证券研究报告/投资策略

报告日期：2025 年 03 月 21 日

推理算力需求持续增长，ASIC 端侧应用前景广阔

半导体行业研究

投资要点

➢ 推理算力需求扩容

近年来大模型持续迭代，大模型参数规模总体呈现增加趋势，参数增加带动算力需求扩容。ChatGPT 3.5加速了生成式人工智能的商业化进程，实现注册用户数量破亿仅耗时两个月，微软、谷歌等科技巨头纷纷接入，之后大模型热度持续火爆，带动算力需求激增。DeepseekR1问世推动大模型平价化，降低了大模型开发成本，利于为下游端侧和应用侧打开市场空间，下游爆发同样将催生大量算力需求，并推动算力需求由训练端向推理端转移。据IDC预测，推理服务器的工作负载占比预计由2020年的51.5%逐年增加至2026年的62.2%，中国人工智能服务器工作负载结构中的推理算力占比总体呈现增加趋势。

➢ Deepseek推动大模型平价化，端侧，应用侧商业化进程有望提速

通过一系列算法优化，Deepseek-V3相较于同类模型，训练成本大幅下降，完成训练仅耗时不到两个月，按H800芯片算力测算， Deepseek-V3预训练阶段的训练时长为266,4万GPU小时，上下文扩展训练耗时11.9万GPU小时，后训练阶段耗时5,000 GPU小时，假设H800每小时的租赁价格为2美元，则模型的总训练成本为557.6万美元，训练成本仅为GPT-4o的十分之一。

➢ ASIC适于端侧部署，市场空间广阔

本地推理不仅可以降低延时、提高吞吐量，摆脱网络限制，还有助于保护数据安全和用户隐私，终端推理任务的本地化运行或是未来的发展趋势，本地推理需求的增加将促进ASIC市场需求扩容。

ASIC芯片专门用来优化神经网络推理或者矩阵运算任务，专注于特定用途或特定模型，相较GPU在功耗、可靠性、性能、成本等方面具备优势，因此更适于在端侧和用户侧部署，如智驾、AI眼镜、智能家居等。随着大模型平价化，预期AI产品将在更多应用场景下实现商业落地，ASIC芯片具备广阔的市场前景。

➢ 投资建议

建议关注产品矩阵丰富，下游应用领域覆盖全面的芯原股份和寒武纪。

➢ 风险提示

建议关注技术迭代风险、下游需求不及预期的风险和中美贸易摩擦加剧的风险。

分析师：吴起涤

执业登记编号：A0190523020001

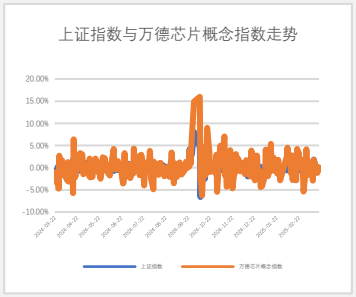
wuqidi@yd.com.cn

分析师：赵毅轩

执业登记编号：A0190124060001

zhaoyixuan@yd.com.cn

上证指数与万德芯片概念指数走势



资料来源：Wind，源达信息证券研究所

目录

一、应用场景有别，性能各有侧重
二、推理端算力需求扩容.....4
1.参数数量总体呈现增加趋势.....4
2.大模型火热，用户量激增.....6
3. Deepseek 推动大模型平价化，利好端侧、应用侧爆发.....6
三、终端定制化特点突出，看好 ASIC 芯片发展前景.....7
四、投资建议.....11
1、芯原股份.....11
2、寒武纪.....11
五、风险提示.....12

图表目录

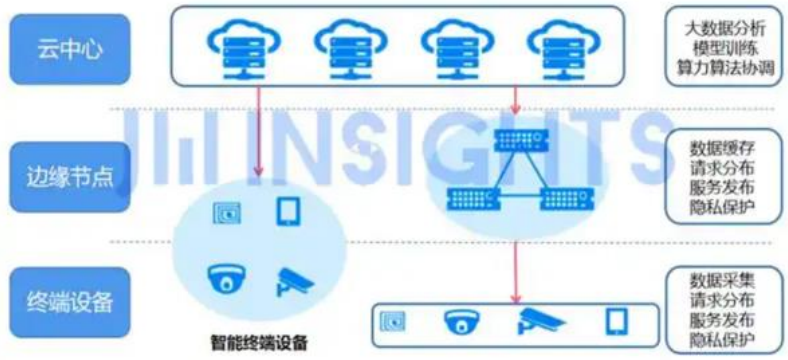
图 1：云端部署、边缘部署、终端部署.....3
图 2：训练与推理环节的性能需求不同.....3
图 3：中国人工智能服务器工作负载预测，2020-2026.....4
图 4：参数量与大模型性能.....5
图 5：增长 1 亿用户花费时间.....6
图 7：推理模型输入输出价格（元/1M Tokens）.....7
图 8：GeForce RTX 50.....8
图 9：ASIC 芯片性能优势.....8
图 10：2021-2025E ASIC 全球市场规模（亿美元）.....10
图 11：2020-2024 营业总收入（亿元）.....11
图 12：2020-2024 扣非归母净利润（亿元）.....11
图 13：2020-2024 营业总收入（亿元）.....12
图 14：2020-2024 扣非归母净利润（亿元）.....12

表 1：不同模型参数规模.....5
表 2：Deepseek-V3 模型训练成本.....7
表 3：英伟达主流产品能耗（W）.....9

一、应用场景有别，性能各有侧重

为应对不同应用场景下的使用需求，芯片可以在云端、边缘或是终端进行部署。大模型训练需要大量算力资源，一般在云端利用大规模算力集群进行训练，但随着大模型提供的服务由文本向图片、视频等多模态扩展，使用人数不断增加，云端推理服务对算力的需求也在不断提升。另一种芯片部署方式为边缘部署，它允许在生成数据的设备附近进行计算，而不是在集中的云计算设施或远程数据中心进行计算。这种本地化处理方式使得设备能够在几毫秒内做出决策，而无需互联网连接或云服务的辅助。随着 AI 眼镜、手机、音箱等端侧需求的增长，及时人机交互、数据的实时采集、低时延等需求逐渐凸显，适应轻量化设备的终端部署迎来更大的发展机会。

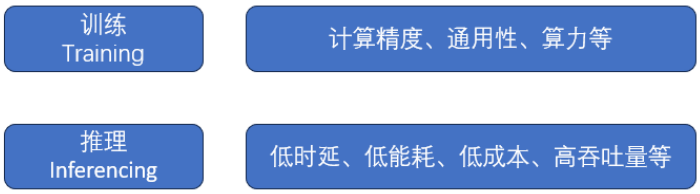
图 1：云端部署、边缘部署、终端部署



资料来源：前瞻产业研究院

训练需要大量地向模型输入训练数据，推理结果，还要调整模型参数和偏置值，如此往复直到模型收敛满足性能要求为止。而推理仅需要向模型输入非训练数据让模型计算出结果即可，推理和训练在工作中有重合的部分，推理可简单理解为简化版的训练过程。训练芯片更关注计算精度、算力等性能指标，而推理芯片更加看重低时延、低能耗、低成本、高吞吐量等指标。

图 2：训练与推理环节的性能需求不同

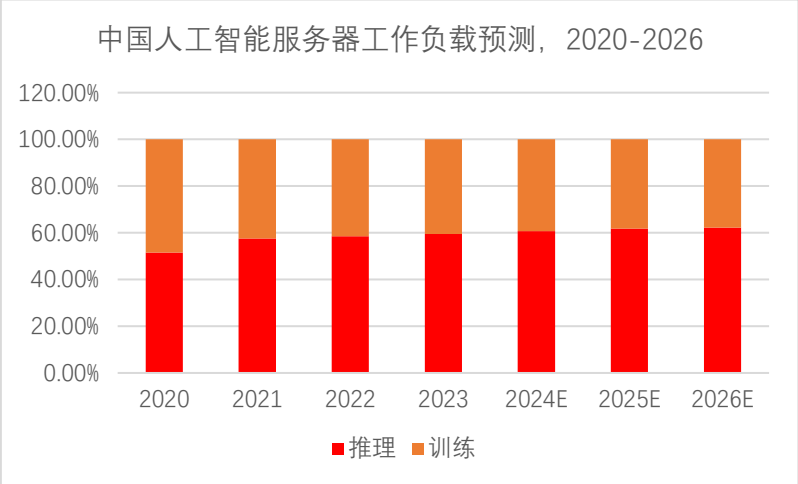


资料来源：源达信息证券研究所

二、推理端算力需求扩容

近年来大模型持续迭代，大模型参数规模总体呈现增加趋势，参数增加带动算力需求扩容。另外，ChatGPT 3.5 的问世加速了生成式人工智能的商业化进程，实现注册用户数量破亿仅仅耗时两个月，微软、谷歌等科技巨头纷纷接入，之后大模型热度持续火爆，带动算力需求扩容。Deepseek 问世推动大模型平价化，降低了大模型开发成本，利于为下游端侧和应用侧打开市场空间，下游爆发同样将催生大量算力需求，并推动算力需求由训练端向推理端转移。据 IDC 预测，推理的服务器工作负载占比预计由 2020 年的 51.5%逐年增加至 2026 年的 62.2%，中国人工智能服务器工作负载结构中的推理算力占比总体呈现增加趋势。

图 3：中国人工智能服务器工作负载预测，2020-2026

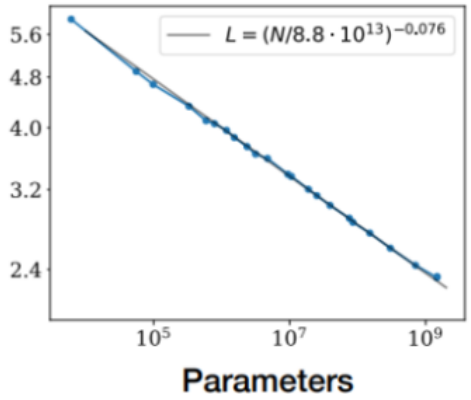


资料来源：IDC，源达信息证券研究所

1.参数数量总体呈现增加趋势

大模型的参数量与算力消耗呈现显著的正相关关系，当参数量不断增加，模型运行对硬件性能要求会大幅增加，能耗及时间成本也会随之上升，性能指标的提升也将同时推动模型优化技术的发展。OpenAI 团队经研究发现，模型性能与模型参数量、训练数据量和计算资源相关，通常大模型性能随着参数量、训练数据量和计算资源的增加而提升，这种现象被称为“Scaling Laws”。具体来说，参数量的增加与性能提升之间存在幂律关系，即参数数量增加的对数与性能提升之间呈近似线性关系。

图 4：参数量与大模型性能



资料来源：《Scaling Laws for Neural Language Models》, Kaplan et al. (2020)

近年来大模型快速迭代，大模型的参数量总体呈现上升趋势，以 OpenAI 发布的大模型为例，公司 2018 年发布的首款大模型 GPT-1 参数量为 1.17 亿，2019 年发布 GPT-2 大模型参数规模达到 15 亿，2019 年发布的 GPT-3 参数规模进一步达到 1,750 亿，2023 年发布 GPT-4 大模型参数规模突破万亿规模，达到 17,600 亿，近似呈现指数级增长，2025 年发布的 GPT-5 参数规模达到 20,000 亿，参数量继续增加但与 GPT-4 保持在同一数量级，参数规模有收敛的趋势。将主要大模型按照发布时间进行排序，参数量呈现出先爆发增长，后趋于收敛的类似变化。参数量与算力需求关系密切，参数量越大，模型复杂度越高，对算力的需求越大。以大模型训练为例，模型训练的总运算量与模型参数规模和 toke 数量的乘积有关，给定单卡运算性能和拟完成训练的时间，参数量越大意味所需芯片数量越多，算力需求越大。

表 1：不同模型参数规模

Models	Release time	Developers	Parameter size (Billion)
GPT-1	2018	OpenAI	1.17
BERT	2018	Google	3.40
GPT-2	2019	OpenAI	15.00
Fairseq	2020	Meta	130.00
GPT-3	2020	OpenAI	1750.00
GlaM	2021	Google	1200.00
LaMDA	2022	Google	1370.00
SparkDesk	2023	iFLYTEK	1700.00
GPT-4	2023	OpenAI	17600.00
Grok 3	2025	xAI	12000.00
GPT-5	2025	OpenAI	20000.00

资料来源：《大语言模型研究现状及趋势》，华尔街见闻，源达信息证券研究所

2.大模型火热，用户量激增

ChatGPT 3.5 于 2022 年重磅推出后收获了极好的市场反馈，发布当天便吸引了超过 10 万用户，五天后注册人数突破百万，获得 1 亿用户仅用时两个月，而知名应用软件 TikTok 达成 1 亿用户共耗时 9 个月，微信耗时 433 天。随后 OpenAI 于 2023 年 3 月发布了不仅能够处理多模态数据且智能水平大幅提高的 ChatGPT 4 大模型，微软、摩根士丹利等一众名企纷纷接入，ChatGPT 实现了 AI 大模型由实验室到商业化应用的历史性转变，大模型热度继续升温，Meta 跟进开源 Lama 大模型，百度发布文心一言大模型、阿里推出通义千问大模型、科大讯飞发布星火大模型，大模型赛道百花齐放。2025 年 1 月 20 日，Deepseek 发布重大更新推出 Deepseek-R1 模型，用户数量出现爆发式增长，2024 年 12 月底至 2025 年 1 月底，用户数由 34.7 万猛增至近 1.2 亿，实现 1 亿用户的增长仅用时 7 天，2 月 8 日国内 APP 端日活用户达到 3,494 万，跃居国内 1 月月均活跃用户数榜首。大模型拥有极高人气，用户规模或将持续增加，推理端算力需求将不断增长。

图 5：增长 1 亿用户花费时间

资料来源：AI 产品榜

3. Deepseek 推动大模型平价化，利好端侧、应用侧爆发

大模型训练成本高企，ChatGPT-4 的训练使用了约 25,000 块 A100 GPU，以 2.15e25 FLOPS 的计算量训练了 90 至 100 天。若 H100 每小时的租用成本为 1 美元，单次训练成本高达 6,300 万美元。为满足大模型训练的算力需求，多家 AI 巨头斥巨资打造万卡集群，即由一万张及以上的计算加速卡（如 GPU、TPU 或其他专用 AI 加速芯片）组成的高性能计算系统，用以支持千亿级甚至万亿级参数规模的大模型训练，而高端算力卡供应几乎被英伟达一家公司垄断，H100 的官方售价大约在 3000 美元左右，由于供需失衡和缺货原因，市场售价远高于官方价格，英伟达毛利率高达 70%以上。高昂的芯片价格拉升了大模型的训练成本，不利于以大模型为底层架构的应用侧及端侧的商业化，限制了 AI 产业的商业化进程。

通过一系列算法优化，Deepseek-V3 相较于同类模型，训练成本大幅下降，完成训练仅耗时不到两个月，按 H800 芯片算力测算， Deepseek-V3 预训练阶段训练时长为 266,4 万 GPU 小时，上下文扩展训练耗时 11.9 万 GPU 小时，后训练阶段耗时 5,000 GPU 小时，假设 H800 每小时的租赁价格为 2 美元，总训练成本为 557.6 万美元，训练成本仅为 GPT-4o 的十分之一。

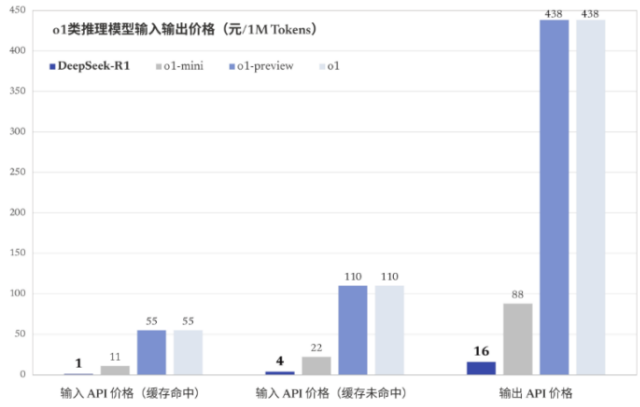
表 2：Deepseek-V3 模型训练成本

训练成本	预训练	上下文扩展	后训练
H800 GPU 小时（万）	2664.000	11.900	0.500
美元（百万）	5.328	0.238	5.576

资料来源：财联社，源达信息证券研究所

Deepseek-V3 模型和 Deepseek-R1 模型在保证模型性能的前提下，通过优化算法减少训练成本实现了 API 服务价格的显著下降，推动大模型平价化。Deepseek-V3 模型 API 服务定价为每百万输入 Token 0.5 元（缓存命中），每百万输入 Token 2 元（缓存未命中），每百万 Token 输出价格为 8 元。Deepseek-R1 模型每百万 tokens 输入为 1 元(缓存命中)，百万 tokens 输入为 4 元（缓存未命中），每百万 tokens 输出为 16 元。GPT-4 每百万输入 Token 约 70 元，大幅高于 Deepseek-V3 模型和 Deepseek-R1 模型的 API 调用价格。

图 6：推理模型输入输出价格（元/1M Tokens）



资料来源：Deepseek 官网

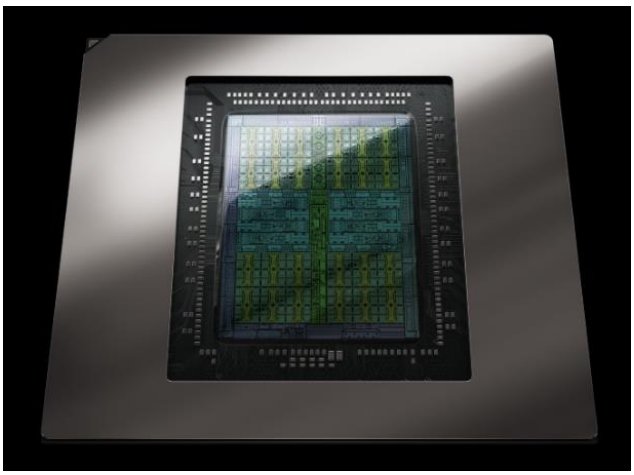
三、终端定制化特点突出，看好 ASIC 芯片发展前景

GPU 与 ASIC 芯片存在显著差别，GPU 在超多核的架构下可以用来处理通用的加速计算任务，如 AI 推理，科学计算，3D 渲染等等，GPU 具有较好的适配能力和通用性，适于在云端服务器部署，以满足不同客户的不同需求，适配不同的模型与任务。当前英伟达在 GPU 领域占据绝对领先地位，公司不仅产品性能优秀并兼具向后兼容能力，当数据中心迭代了新的 GPU，老式 GPU 则可用于训练，实现基础设施和代码的复用，能够帮助用户节省资本开支，提升算力基础设施投资的经济性。

此外，Cuda 能够深度赋能大模型开发，因训练和推理在代码层面有较高的重合度，因此使

用英伟达 GPU 进行训练的企业仅需复用其中的部分代码用于训练，不必再依托新的平台开发程序，大幅减少了开发成本，节约了开发时间。上述优势使英伟达在 GPU 领域构筑了护城河，竞争者在短期内或较难颠覆其行业地位，但随着 AI 的发展，端侧和应用侧出现了细分需求，轻量化、定制化、低功耗、低时延等需求凸显，为 ASIC 架构创造了机会。

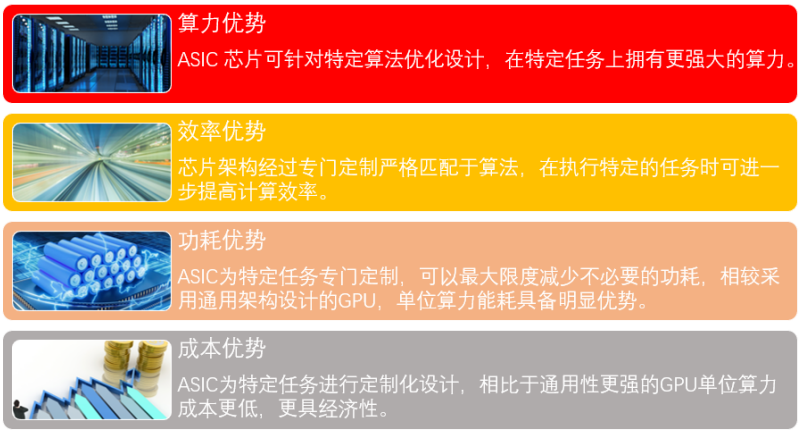
图 7：GeForce RTX 50



资料来源：英伟达官网

ASIC 芯片则专门用来优化神经网络推理或者矩阵运算任务，专注于特定用途或特定模型，相较 GPU 在功耗、可靠性、性能、成本等方面具备优势，因此更适于在端侧和用户侧部署，如智驾、AI 眼镜、智能家居等。随着大模型平价化，预期 AI 产品将在更多应用场景下实现商业落地，ASIC 芯片具备广阔的市场前景。

图 8：ASIC 芯片性能优势



资料来源：源达信息证券研究所

将目光转向芯片巨头英伟达，其核心产品功耗普遍在百瓦至几百瓦不等，这样的能耗很难在如手机、AI 眼镜甚至是汽车这样的终端使用。目前而言英伟达这样的头部玩家并没有进军这一赛道，且英伟达远离终端市场，不具备深度理解端侧不同应用场景下客户具体需求的先发优势，利于 ASIC 芯片厂商在这一赛道的布局和发力。

表 3：英伟达主流产品能耗（W）

型号	能耗 W
GeForce RTX 5090	575
GeForce RTX 5080	360
GeForce RTX 5070 Ti	300
GeForce RTX 5070	250
GeForce RTX 4090	425
GeForce RTX 4080	320
GeForce RTX 4070 Ti	285
GeForce RTX 4070	200

资料来源：英伟达官网，源达信息证券研究所

本地推理不仅可以降低延时、提升吞吐量，摆脱网络限制，还有助于增强数据安全和保护用户隐私，终端推理任务的本地化运行或是未来的发展趋势，本地推理预期将增加 ASIC 芯片的市场需求。此外，ASIC 芯片应用场景众多，并不限于人工智能领域，在国防、办公、安防、家居等行业都有广泛应用，随着智能化升级趋势的深入，ASIC 芯片的市场需求将持续扩容。

国防军工：ASIC 芯片为特定军事用途定制化设计，能够更好的满足军事用户在武器制导，精确打击方面的需求。且 ASIC 芯片具备高可靠性，高保密性的特点，符合军方对可靠性和信息安全方面的需要。

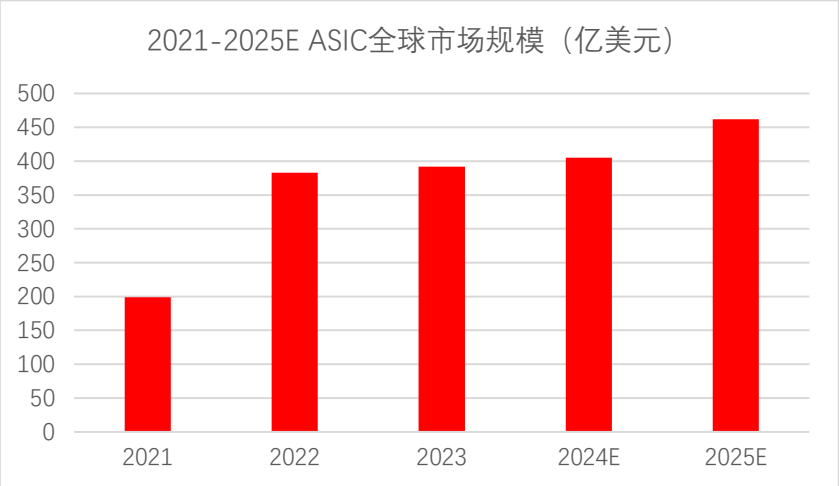
智慧办公：ASIC 芯片能够实现办公设备的智能化升级，赋能流程管理、决策、执行等不同环节，提升使用者的工作效率。

智慧安防：专业化的 ASIC 芯片能够高效进行图像识别、行为分析、视频的结构化分析等，提升安防的智能化程度，更好的完成安防任务。

智能家居：智能家居是 ASIC 芯片另一大应用场景，通过家居的智能化升级实现设备的互联互通，建立更便捷的人机交互，让客户获得更佳的使用体验。

得益于 AI 浪潮，ASIC 芯片市场近年来高速增长，据 IDC 数据显示，2021 年，全球 ASIC 芯片市场规模为 199 亿美元，2022 年达到 383 亿美元，2023 年达到 392 亿美元，随着 AI 商业化进程提速，预期市场需求将加速扩容，至 2025 年预计将达到 462 亿美元，符合增速高达 23.4%。

图 9：2021-2025E ASIC 全球市场规模（亿美元）



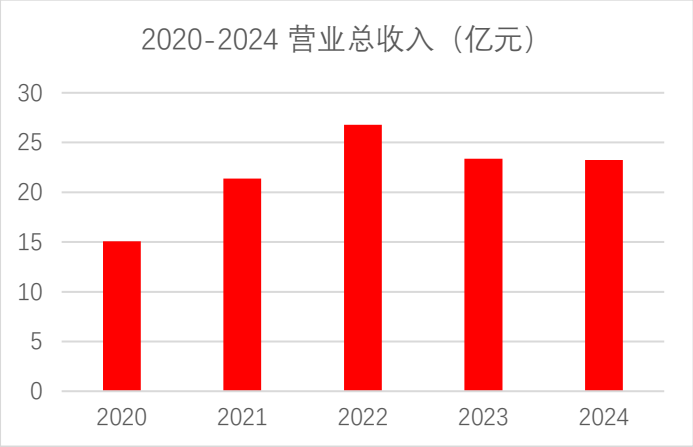
资料来源：IDC，源达信息证券研究所

四、投资建议

1、芯原股份

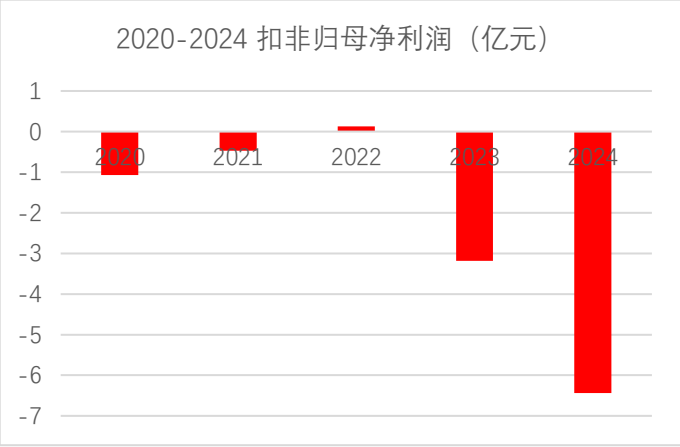
公司是一家依托自主半导体 IP，为客户提供平台化、全方位、一站式芯片定制服务和半导体 IP 授权服务的企业。公司通过基于自主半导体 IP 搭建的技术平台，可在短时间内打造出从定义到测试封装完成的半导体产品，为包含芯片设计公司、半导体垂直整合制造商(IDM)、系统厂商、大型互联网公司和云服务提供商在内的各种客户提供高效经济的半导体产品替代解决方案。公司业务范围覆盖消费电子、汽车电子、计算机及周边、工业、数据处理、物联网等行业应用领域。

图 10：2020-2024 营业总收入（亿元）



资料来源：Wind，源达信息证券研究所

图 11：2020-2024 扣非归母净利润（亿元）

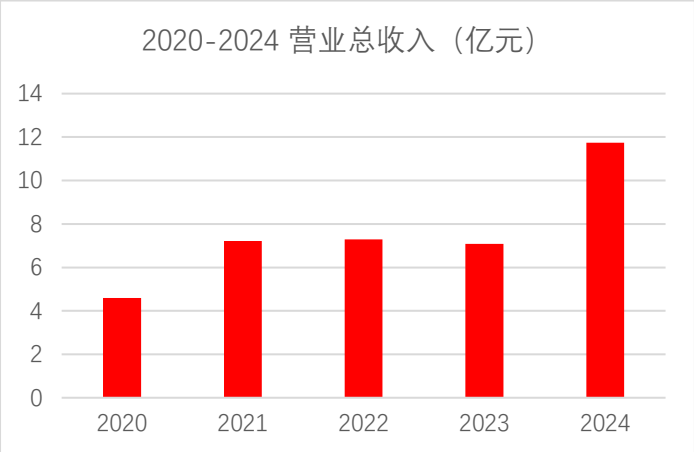


资料来源：Wind，源达信息证券研究所

2、寒武纪

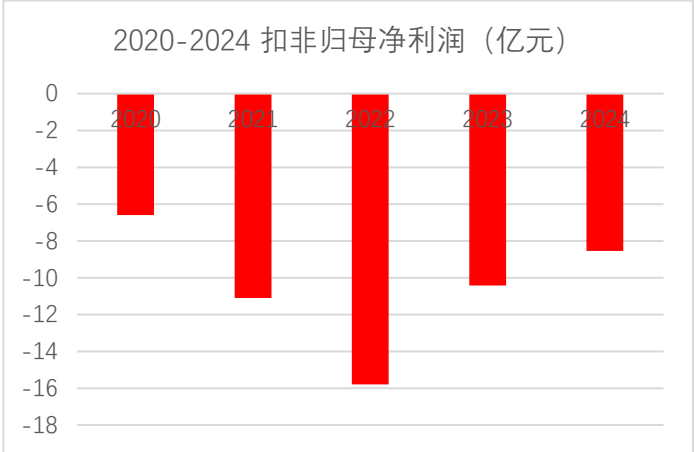
公司产品广泛应用于消费电子、数据中心、云计算等诸多场景。采用公司终端智能处理器 IP 的终端设备已出货过亿台；云端智能芯片及加速卡也已应用到国内主流服务器厂商的产品中，并已实现量产出货；边缘智能芯片及加速卡的发布标志着公司已形成全面覆盖云端、边缘端和终端场景的系列化智能芯片产品布局。

图 12：2020-2024 营业总收入（亿元）



资料来源：Wind，源达信息证券研究所

图 13：2020-2024 扣非归母净利润（亿元）



资料来源：Wind，源达信息证券研究所

五、风险提示

技术迭代风险

下游需求不及预期的风险

中美贸易摩擦加剧的风险

投资评级说明

行业评级	以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，投资建议的评级标准为：		
	看	好：	行业指数相对于沪深 300 指数表现＋10%以上
	中	性：	行业指数相对于沪深 300 指数表现 - 10%~＋10%以上
	看	淡：	行业指数相对于沪深 300 指数表现 - 10%以下
公司评级	以报告日后的 6 个月内，行业指数相对于沪深 300 指数的涨跌幅为标准，投资建议的评级标准为：		
	买	入：	相对于沪深 300 指数表现＋20%以上
	增	持：	相对于沪深 300 指数表现＋10%~＋20%
	中	性：	相对于沪深 300 指数表现 - 10%~＋10%之间波动
	减	持：	相对于沪深 300 指数表现 - 10%以下

办公地址

石家庄

河北省石家庄市长安区跃进路 167 号源达办公楼

上海

上海市浦东新区峨山路 91 弄 100 号陆家嘴软件园 2 号楼 701 室

分析师声明

作者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，以勤勉的职业态度，独立、客观地出具本报告。分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点。作者所得报酬的任何部分不曾与，不与，也不将与本报告中的具体推荐意见或观点而有直接或间接联系，特此声明。

重要声明

河北源达信息技术股份有限公司具有证券投资咨询业务资格，经营证券业务许可证编号：911301001043661976。

本报告仅限中国大陆地区发行，仅供河北源达信息技术股份有限公司（以下简称：本公司）的客户使用。本公司不会因接收人收到本报告而视其为客户。本报告的信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证，也不保证所包含信息和建议不发生任何变更。本公司已力求报告内容的客观、公正，但文中的观点、结论和建议仅供参考，不包含作者对证券价格涨跌或市场走势的确定性判断。本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估。

本报告仅反映本公司于发布报告当日的判断，在不同时期，本公司可以发出其他与本报告所载信息不一致及有不同结论的报告；本报告所反映研究人员的不同观点、见解及分析方法，并不代表本公司或其他附属机构的立场。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司及作者在自身所知情范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。