务收入。公司全资子公司励飞科技与德元方惠于 2024 年 7 月初签署合作正式合同，目前处于合同履约中，将对公司经营业绩产生积极影响。

表13：与德元签订合同细节

合作细节	主要内容
服务内容	乙方提供 AI 训练及推理算力服务及对应的交付物，包括智能算力调度及 AI 大模型开发服务平台配套服务。服务总规模约为 4000PFLOPS。 服务内容还包括组网设备及实施、算力调优、运维服务以及智能算力调度及 AI 大模型开发服务平台，不包括 IDC 托管服务，IDC 托管费用由甲方另行承担。

服务价格	(1) 在服务期限内，以月为结算周期，收取月度服务费，具体内容如下： 月度服务费：44,631,250 元（含税价，税率 6%） 使用期：36 个月
	(2) 对于发生故障的算力服务，乙方应积极进行排障、修复或更换硬件等措施，甲方应积极配合，对于故障发生后的乙方责任及月费用核减，以双方签署的补充协议约定内容为准。
	(3) 服务起始日以验收通过之日为准。
服务期开始日	不晚于 2024 年 10 月 1 日（若乙方交付日期延迟，则服务期开始日相应顺延）。
付款	甲方应按本合同的约定按月向乙方支付月度服务费，和续期月度服务费（如涉及续期）、一次性服务费（如涉及且双方达成一致）。乙方应于每个 结算周期结束后，向甲方出具结算单明细。甲方应于结算单出具后的 7 日内予以确认或提出异议，乙方对异议点进行复核后重新输出结算单明细。

资料来源：公司公告，民生证券研究院

公司子公司成都云天励飞中标天府智算中心项目采购，轻资产模式算力租赁亦增加营收。公司子公司中标天府智算中心项目设备采购，中标项目金额为 1.30 亿元，该项目为轻资产模式，公司交付智算整体方案，而不持有算力卡等硬件设施。与重资产模式相比，轻资产模式毛利率偏低，但无折旧费用成本，亦可补充营收现金流。

3.3 行业级应用场景：智慧解决方案经验成熟，多行业拓展拓宽渠道

行业级业务，凭借软硬件通用平台，在六大行业开展智慧行业解决方案。公司行业级业务，面向政府机构等终端客户、企事业单位等集成商客户，基于自研的算法软件、芯片等核心产品，搭配外购的定制化或标准化硬件产品，按需求交付硬件产品或解决方案。公司所涉及的细分行业为六类，分别为智慧警务、城市治理、智慧交通、人居生活、低空经济、智慧教育。**业务规划：未来续优化现有城市治理、公共安全、商业园区、智慧交通等成熟领域解决方案的同时，积极布局智慧教育、低空经济等新兴领域，实现应用场景延展。**

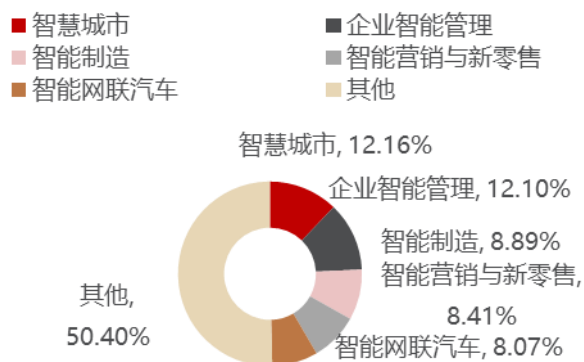
图41：公司行业级业务



资料来源：公司公告，民生证券研究院

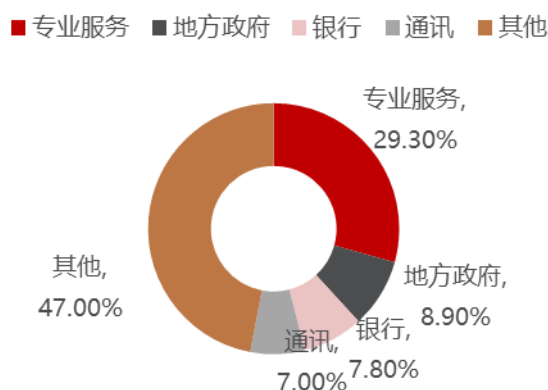
AI 应用领域智慧城市与企业管理需求占比高，专业服务与地方政府占据需求主导。未来随着 AI 大模型技术的进步，AI 对各行各业的影响逐步扩大，对应的 AI 应用市场有望迎来快速扩容期，细分市场对应的行业及客户或成为 AI 应用企业重点布局领域。据新一代人工智能发展战略研究所数据，2024 年中国人工智能行业下游中智慧城市、企业智能管理、智能制造为占比前三细分行业，分别为 12.16%、12.10%、8.89%。据 IDC 预测，2026 年中国人工智能市场支出预测（行业）前三分别为专业服务、地方政府、银行，占比分别为 29.3%、8.9%、7.8%。AI 行业应用逐步落地过程中，精准定位细分市场与客户的企业有望因此受益。

图42：2024 年中国人工智能行业下游应用领域占比 (%)



资料来源：新一代人工智能发展战略研究所、中商产业研究院、民生证券研究院

图43：2026 年中国人工智能市场支出预测 (%)



资料来源：IDC、民生证券研究院

公司智慧行业解决方案主要分为 4+2 应用场景分类,主要分类为**公安、城治、交通、人居、低空、教育**。每个人和家庭都有 N 个 AI 助手。

智慧公安，从被动警务到预防警务。公司利用 AI 技术结合视频巡逻，打造面向全场景主动警务和预防警务的“新引擎”，包含“一平台、一机制、N 场景”。

一平台实现导防导稳导侦，包含数据驾驶舱、大模型分析报告、算力部署建议、勤务考核管理。一机制实现事件快速处理，主城区 1 分钟快速反应、重点区域 3 分钟现场处置、人群密集场所 5 分钟现场处置，事后进行人工复审。N 场景实时感知异常事件，在治安置点、涉政涉恐、专项勤务、警情高发区域、灾害易发区域、涉警预警场景进行部署，实现从被动警务到预防警务。

图44：智慧行业解决方案-智慧警务



资料来源：公司宣传材料，民生证券研究院

城市治理，共建共治共享的“以人为本的城市”。城治场景下，采取“发现、分拨、处置、复核、归档”程序，高效处置一件事。从而实现日均发现事件 1000+，识别准确率>95%，事件闭环率>97%,事件闭环周期从 8-10 小时缩减至 3 小时内。

图45：智慧行业解决方案-城治



资料来源：公司宣传材料，民生证券研究院

智慧交通，“空天地一体”的智能交通网络。其中公交 OD 智能分析，落地巴士集团公交 5800+辆，提供 OD 采集终端 10000+台，车辆周转率提升 10%，大站快车线市民出行时间下降 20%，区间快车线营收提升 20%。城市 CT 巡检，道路病害检出率高于 90%，病害检出准确率大于 95%，新算法上线周期小于 2 周，病害算法检测种类大于 20 种，巡检人力需求下降 50%，端云协同方案节省带宽 80%+。

图46：智慧行业解决方案-交通



资料来源：公司宣传材料，民生证券研究院

低空经济，赋能百业，带来产业的转型、融合。公司边缘 AI 赋能应用，实现飞行监测、自主导航、自动避障，为低空下的物流、文旅、载人、城市服务赋能。通过低空行业布局，公司收集采取城市低空数据源，包含地面基建数据、低空无人机数据、低空视频图像及分析数据、行业业务数据，进一步完善 AI 产业与应用场景结合。

图47：智慧行业解决方案-低空



资料来源：公司宣传材料，民生证券研究院

智慧教育，未来 AI 时代世界变革的引领者。凭借在 AI 领域近 10 年的深厚积累，公司提出了 Fly-X 未来 AI 人才培养四步计划，致力于构建一个全面、立体的人工智能教育生态系统，为学生、教师和学校带来颠覆式的教学体验。

图48：智慧行业解决方案-教育



资料来源：公司宣传材料，民生证券研究院

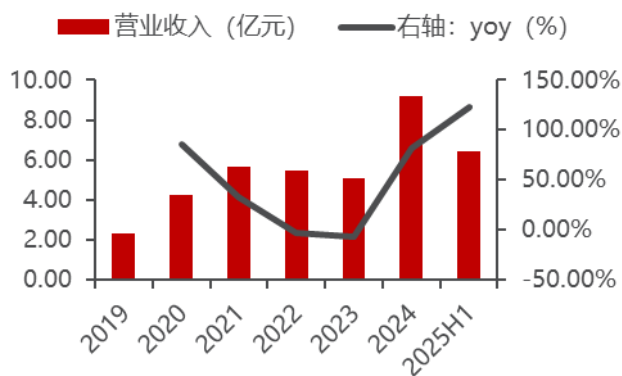
4 营收维持高增，利润端持续减亏

4.1 营收进入高速增长期，利润端大幅减亏

营业收入增速恢复高速增长。2019-2024 年，公司营业收入从 2.30 亿元增长至 9.18 亿元，CAGR 达 31.89%。其中 2022、2023 年营收出现下滑，2022 年主要系部分项目推进落地不及预期，叠加业务受公共卫生事件影响；2023 年主要系数字城市业务收入下降所致。公司 2024 年收入为 9.18 亿元，同比增长 81.42%，营业收入恢复快速增长态势，主要系企业级及消费级的销售收入增加所致。未来公司凭借算法+芯片技术拓展消费级与企业级业务增长，营业收入有望维持高速增长。公司 2025 年 H1 营收为 6.46 亿元，同比+123.10%；半年度高速增长系企业级业务与消费级业务兑现。

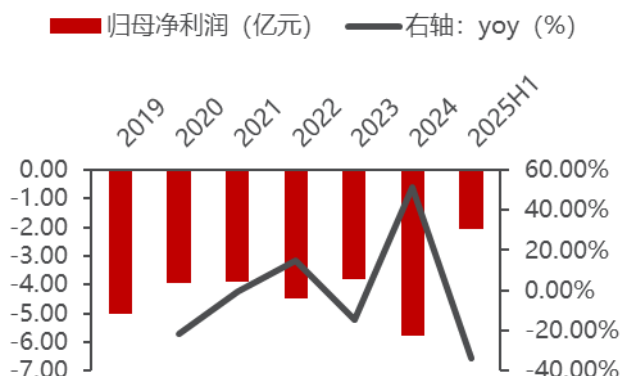
利润端受股份支付额度及研发投入较高承压，利润端尚未盈利。2019-2024，公司归母净利润分别为-5.00/-3.93/-3.90/-4.47/-3.83/-5.79 亿元。公司所处行业属于人工智能新兴产业，前期需要投入大量研发资金，公司同时需支付股权激励费用，从而导致了利润端短期内仍呈现压力。未来随着公司对低毛利率项目的把控，公司有望持续减亏。公司 2025H1 归母净亏损为 2.06 亿元，较上年同期减亏 1.04 亿元，公司盈利能力提升显著。

图49：公司营业收入（亿元）&yoy（%）



资料来源：公司公告，民生证券研究院

图50：公司归母净利润（亿元）&yoy（%）



资料来源：公司公告，民生证券研究院

公司从上市初的行业解决方案为主转换为智能硬件、智算运营、智能应用三方向发展为主。公司上市初，主要业务以 AI 应用服务为主，公司主要披露的营业收入主要以人居生活智慧化升级 AI 产品及整体解决方案行业和数字城市运营管理 AI 产品及整体解决方案行业为主，2018 年-2023 年，两项业务分别从 1.22、0.12 亿元增长至 3.92、0.85 亿元，CAGR 分别为 126.29%、147.93%，行业呈现高增速。数字城市业务在 2023 年受经济环境影响下滑，人居生活自 2022 年受公共卫生事件影响出现下滑。

进入 2024 年以来，公司积极拓展企业级与消费级 AI 项目与产品，叠加并表峴丞技术，公司将营业收入重新进行分类，并分类为企业级、消费级、行业级三类

业务进行披露。三项业务 2025 年上半年营业收入分别为 2.69/2.98/0.76 亿元，占比分别为 41.84%/46.35%/11.82%。

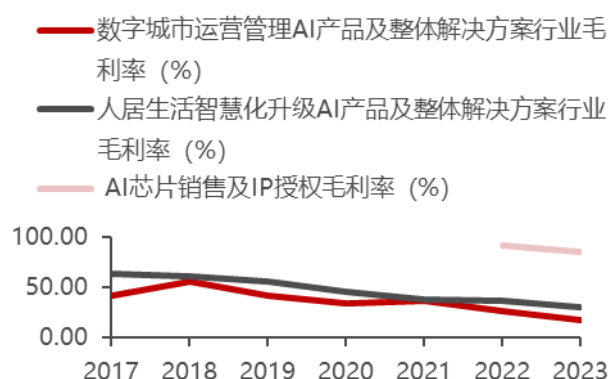
利润端承担压力，不同业务毛利率呈下降趋势。公司的数字城市业务在 2018 年达到毛利率高点，为 55.93%；人居生活业务在 2017 年达到毛利率高点，为 63.16%。后续两项业务毛利率逐步呈现下降趋势，系公司为保持营收规模高速扩张，低毛利项目增多所导致。我们认为未来随着公司转变整体市场策略，稳定维持高毛利项目增长，利润端有望持续减亏。

图51：2017-2023 年公司分业务营收（亿元）



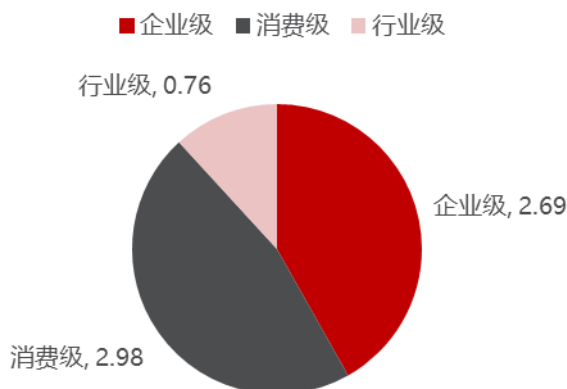
资料来源：公司公告，民生证券研究院

图52：2017-2023 年公司分业务毛利率（%）



资料来源：公司公告，民生证券研究院

图53：2025H1 公司按新行业模式收入分类（亿元）

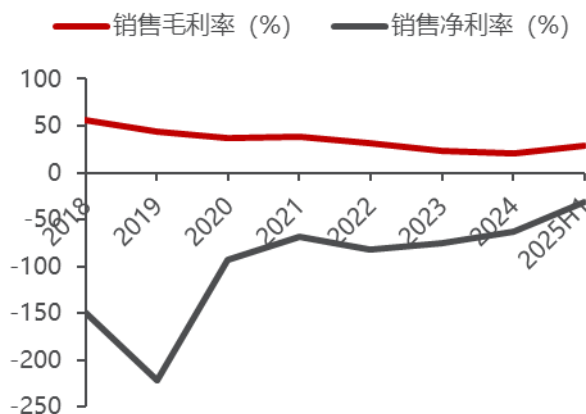


资料来源：公司公告，民生证券研究院

公司利润端承压，三费+研发费用率占比出现下降。公司毛利率在 2018 年达到高点，为 56.34%，后续呈现下降趋势。2025H1，毛利率为 28.50%，毛利率下降系公司新增业务中硬件比例提高，硬件产品相较软件产品毛利率偏低。利润端由于公司仍处于高研发投入阶段，公司尚未盈利，2025H1，公司净利率为-31.51%，

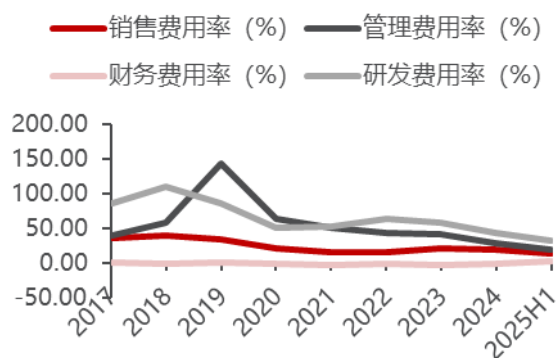
净利端亏损幅度大幅减少。公司费用率近年来维持相对稳定，2025H1，公司销售费用率、管理费用率、财务费用率、研发费用率分别为13.41%/19.79%/1.99%/31.36%，较2024H1，费用率分别同比变动-15.23pct/-15.74pct/+3.74pct/-31.18pct，整体费用率出现下降，公司因营收规模扩大带来规模效应，费用率呈现下降趋势。

图54：公司毛利率 (%) & 净利率 (%)



资料来源：公司公告，民生证券研究院

图55：公司期间费用率 (%)



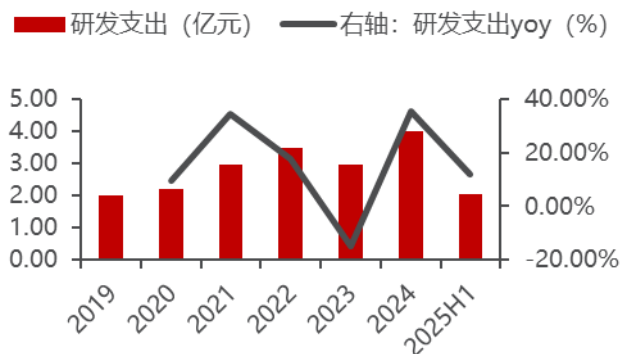
资料来源：公司公告，民生证券研究院

4.2 重研发投入，人均创收恢复高增速

重视研发与人才储备，人均创收恢复高增速。公司研发投入恢复快速增长，2019-2024年，公司研发投入从2.00亿元增长至4.00亿元，CAGR达14.87%。2023年受整体经济环境下行影响，研发投入出现下滑。2025H1，公司研发支出为2.03亿元，同比增长12.15%，研发投入持续增长。公司员工人数恢复增长态势，公司员工人数从2019年的752人增长至2024年的920人，其中2022、2023年人数出现下降，2024年重新恢复快速增长态势，公司整体呈现扩张趋势。

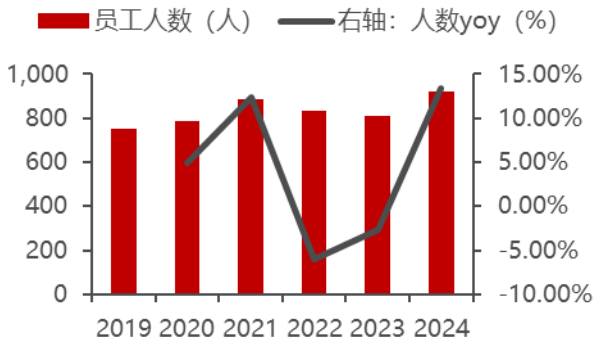
公司人均创收重新恢复增长趋势。我们将公司营收与公司总人数进行对比，可以发现公司人均创收在2023年出现小幅下降，2024年，人均创收重新恢复增长趋势，同比增长达59.93%，公司人效进一步提升。

图56：研发支出 (亿元) & 右轴：研发支出 yoy (%)



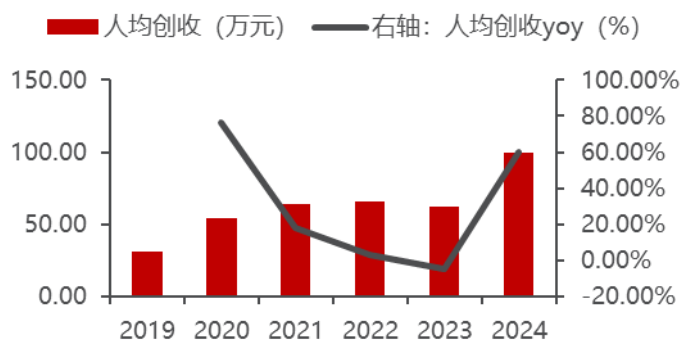
资料来源：公司公告，民生证券研究院

图57：员工人数 (人) & 右轴：人数 yoy (%)



资料来源：公司公告，民生证券研究院

图58：人均创收（万元）&yoy（%）



资料来源：公司公告，民生证券研究院

5 盈利预测与投资建议

5.1 盈利预测与业务拆分

由于公司于 2024 年调整业务营收分类，将公司业务主要分为消费级、企业级、行业级，我们按照公司最新业务分类进行盈利预测。

消费级：1) 收入：公司消费级产品主要为多种 AI 端侧设备，包括可穿戴设备类产品、AI 交互类产品等。随着大模型技术的进步，端侧设备的实效性有望提升，端侧设备消费端接受度提升，端侧设备市场规模有望迎来高速扩容期，公司相关产品有望迎来快速增长。受子公司岢丞技术收购协议营收目标与 2025 年公司加大子公司品牌噜卡博士宣发原因，相关业务增速呈现先升后降趋势。**因此我们预计公司消费级营业收入 2025-2027 年增速分别为 16%、26%、21%。**2) **毛利率：**考虑到公司端侧设备销售业务模式稳定，预计毛利率会延续历史水平。**预计 2025-2027 年毛利率分别为 12%、12%、12%。**

企业级：1) 收入：公司企业级产品以推理芯片为核心，拓展系列产品，同时凭借算力储备提供算力支持与运维服务。公司该项业务营收为算力服务业务营收叠加以芯片为核心的相关的推理卡、边缘计算盒子、一体机等产品。公司算力服务业务营收规模较大，并于 2024 年下半年计入收入，2025 年全年计入收入，因此预计 2025 年企业级业务增速提升较快，后续增速放缓。**我们预计公司企业级营业收入 2025-2027 年增速分别为 120%、31%、31%。**2) **毛利率：**考虑到公司核心芯片产品加速卡推出，加速卡为高毛利率产品，相关产品起量有望带动毛利率提升，**预计 2025-2027 年毛利率分别为 40%、42%、44%。**

行业级：1) 收入：公司行业级业务为凭借软硬件通用平台，在六大行业开展智慧行业解决方案。公司此项业务主要在智慧警务、城市治理、智慧交通、人居生活、低空经济、智慧教育六大领域，向政府机构等终端客户、企事业单位等集成商客户，按需求交付硬件产品或解决方案。公司此项业务未来规划维持相对稳定，**因此我们预计行业级营业收入 2025-2027 年增速分别为 2%、2%、2%**2) **毛利率：**考虑到公司该业务营收模式相对稳定，因此预计毛利率会延续历史水平，**预计 2025-2027 年毛利率分别为 15%/15%/15%。**

表14：公司营收测算汇总（百万元）

	2024A	2025E	2026E	2027E
企业级（百万元）	249	546	716	936
yoy (%)		119%	31%	31%
毛利率 (%)	40%	40%	42%	44%
消费级（百万元）	406	471	595	722
yoy (%)		16%	26%	21%
毛利率 (%)	11%	12%	12%	12%
行业级（百万元）	254	259	264	269
yoy (%)		2%	2%	2%

毛利率 (%)	15%	15%	15%	15%
其他	9	9	9	9
yoy (%)		0%	0%	0%
毛利率 (%)	82%	70%	70%	70%
总营收 (百万元)	917	1285	1584	1936
yoy (%)		40%	23%	22%
毛利润 (百万元)	192	321	419	546
毛利率 (%)	21%	25%	26%	28%

资料来源：Wind，民生证券研究院预测

主要费用率：

1) 销售费用率：随着公司营收规模扩张叠加公司企业级、消费级、行业级三方业务销售渠道建设稳定，销售团队已组件完毕，后期公司销售费用率有望合理收敛，**预计 2025-2027 年销售费用率分别为 14%、12%、10%；**

2) 管理费用率：公司前期管理费用率较高涉及上市费用及股权激励费用，未来营收规模扩张及无相关费用影响，管理费用率有望合理收敛，**预计 2025-2027 年管理费用率分别为 18%、15%、12%。**

3) 研发费用率：公司作为 AI 芯片和算法企业，研发投入需求较大。营收规模虽大幅增长，但公司全力转向推理市场，大量研发费用投入 AI 推理芯片研发，公司研发费用或呈现缓步下降，仍维持较高水平。**预计 2025-2027 年研发费用率分别为 32%、28%、23%。**

表15：公司主要费用率预测

	2023	2024	2025E	2026E	2027E
销售费用率	20%	19%	14%	12%	10%
管理费用率	41%	28%	18%	15%	12%
研发费用率	58%	44%	32%	28%	23%

资料来源：Wind，民生证券研究院预测

5.2 估值分析

公司作为人工智能推理侧的领先企业，凭借自身软件算法与硬件设计等优势，从企业级、消费级、行业级三方渠道拓展公司产品销售。公司核心主业为推理芯片设计、并以推理芯片为核心生产销售推理卡、一体机、端侧设备等，在销售硬件设备的同时，也凭借算法优势销售软硬一体的行业解决方案。我们选取与公司同样销售 AI 芯片的芯片设计公司寒武纪、海光信息以及主业为 AI 行业解决方案的公司虹软科技、云从科技-UW 作为可比公司。可比公司 2025-2027 年平均 PS 分别为 46X、31X、22X，公司对应的 PS 分别为 23X、19X、15X，公司 2025-2027 年 PS 均低于可比公司均值水平。

表16：可比公司盈利预测与估值比较

股票代码	公司简称	股价（元/股）	营收（亿元）			PS（倍）		
			2025E	2026E	2027E	2025E	2026E	2027E
688041.SH	海光信息	255.84	141.11	199.42	264.86	42	30	22
688088.SH	虹软科技	57.20	10.12	12.73	15.89	23	18	14
688327.SH	云从科技-UW	16.94	4.84	6.34	8.68	36	28	20
688256.SH	寒武纪-U	1,323.50	67.21	118.43	178.03	82	47	31
	平均值					46	31	22
688343.SH	云天励飞-U	82.77	12.85	15.84	19.36	23	19	15

资料来源：Wind，民生证券研究院预测；

（注：可比公司数据采用 Wind 一致预期，股价时间为 2025 年 9 月 29 日）

5.3 投资建议

公司部署 AI 算力到应用全产业链，多项平台技术优势明显。此外，公司多方向开拓 AI 市场，通过企业级、消费级、行业级三方向渠道开拓，销售额增长显著。后续随着自研芯片及相关产品在云端推理市场，公司营收有望迎来快速增长。我们预计公司 2025-2027 年营业收入为 12.85、15.84、19.36 亿元，对应 PS 分别为 23X、19X、15X，首次覆盖，给予“推荐”评级。

6 风险提示

1) 核心芯片产品销售不及预期。公司的核心增长业务为芯片设计，该行业核心产品迭代快、市场竞争激烈，公司最新芯片及其相关产品如销售不及预期，或影响公司销售额变动。

2) 技术研发不及预期。公司所研发芯片为 NPU 芯片，该芯片主要服务于人工智能行业推理端，行业需求变化快。为面对需求端的快速变化，公司已公布未来几年芯片设计规划。如公司芯片研发进度不及预期，或影响公司核心竞争力。

3) 端侧设备产品销售不及预期风险。公司消费级业务主要为销售带有 AI 功能的端侧设备，其中包含拍学机、AI 眼镜、AI 玩具等。如端侧设备销售不及预期，或影响公司消费级业务营收。

公司财务报表数据预测汇总

利润表 (百万元)	2024A	2025E	2026E	2027E
营业总收入	917	1,285	1,584	1,936
营业成本	725	964	1,165	1,390
营业税金及附加	5	8	10	12
销售费用	176	180	190	194
管理费用	255	229	235	229
研发费用	400	411	444	445
EBIT	-620	-434	-366	-216
财务费用	-7	15	37	58
资产减值损失	-12	-10	-12	-14
投资收益	34	19	24	29
营业利润	-579	-439	-391	-259
营业外收支	3	0	0	0
利润总额	-576	-439	-391	-259
所得税	-4	-1	-1	-1
净利润	-572	-438	-390	-259
归属于母公司净利润	-579	-438	-390	-259
EBITDA	-468	-232	-76	137

资产负债表 (百万元)	2024A	2025E	2026E	2027E
货币资金	927	1,441	1,483	1,693
应收账款及票据	731	868	941	1,047
预付款项	44	29	35	42
存货	249	231	280	334
其他流动资产	1,034	1,013	1,016	1,021
流动资产合计	2,984	3,582	3,756	4,136
长期股权投资	17	17	17	17
固定资产	1,558	1,758	1,908	2,008
无形资产	35	35	35	35
非流动资产合计	1,934	2,098	2,249	2,349
资产合计	4,918	5,679	6,005	6,485
短期借款	66	66	66	66
应付账款及票据	465	641	775	925
其他流动负债	256	800	1,081	1,370
流动负债合计	787	1,508	1,923	2,361
长期借款	86	586	886	1,186
其他长期负债	99	49	50	50
非流动负债合计	185	635	936	1,236
负债合计	973	2,143	2,859	3,598
股本	355	359	359	359
少数股东权益	9	9	9	9
股东权益合计	3,946	3,536	3,146	2,888
负债和股东权益合计	4,918	5,679	6,005	6,485

资料来源：公司公告、民生证券研究院预测

主要财务指标	2024A	2025E	2026E	2027E
成长能力 (%)				
营业收入增长率	81.30	40.05	23.29	22.24
EBIT 增长率	-41.52	29.96	15.79	40.88
净利润增长率	-51.14	24.28	11.04	33.68
盈利能力 (%)				
毛利率	20.94	24.97	26.44	28.20
净利率	-63.12	-34.13	-24.63	-13.36
总资产收益率 ROA	-11.77	-7.72	-6.50	-3.99
净资产收益率 ROE	-14.71	-12.43	-12.43	-8.99
偿债能力				
流动比率	3.79	2.38	1.95	1.75
速动比率	3.06	2.02	1.65	1.47
现金比率	1.18	0.96	0.77	0.72
资产负债率 (%)	19.77	37.73	47.61	55.47
经营效率				
应收账款周转天数	238.99	222.87	204.67	183.91
存货周转天数	95.63	89.77	78.94	79.41
总资产周转率	0.19	0.24	0.27	0.31
每股指标 (元)				
每股收益	-1.61	-1.22	-1.09	-0.72
每股净资产	10.97	9.83	8.74	8.02
每股经营现金流	-0.82	0.14	0.04	0.58
每股股利	0.00	0.00	0.00	0.00
估值分析				
PE	/	/	/	/
PB	7.5	8.4	9.5	10.3
EV/EBITDA	/	/	/	215.50
股息收益率 (%)	0.00	0.00	0.00	0.00

现金流量表 (百万元)	2024A	2025E	2026E	2027E
净利润	-572	-438	-390	-259
折旧和摊销	152	202	290	353
营运资金变动	87	244	42	22
经营活动现金流	-295	52	13	209
资本开支	-1,563	-385	-423	-435
投资	-147	0	0	0
投资活动现金流	601	-325	-399	-406
股权募资	0	0	0	0
债务募资	138	850	500	500
筹资活动现金流	79	787	428	407
现金净流量	385	514	43	210

插图目录

图 1: 公司产品出货形态及面向的终端场景	3
图 2: 公司发展历程	4
图 3: 云天励飞股权结构	4
图 4: AI 产业链	7
图 5: 中国人工智能产业市场规模 (亿元) &右轴: yoy (%) &右轴: 中国/全球占比 (%)	7
图 6: 中国边缘计算市场规模 (亿元) &yoy (%)	8
图 7: 2023 年中国边缘计算市场结构	8
图 8: 中国端侧 AI 行业市场规模 (亿元) &yoy (%)	9
图 9: 端侧 AI 应用场景分类	9
图 10: 算法芯片化基础技术	10
图 11: 公司指令集进步路线	10
图 12: 公司多场景多层级销售产品	13
图 13: 公司消费级业务	14
图 14: 岼丞技术销售支持产品	15
图 15: 岼丞技术硬件设计能力	15
图 16: 岼丞技术收购协议营收目标 (万元) &yoy (%)	15
图 17: 噜味博士拍学机	16
图 18: AI 拍学机 618 销售情况	16
图 19: 大模型赋能闪极 AI 眼镜	17
图 20: 公司企业级产品及服务	17
图 21: 中国 AI 芯片市场规模 (亿元) &yoy (%)	18
图 22: AI 推理芯片主要分类及部署地点	20
图 23: 中国 AI 推理芯片相关产品及服务行业市场规模 (十亿元) &占比 (%)	20
图 24: 公司 AI 推理芯片发展规划	21
图 25: DeepEdge 推理芯片系列	21
图 26: 云天励飞 DeepEdge10 系列	21
图 27: Edge10Series 芯片平台支持大模型	22
图 28: DeepEdge10Max 架构图	23
图 29: DeepEdge10 架构图	23
图 30: C2C Mesh Torus 互联扩展	24
图 31: D2D Chiplet 先进封装架构	24
图 32: D2D 和 C2C 推理架构支撑推理算力积木	24
图 33: 云天励飞 AI 推理芯片迭代规划	25
图 34: DeepEdge10 已完成 DeepSeek R1 系列模型适配	25
图 35: 云天励飞支撑性平台	28
图 36: 云天励飞 X6000 推理卡	29
图 37: X6000 mesh 一体机	30
图 38: 中国智能算力规模及预测 (EFLOPS)	30
图 39: 全球人工智能服务器市场规模预测 (\$M)	30
图 40: 数据中心市场规模 (亿元) &yoy (%)	31
图 41: 公司行业级业务	33
图 42: 2024 年中国人工智能行业下游应用领域占比 (%)	33
图 43: 2026 年中国人工智能市场支出预测 (%)	33
图 44: 智慧行业解决方案-智慧警务	34
图 45: 智慧行业解决方案-城治	35
图 46: 智慧行业解决方案-交通	35
图 47: 智慧行业解决方案-低空	36
图 48: 智慧行业解决方案-教育	36
图 49: 公司营业收入 (亿元) &yoy (%)	37
图 50: 公司归母净利润 (亿元) &yoy (%)	37
图 51: 2017-2023 年公司分业务营收 (亿元)	38
图 52: 2017-2023 年公司分业务毛利率 (%)	38
图 53: 2025H1 公司按新行业模式收入分类 (亿元)	38
图 54: 公司毛利率 (%) &净利率 (%)	39

图 55: 公司期间费用率 (%)	39
图 56: 研发支出 (亿元) & 右轴: 研发支出 yoy (%)	39
图 57: 员工人数 (人) & 右轴: 人数 yoy (%)	39
图 58: 人均创收 (万元) & yoy (%)	40

表格目录

盈利预测与财务指标	1
表 1: 公司管理层结构	5
表 2: 2023 年公司股权激励情况	5
表 3: 募集资金项目	6
表 4: 边缘计算行业产品结构	8
表 5: DeepseekR1 技术报告总结	11
表 6: DeepSeek-R1: 冷启动与多阶段优化	11
表 7: 算力芯片应用场景分类	18
表 8: AI 芯片分类	19
表 9: 云天励飞系列芯片	22
表 10: Deepseek 开源日开放项目	26
表 11: 芯片对比	27
表 12: 国产推理卡对比	29
表 13: 与德元签订合同细节	31
表 14: 公司营收测算汇总 (百万元)	41
表 15: 公司主要费用率预测	42
表 16: 可比公司盈利预测与估值比较	43
公司财务报表数据预测汇总	45

分析师承诺

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登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

评级说明

投资建议评级标准	评级	说明
以报告发布日后的 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。	推荐	相对基准指数涨幅 15%以上
	谨慎推荐	相对基准指数涨幅 5% ~ 15%之间
	中性	相对基准指数涨幅-5% ~ 5%之间
	回避	相对基准指数跌幅 5%以上
	推荐	相对基准指数涨幅 5%以上
	中性	相对基准指数涨幅-5% ~ 5%之间
	回避	相对基准指数跌幅 5%以上

免责声明

民生证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。

本报告仅供本公司境内客户使用。本公司不会因接收人收到本报告而视其为客户。本报告仅为参考之用，并不构成对客户的投资建议，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑获取本报告的机构及个人的具体投资目的、财务状况、特殊状况、目标或需要，客户应当充分考虑自身特定状况，进行独立评估，并应同时考量自身的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见，不应单纯依靠本报告所载的内容而取代自身的独立判断。在任何情况下，本公司不对任何人因使用本报告中的任何内容而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。

本报告的版权仅归本公司所有，未经书面许可，任何机构或个人不得以任何形式、任何目的进行翻版、转载、发表、篡改或引用。所有在本报告中使用的商标、服务标识及标记，除非另有说明，均为本公司的商标、服务标识及标记。本公司版权所有并保留一切权利。

民生证券研究院：

上海：上海市虹口区杨树浦路 188 号星立方大厦 7 层； 200082

北京：北京市东城区建国门内大街 28 号民生金融中心 A 座 18 层； 100005

深圳：深圳市福田区中心四路 1 号嘉里建设广场 1 座 10 层 01 室； 518048