寒武纪-U（688256）

寒武破晓，算力腾飞

买入（首次）

盈利预测与估值	2022A	2023A	2024E	2025E	2026E
营业总收入（百万元）	729.04	709.39	1,115.85	3,587.61	5,411.63
同比（%）	1.11	(2.70)	57.30	221.51	50.84
归母净利润（百万元）	(1,256.35)	(848.44)	(461.17)	213.11	406.79
同比（%）	(52.29)	32.47	45.65	146.21	90.88
EPS-最新摊薄（元/股）	(3.01)	(2.03)	(1.10)	0.51	0.97
P/E（现价&最新摊薄）	(262.62)	(388.89)	(715.46)	1,548.24	811.11

投资要点

- **厚积薄发，打造人工智能产业核心引擎。**寒武纪作为中国智能芯片领域的标杆企业，专注于人工智能芯片的研发与技术创新，产品矩阵覆盖云、边缘和终端三大场景，逐步构建出完整的生态体系。公司股权结构稳定，核心管理层具备深厚的行业经验，同时通过激励机制保障团队活力，展现出对市场拓展和业务规模增长的信心。2024 年公司预计实现营业收入 10.7-12.0 亿元，同比增长 50.8%到 69.2%。

- **国产算力腾飞在即，寒武纪迎来黄金发展期。**在政府政策支持和企业需求激增的双重推动下，国产算力市场空间广阔。2025 年或将成为政府和运营商算力采购的大年，六部门定调到 25 年建设 105EFlops 智能算力，中国移动计划 24-25 年采购 AI 服务器 7994 台。互联网企业，特别是字节跳动等公司，在资本开支和 AI 推理需求上持续加码，这为国产算力厂商提供了历史性机遇。字节 CapEx 自 24 年 800 亿元飙升至 25 年 1600 亿元，对比北美云厂商 24 年平均 CapEx 约合 3800 亿人民币左右。我们预计国内云厂商 CapEx 上升空间依然广阔，有望带动算力芯片需求增长。我们认为寒武纪凭借技术优势和产品布局，有望在新一轮增长周期中获得显著市场份额。

- **智算未来先锋，寒武纪引领国产算力新格局。**寒武纪在技术路径上采取通用型智能芯片的开发路线，兼具高性能和低功耗，适配多场景应用，与国内外竞争者相比具备显著优势。具体体现在 1）“领跑者“计划推动数据中心算力国产替代；2）公司通过不断推出高性能芯片，强化产品迭代能力，进一步巩固其在国产算力领域的领先地位；3）差异化设计架构，凭借多样化运算的高效适配能力，与 Google TPU 对比各有千秋。

- **盈利预测与投资评级：**公司为国内 AI 芯片领军者，随着 AI 芯片市场规模的快速增长，公司有望依托不断扩展的产品布局和生态建设逐步抢占市场。我们预计公司 2024-2026 营业收入 11.16 / 35.88 /54.12 亿元，对应当前 PS 估值 292/91/60 倍。截至目前，可比公司 2024-2026 年 PS 估值为 52/36/28 倍。我们认为公司作为国产 AI 芯片稀缺标的，有望同时受益于 AI 行业的蓬勃发展以及算力国产替代的双重逻辑，故可享受一定估值溢价。首次覆盖，给予“买入”评级。

- **风险提示：**AI 需求不及预期风险，公司持续稳定经营风险，客户集中度较高风险，供应链稳定相关风险。

2025 年 02 月 26 日

证券分析师 陈海进
执业证书：S0600525020001
chenhj@dwzq.com.cn
研究助理 李雅文
执业证书：S0600125020002
liyw@dwzq.com.cn

股价走势

市场数据

收盘价(元)	790.38
一年最低/最高价	130.02/818.87
市净率(倍)	64.21
流通 A 股市值(百万元)	329,949.47
总市值(百万元)	329,949.47

基础数据

每股净资产(元,LF)	12.31
资产负债率(%,LF)	15.56
总股本(百万股)	417.46
流通 A 股(百万股)	417.46

相关研究

请务必阅读正文之后的免责声明部分

内容目录

请务必阅读正文之后的免责声明部分

图表目录

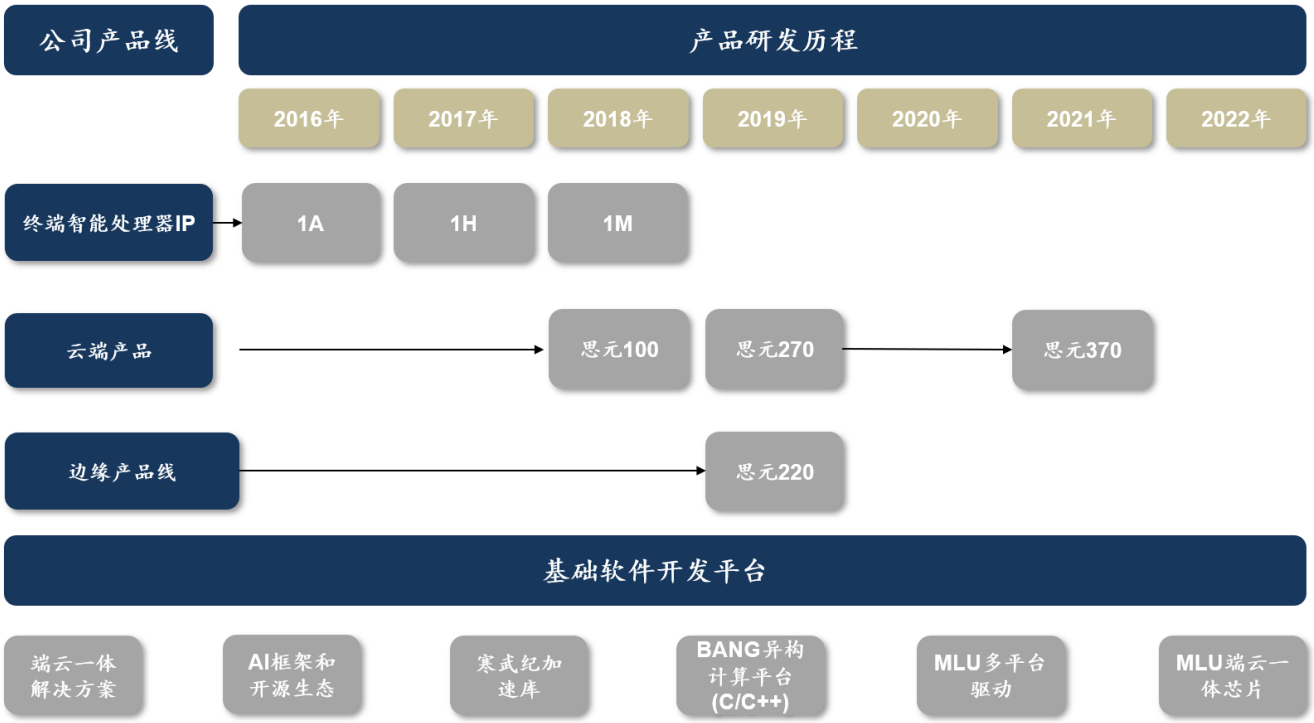
请务必阅读正文之后的免责声明部分

1. 公司介绍：厚积薄发，25 年剑指放量

1.1. 公司深耕产品开发，产品矩阵迅速扩充

寒武纪是中国最具代表性的智能芯片厂商之一。中科寒武纪科技股份有限公司于 2016 年成立于北京，目前已在上海、深圳等地成立分部。寒武纪是智能芯片领域全球知名的新兴公司，研发、设计、销售云边端一体、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，为客户提供丰富的芯片产品与系统软件解决方案。公司自成立以来一直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，让机器更好地理解和服务人类。公司的主营业务是应用于各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的研发、设计和销售，以及为客户提供丰富的芯片产品。目前，公司的主要产品线包括云端产品线、边缘产品线、IP 授权及软件。

图1：公司主要产品介绍



数据来源：公司 2022 年年度报告，公司官网，东吴证券研究所

1.2. 公司股权结构稳定，核心人员激励到位

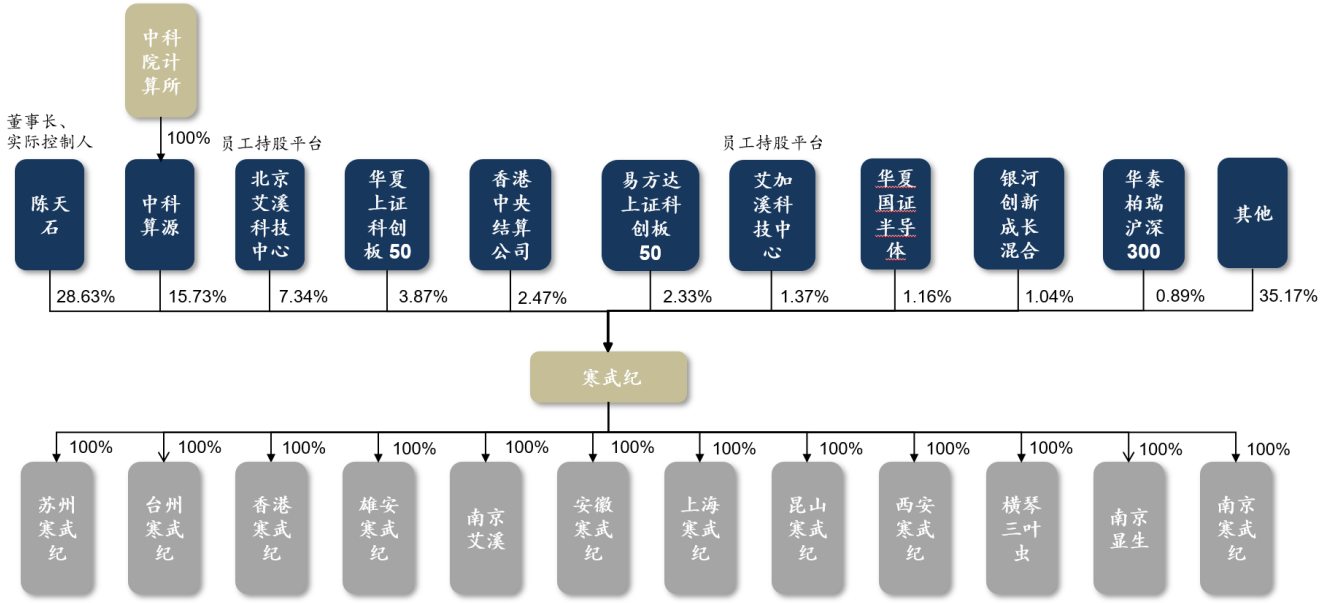
公司股权结构稳定，实控人持股比例 28.63%。截至 24 年三季度报，公司实控人为公司董事长、总经理陈天石，持股比例为 28.63%，系公司第一大股东。陈天石博士曾在中科院计算所担任研究员（正高级职称），在人工智能及处理器芯片领域从事基础科研工作十余年，积累了坚实的理论功底及研发经验。中科院计算技术研究所独资公司北京

请务必阅读正文之后的免责声明部分

中科算源资产管理有限公司为公司第二大股东，持股比例达 15.73%。第三大持股股东为员工持股平台北京艾溪科技中心，持股比例为 7.34%。

公司建立、健全长效激励机制，推动公司长远发展。公司 2020 年度、2021 年度和 2023 年度限制性股票激励计划仍在有效期内。 2023 年激励计划在第一个归属期的考核目标为 2024 年营业收入值不低于 11 亿元；第二个归属期的考核目标为 2024-2025 年累计营业收入值不低于 26 亿元；第三个归属期的考核目标为 2024-2026 年累计营业收入值不低于 46 亿元。高营收的考核指标体现公司对未来规模成长的信心。

图2：公司股权结构（截至 2024 三季度报）



数据来源：公司公告，东吴证券研究所

注 1：本图仅选取控股比例为 100%的子公司

注 2：子公司信息来源公司 2023 年年度报告

图3：公司仍在实施的股权激励计划

仍在实施的股权激励计划年份	2020年（预留授予部分）	2021年（首次授予部分）	2023年（首次授予部分）
目标值（公司层面归属系数100%）	2024年各类智能芯片及加速卡、训练整机销售收入21.53亿	2021-2024年累计营业收入值不低于46亿元	2024年营业收入值不低于11亿元
触发值（公司层面归属系数80%）	2024年各类智能芯片及加速卡、训练整机销售收入15.07亿	2021-2024年累计营业收入值不低于36.8亿元	2024年营业收入值不低于8.8亿元

数据来源：公司公告，东吴证券研究所

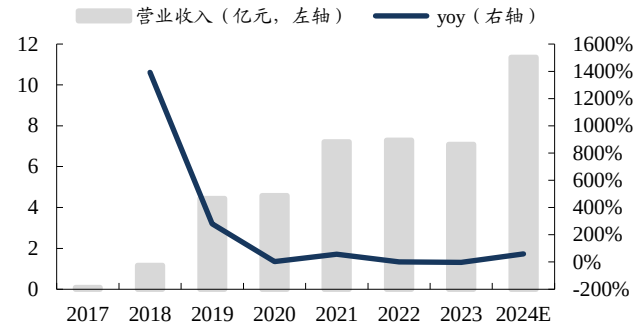
1.3. 财务分析：持续拓展市场，预计 2024 收入显著增长

公司营收长期增长，预计 2024 年度营收可观。公司营收 2017-2023 年基本保持增长趋势，根据公司 2024 年年度业绩预告，2024 年度预计实现营业收入 10.7-12.0 亿元，

请务必阅读正文之后的免责声明部分

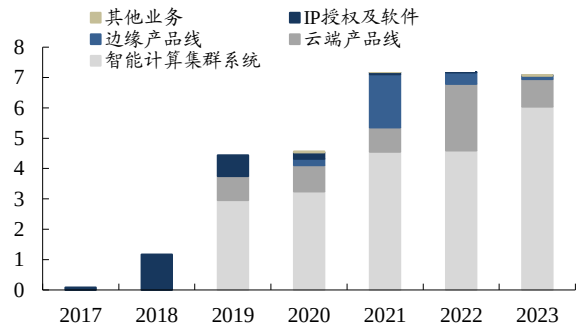
yoy+50.8%到 69.2%，这主要得益于公司持续拓展市场，积极助力人工智能应用落地。**从历史数据来看，智能计算集群系统构成营收主力。**2023 年智能计算集群系统营收较上年同期增长 31.9%，达到 6.1 亿元，贡献了 85.2%的营业收入，主要是因为公司成功在沈阳、台州实施智能计算集群项目，保持了智能计算集群系统业务收入的持续增长。**2024 年预计在互联网场景取得更多进展。**公司 24Q3 在大模型适配方面做出积极努力，8 月，飞桨新一代框架 3.0-beta 寒武纪版与 PaddleX 3.0-beta 寒武纪版在飞桨官网上线；9 月，公司开源了 PyTorch 设备后端扩展插件 Torch-MLU，并实现了硬件对于 PyTorch 的原生支持，充分提升了开发者的使用体验和集成效率。我们认为公司在软件生态上频频更新动态，或意味着客户端更多切实需求出现。

图4：公司营业收入情况



数据来源：公司公告，iFinD，东吴证券研究所；
注：2024E 由年度业绩预告取中值得出

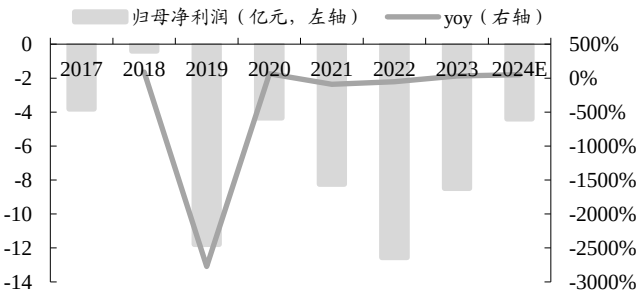
图5：公司营收结构情况（亿元）



数据来源：公司公告，iFinD，东吴证券研究所

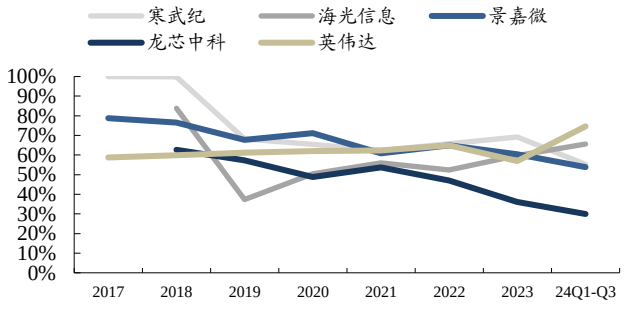
公司亏损面收窄，毛利率在可比公司中领先。净利率方面，2017 年至 24Q3 公司持续处于亏损状态。公司预计 2024 年度预计亏损 4.0 亿元到 4.8 亿元，同比亏损收窄 43.0%到 53.3%。公司尚未实现盈利且存在累计未弥补亏损，主要原因是公司为确保智能芯片产品及基础系统软件平台的高质量迭代，在竞争激烈的市场中保持技术领先优势，持续进行了大量的研发投入。**毛利率方面，**24Q1-Q3 公司毛利率为 55.2%，虽然相较 2023 年降低 13.9pcts，但仍在同行业较高水平。分析公司仍处于业绩释放初期，毛利率受多方因素影响，我们预计随着公司产品规模上量，毛利率波动有望逐渐缩小。

图6：公司归母净利润（亿元）



数据来源：公司公告，iFinD，东吴证券研究所
注：2024E 由年度业绩预告取中值得出

图7：可比公司毛利率

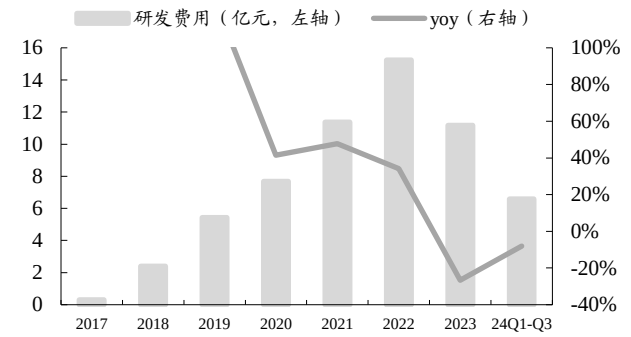


数据来源：各公司公告，iFinD，东吴证券研究所

请务必阅读正文之后的免责声明部分

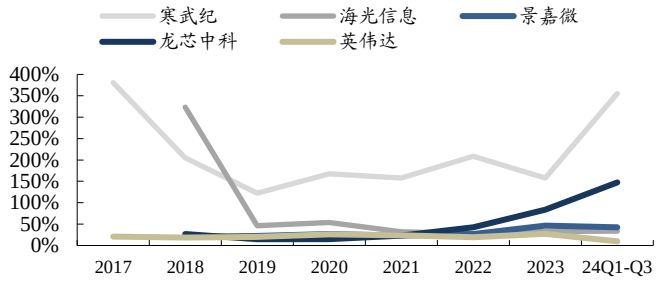
公司研发费用减少，费用率有望逐步恢复。寒武纪研发费用近年连续大幅度减少，主要原因是公司于 2022 年 12 月被美国商务部工业和安全局（BIS）以国家安全和外交利益为由列入“实体清单”，受此影响，公司调整战略，陆续暂停了部分预期毛利率较低的研发项目。24Q1-Q3 公司研发费用为 6.6 亿元，较去年同期减少 8.0%；研发费用率 355.7%，相对同行业仍然较高，我们预计随着公司产品大规模放量，研发费用率有望逐步恢复正常水平。

图8：公司研发费用（亿元）



数据来源：公司公告，iFinD，东吴证券研究所

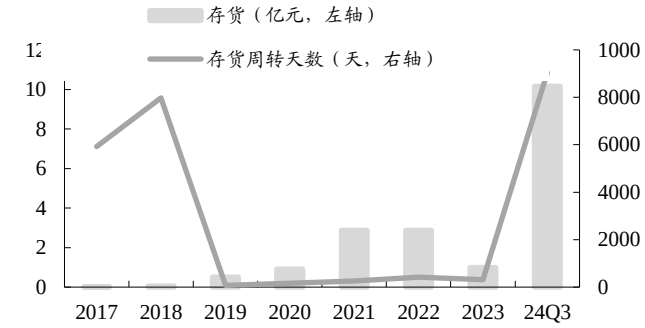
图9：可比公司研发费用率



数据来源：各公司公告，iFinD，东吴证券研究所

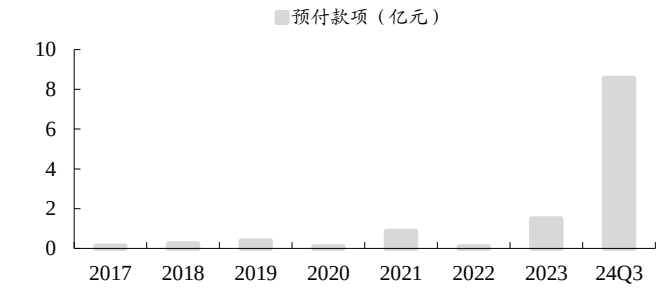
公司资产端实现大幅增长，或预示营收逐步释放。24Q3 公司存货大幅度增加，达到 10.2 亿元，此前 24Q1/Q2 分别为 1.3/2.4 亿元，环比数倍增长。24Q3 预付款项达到 8.5 亿，24Q1/Q2 分别为 2.1/5.5 亿，环比+55%。结合公司营收增长、研发投入较大的发展状况，我们预计公司订单数也有望逐渐增加。

图10：公司存货及存货周转天数（亿元，天）



数据来源：公司公告，iFinD，东吴证券研究所

图11：公司预付款项（亿元）



数据来源：公司公告，iFinD，东吴证券研究所

2. 如何看国产算力市场空间？

2.1. 政府&运营商：2025 年或为采购大年

政府：六部门定调到 25 年建设 105EFlops 智能算力，24 年底上海市智算需求目标大超预期。23 年 10 月，六部门联合印发《算力基础设施高质量发展行动计划》，提出到 25 年，算力规模超过 300EFlops，智能算力占比达到 35%，对应 105EFlops。24 年 3 月，

请务必阅读正文之后的免责声明部分

《上海市智能算力基础设施高质量发展“算力浦江”智算行动实施方案》要求至 25 年智能算力规模超 30EFlops；而在 24 年 12 月，上海市重新调整目标，印发《关于人工智能“模塑申城”的实施方案》，到 25 年底力争全市智能算力规模突破 100EFlops，建设 3-5 个大模型创新加速孵化器等。我们认为，此次上海市算力建设目标大超预期，且大多数省份算力规划的目标年均定在 25 年，因此 25 年政府算力采购方面值得重点关注。

图12：各省算力规划

省份	文件名	发布时间	算力需求	2025
湖南	《湖南省强化“三力”支撑规划（2022-2025年）》	2022-05	2025年10EFLOPS	10
广西	《广西信息通信业“畅联八桂”行动计划（2023-2025年）》	2023-06	2025年2.5EFLOPS	3
福建	《福建省新型基础设施建设三年行动计划（2023-2025年）》	2023-07	2024年6.5EFLOPS，2025年8EFLOPS	8
湖北	《湖北省加快发展算力与大数据产业 三年行动方案（2023-2025年）》	2023-08	2025年8EFLOS	8
重庆	《重庆市算力网络发展“算力山城 强算赋能”行动计划（2023-2025年）》	2023-12	2024年8EFLOPS，2025年10EFLOPS	10
安徽	《安徽省数字基础设施建设发展三年行动方案（2023-2025 年）》	2023-12	2025年12EFLOPS	12
山西	《山西省算力基础设施高质量发展实施方案》	2024-01	2025年9EFLOPS	9
青海	《青海省绿色算力基地建设方案》	2024-02	2025年2.06EFLOPS	2
上海	《上海市智能算力基础设施高质量发展“算力浦江”智算行动实施方案（2024-2025年）》	2024-03	2025年30EFLOPS	30
浙江	《浙江省智能物联产业集群建设行动方案》	2024-03	2027年40EFLOPS	20
河南	《河南省加快制造业“六新”突破实施方案》	2024-03	2025年2EFLOPS	2
广东	《广东省算力基础设施高质量发展行动暨“粤算”行动计划（2024-2025年）》	2024-03	2024年28EFLOPS，2025年38EFLOPS	38
北京	《北京市算力基础设施建设实施方案（2024-2027年）》	2024-04	2025年45EFLOPS	45
江苏	《江苏省算力基础设施发展专项规划》	2024-04	2025年24EFLOPS，2030年50EFLOPS	24
陕西	《陕西省加快推动人工智能产业发展实施方案（2024-2026年）》	2024-05	2026年3EFLOPS	2
河北	《河北省人民政府办公厅关于进一步优化算力布局推动人工智能产业创新发展的意见》	2024-05	2025年35EFLOPS	35
山东	《山东省算力基础设施高质量发展行动方案》	2024-05	2025年12.5EFLOPS	13
甘肃	《甘肃算力基础设施高质量发展三年行动计划（2024-2026年）》	2024-05	2026年30EFLOPS	
西藏	《“算力珠峰”高质量发展行动计划（2024-2026）》	2024-06	2026年0.1EFLOPS	
天津	《天津市算力产业发展实施方案（2024-2026年）》	2024-07	2026年10EFLOPS	7
贵州	《贵州省“千兆黔省、万兆筑城”行动计划（2024-2025年）》	2024-07	2024年70EFLOPS，2025年200EFLOPS	200
四川	《四川省算力基础设施高质量发展行动方案（2024-2027年）》	2024-11	2027年40EFLOPS	20
上海	《关于人工智能“模塑申城”的实施方案》	2024-12	2025年100EFLOPS	100

数据来源：各省政府官方，上海证券报，中国证券网，浙江省民营经济研究中心，新京报，湖北网络广播电视台，东吴证券研究所；注：所列数据为总的算力规模，包含智算和超算算力

运营商：23H2-24H1 AI 服务器集采需求超过 1.7 万台，约对应 14 万张 GPU 卡需求。中国移动 24 年 4 月发布公告，计划 24-25 年采购 AI 服务器 7994 台。此前，移动 23-24 年招标采购 AI 服务器计划已达到 2454 台，招标量合计已超过万台。联通和电信也分别在 24 年 3 月/23 年 10 月公告智算采购计划，约为 2503/4175 台。我们假设单台服务器采用 8 张 GPU 卡，由此测算以上四次 AI 服务器采购招标合计大约产生 13.7 万张 GPU 卡需求。

图13：三大运营商 23H2-24H1 AI 服务器采购情况

运营商	公告日期	公告文件	服务器采购数（台）	GPU需求量（万张）
移动	23Q3	《2023至2024年新型智算中心（试验网）采购招标公告》	2454	2.0
	24Q2	《2024至2025年新型智算中心采购招标公告》	7994	6.4
电信	23Q3	《AI算力服务器（2023-2024年）集中采购项目》	4175	3.3
联通	24Q1	《2024年中国联通人工智能服务器集中采购项目》	2503	2.0
合计			**17126**	**13.7**

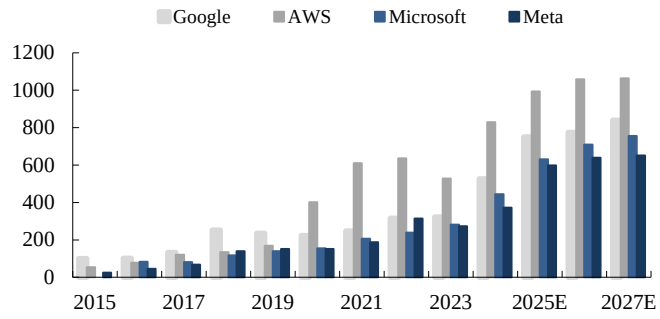
数据来源：科创板日报，证券时报，集微网，中国工信新闻网，格隆汇等，东吴证券研究所；注：图中 GPU 需求量为测算数据，我们假设单台服务器采用 8 张 GPU 卡

请务必阅读正文之后的免责声明部分

2.2. 互联网：字节 CapEx 大增，重视互联网推理需求

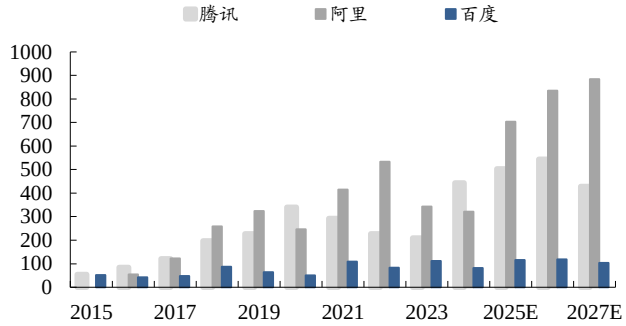
字节引领国内云厂商加大 CapEx 投入，供给格局进一步清晰。字节跳动在 AI 上积极投入，仅 2024 年就达到 800 亿元，百度、阿里、腾讯三家资本开支的总和为 845 亿元。2025 年字节跳动的资本开支预算飙升至近 1600 亿人民币，其他三家合计为 1322 亿元。由此，可粗略计算 2025 年国内四大云厂商资本开支合计为 2922 亿元(yoy: +78%)。横向对比北美四大云厂商 2024 年 CapEx 合计为 2172 亿美元，各家厂商平均投入在 543 亿美元左右，约合 3837 亿人民币左右。我们预计国内云厂商在资本开支方面的上升空间依然广阔。供给格局来看，随着 2024 年“领跑者计划”（H20 出货受限）、AI 曼哈顿计划等陆续颁布，中美双向制裁共同促成国产算力的加速替代，我们预计 2025 年国产算力头部厂商有望迎来大幅放量。

图14：北美云厂商 CapEx 及预测（亿美元）



数据来源：Bloomberg，东吴证券研究所
注：预测数据为 Bloomberg 一致预期；以上均为财年

图15：国内云厂商 CapEx 及预测（亿元）



数据来源：Bloomberg，东吴证券研究所
注：预测数据为 Bloomberg 一致预期；2024 年腾讯 CapEx 数据采用 Bloomberg 一致预期；以上均为财年

请务必阅读正文之后的免责声明部分

3. 竞争地位：智算未来先锋，寒武纪引领国产算力新格局

3.1. 对比英伟达 H20：“领跑者”计划推动数据中心算力国产替代

在严格控制能效的政策背景下，H20 在国内市场的应用或将受限。2024 年 7 月，国家发改委、工信部、国家能源局、国家数据局等四部门印发《数据中心绿色低碳发展专项行动计划》，提出新建及改扩建数据中心应采用能效达到**《服务器和数据存储设备能效“领跑者”评价要求》(T/CECA-G 0284-2024)**规定的节能水平及以上服务器产品。评价要求提到，对于服务器 GPU，14nm 以下芯片至少达到 0.5 TFLOPS/W（节能水平），先进水平则要求达到 1.0 TFLOPS/W。而英伟达此前对华特供的 H20 芯片，FP16 算力达 148 TFLOPS，功耗为 400W，对应 0.37 TFLOPS/W，不满足评价要求的节能水平。

图16：NVIDIA H20 参数信息

NVIDIA H20	
GPU Architecture	NVIDIA Hopper
GPU Memory	96 GB HBM3
GPU Memory Bandwidth	4.0 TB/s
INT8 | FP8 Tensor Core	296 | 296 TFLOPS
BF16 | FP16 Tensor Core	148 | 148 TFLOPS
TF32 Tensor Core	74 TFLOPS
FP32	44 TFLOPS
FP64	1 TFLOPS
RT Core	N/A
MIG	Up to 7 MIG
L2 Cache	60 MB
Media Engine	7 NVDEC
	7 NVJPEC
Power	400 W
Fom Factor	8-way HGX
Interconnect	PCIe Gen5 x16:128 GB/s
	NVLink: 900GB/s

数据来源：弘信电子官网，东吴证券研究所

3.2. 对比国产竞争者：寒武纪占据第一梯队

在人工智能数十年的发展历程中，**传统芯片**曾长期为其提供底层计算能力。这些传统芯片包括 CPU、GPU、DSP、FPGA 等，代表公司有 Intel、AMD、Nvidia、ARM 等，但传统芯片在芯片架构、性能、能效等方面并不能适应人工智能技术与应用的快速发展。而**智能芯片**是专门针对人工智能领域设计的芯片，包括**通用型智能芯片**与**专用型智能芯片（ASIC）**两种类型。

请务必阅读正文之后的免责声明部分

图17：AI 芯片技术路线对比

芯片类型		技术原理	技术发展情况与技术特点	技术优势与技术局限性	市场需求情况	未来发展、演化或融合的趋势	代表芯片类型与厂商	市场渗透率
传统芯片	CPU	（1）CPU的基本原理为：通过灵活的控制单元、细粒度的运算单元、多层次的缓存、多发射流水线，实现对于通用计算任务灵活和高效的支持 （2）具体对于智能训练和推理应用，通过CPU的基本指令组合出训练或推理需要的运算操作，从而实现对智能算法的支持	技术成熟，通用性最强，可执行各种类型的计算机应用程序，非常适合传统的控制密集型计算任务	人工智能应用开发生态成熟，但性能已无法满足人工智能快速增长的计算能力需求	通用计算市场需求大且稳定	CPU的演化趋势为集成更多更高速的外部接口，长期看仍将主要应用于通用计算	（1）云端服务器和PC市场多使用X86 CPU，已经迭代成熟，主流芯片基于X86-64架构，主流工艺包括7nm、10nm、14nm、16nm （2）终端和边缘端多使用ARM CPU，也已经迭代成熟，主流芯片基于ARMv8架构，主流工艺范围较广，从7nm到60nm工艺均有产品	广泛应用于个人电脑、移动终端、传统服务器等领域，在人工智能芯片市场渗透率相对较低
	GPU	（1）GPU的基本原理为：通过简化控制单元并集成大规模的并行运算单元，实现对图形渲染等并行任务的良好支持 （2）具体对于智能训练和推理应用，通过GPU的向量等指令组合出训练或推理需要的运算操作，从而实现对智能算法的支持	技术成熟，通用性较好，擅长数据级并行处理，为图形处理、科学计算等传统任务提供了良好的硬件支持	峰值运算性能高，但整体能耗较高；在云端具备成熟的应用开发生态，但在终端生态尚不成熟	图形渲染和科学计算市场需求大且稳定，但在人工智能领域面临通用型智能芯片的挑战	GPU的演化趋势为持续保持其在图形渲染和科学计算领域的技术优势，加强对人工智能领域的支持	（1）云端主流产品为AMD和Nvidia产品，主流工艺为7/12/16nm （2）边缘端或终端主流产品为SoC集成的GPUIP，主流厂商包括ARM、Imagination等	在人工智能领域，GPU多用于服务器与数据中心，是目前渗透率最高且最主流的芯片类型，在终端应用较少
智能芯片	通用型智能芯片	（1）通用型智能芯片的原理是：通过对各类智能应用和算法的计算和访存特点进行抽取和抽象，定义出一套适用于智能算法且相对灵活的指令集和处理器架构，从而广泛支持多样化的人工智能算法和应用 （2）智能芯片的指令通常与人工智能算法中的关键运算操作相匹配 （3）在具体的训练和推理应用中，对于关键运算操作，智能芯片指令可直接支持，从而实现高效的训练和推理	相关技术持续发展中，全新指令集完备高效，可覆盖各类智能算法所需的基本运算操作	性能、功耗比较传统芯片优势明显，可适应各种场景和规模的人工智能计算需求	人工智能市场需求潜力大，未来将成为该市场主流产品	云端智能芯片将集成更高计算能力、更高速的外围接口及更先进的集成电路工艺；边缘及终端智能芯片将集成多样化的模块，沿SoC技术路径继续深度发展	（1）云端和边缘通用型智能芯片处于应用推广期，主要厂商和产品为寒武纪（思元100/270/220）、华为海思（Ascend 310/910）、Google（TPU V1/V2/V3、TPU EDGE）等 （2）终端通用型智能处理器多集成于手机SoC等芯片中，已实现大规模应用，主要厂商和产品为华为海思（麒麟970/980/990）等	在云端、边缘端和消费类电子终端都开始出现广泛应用，渗透率将逐渐提升
	专用型智能芯片（ASIC）	专用型智能芯片的原理是：针对面向特定的、具体的、相对单一的人工智能应用专门设计的芯片，具体实现方法为在架构层面对特定智能算法作硬化支持，指令集简单或指令完全固化	相关技术持续发展中，在架构层面对特定智能算法作硬化支持，指令集简单或指令完全固化	成本相对较低，软件栈相对简单，设计和生产周期短，但通用性较差	应用细分市场需求大且分散，成本敏感	专用型智能芯片逐渐融入各类行业专用SoC芯片（如智能音箱芯片）中	目前主要应用于终端，主要形态为行业专用SoC，较多集中于语音处理领域	常用于在低功耗、成本敏感的终端上支撑特定的智能应用，在云端、边缘端等场景渗透率相对较低

数据来源：公司招股书（披露于 20 年 7 月），东吴证券研究所

区别于国产算力的其他竞争者，公司优势显著。从技术路径上看，公司主要研发**通用型智能芯片**，这类芯片的特点是可覆盖各类智能算法所需的基本运算操作，同时性能功耗比较传统芯片优势明显，可适应各种场景和规模的人工智能计算需求。国内竞争者中，根据海光信息招股书披露，**海光 DCU** 属于 GPGPU 的一种，其结构逻辑相对 CPU 简单，但计算单元数量较多；**华为昇腾利**用 ASIC 技术，常用于在低功耗、成本敏感的终端上支撑特定的智能应用，芯片架构相对简单，技术门槛相对较低。公司认为未来人工智能市场需求潜力大，通用型智能芯片将成为市场主流产品。

公司产品不断迭代，有望成为国产算力“先锋”。2022 年 3 月 21 日，寒武纪正式发布新款训练加速卡 MLU370-X8。MLU370-X8 搭载双芯片四芯粒思元 370，集成寒武纪 MLU-Link™多芯互联技术，主要面向训练任务，在业界应用广泛的 YOLOv3、Transformer

请务必阅读正文之后的免责声明部分

等训练任务中，8 卡计算系统的并行性能平均达到 350W RTX GPU 的 155%。MLU370-X8 加速卡与国内主流服务器合作伙伴的适配工作已经完成，并已对客户实现小规模出货。

图18：思元 370 产品线参数对比

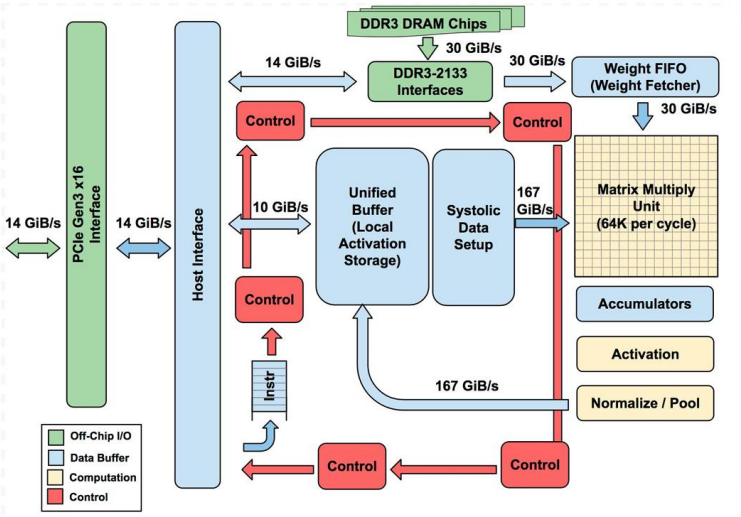
板卡型号	MLU370-S4/S8	MLU370-X4	MLU370-X8
制程工艺	7nm		
计算精度支持	FP32、FP16、BF16、INT16、INT8、INT4		
峰值性能	192 TOPS (INT8)	256 TOPS (INT8)	256 TOPS (INT8)
	96 TOPS (INT16)	128 TOPS (INT16)	128 TOPS (INT16)
	72 TFLOPS (FP16)	96 TFLOPS (FP16)	96 TFLOPS (FP16)
	72 TFLOPS (BF16)	96 TFLOPS (BF16)	96 TFLOPS (BF16)
	18 TFLOPS (FP32)	24 TFLOPS (FP32)	24 TFLOPS (FP32)
内存类型	LPDDR5		
内存容量	24GB/48GB	24GB	48GB
内存带宽	307.2 GB/s	307.2 GB/s	614.4 GB/s
最大热功耗	75W	150W	250W

数据来源：公司官网，东吴证券研究所

3.3. 对比 Google TPU：架构设计各有千秋

除了公司研发的智能芯片，通用型智能芯片的代表性产品还有 Google TPU。寒武纪产品在处理器架构上采用了不同的路线。**Google TPU 的核心是经典的脉动阵列机技术**，脉动阵列本身对于卷积类运算的效率较高，但是对于相对低频的部分运算操作（如全连接运算、激活运算）的效率不高。对于后者，Google TPU 引入了额外的硬件单元作为补充。而寒武纪的芯片架构，则直接将算法的基本操作区分为高位张量运算、向量运算、算数逻辑运算，并在处理器中分别通过高维张量计算部件、向量计算部件、传统算术逻辑计算部件予以处理：高维张量计算部件可高效支持卷积运算、全连接运算，而向量计算部件则可以支持激活等运算，传统算术逻辑计算部件则可以支持分支跳转等。

图19：TPU 设计架构图



数据来源：Norman P. Jouppi《In-Datacenter Performance Analysis of a Tensor Processing Unit》，东吴证券研究所

请务必阅读正文之后的免责声明部分

4. 盈利预测与投资建议

4.1. 盈利预测

公司为国内 AI 芯片领军者，随着 AI 芯片市场规模的快速增长，公司有望依托不断扩展的产品布局和生态建设逐步抢占市场。我们预计公司 2024-2026 营业收入 11.16 / 35.88 /54.12 亿元。分业务假设如下：

（1）**云端产品线**：2023 年受供应链影响，公司云端业务营收同比-59%，存在一定程度的下降。根据公司 2023 年报，云端产品线目前包括云端智能芯片、加速卡及训练整机，训练整机产品与智能计算集群系统业务的区别在于训练整机主要提供计算集群中的单体训练服务器，而不提供全集群搭建服务，主要面向有一定技术基础的商业客户群体。**在软件生态方面**，8 月，飞桨新一代框架 3.0-beta 寒武纪版与 PaddleX 3.0-beta 寒武纪版在飞桨官网上线；9 月，公司开源了 PyTorch 设备后端扩展插件 Torch-MLU，并实现了硬件对于 PyTorch 的原生支持。公司在大模型适配方面做出积极努力，我们认为这意味着互联网场景将有更多切实需求出现。**从地缘政治角度，**美国陆续颁布“AI 曼哈顿计划”与新一轮芯片出口管制，而《服务器和数据存储设备能效“领跑者”评价要求》提出更严格的能效控制要求，英伟达对华特供的 H20 芯片或将受到限制，中美双向制裁共同促成国产算力的加速替代。结合国内云厂商相对北美在资本支出上的上升空间，我们预计云端业务有望在 24 年修复，并在 25 年见到大规模放量。看好行业高景气度带动公司毛利率端持续提升，预计公司云端产品线业务 2024-2026 年毛利率为 61%/58%/58%。

（2）**智能计算集群系统**：2023 年公司智能计算集群系统营收大幅度提升，同比+32%，从营收占比上看，目前智能计算集群系统占据绝大部分营收份额。随着集群算力在 AI 行业的重要性持续提升，我们预测公司未来以智能计算集群系统出货的形式仍有望占据较大比重。根据公司 2023 年报，公司智能计算集群系统业务是将公司自研的加速卡或训练整机产品与合作伙伴提供的服务器设备、网络设备与存储设备结合，并配备公司的集群管理软件组成的数据中心集群，其核心算力来源是公司自研的云端智能芯片。该业务项目规模一般较大，订单量对公司营收体量或将产生较大影响。公司 2022 年拿下南京智算项目，2023 年拿下沈阳、台州两个项目（其中台州项目已履行完毕）。伴随《北京市算力基础设施建设实施方案（2024—2027 年）》、《算力基础设施高质量发展行动计划》等政策的提出与陆续实施，**我们看好 2025 年将是政府的国产算力采购大年**，公司或将拿到并执行更多新增项目，支撑集群营收持续快速增长，我们看好行业高景气度带动公司毛利率端持续提升，预计 2024-2026 智能计算集群业务毛利率为 72%/70%/68%。

（3）**边缘产品线**：公司边缘端业务的产品为思元 220 系列。截至 2023 年年报，思元 220 自发布以来，累计销量突破百万片。根据本文图表 1 所示，思元 220 为 2019 年推出产品，当前尚未看到公司在边缘产品线的迭代产品，我们认为在国产芯片行业竞争持续加剧的情况下，公司边缘产品线或存在一定的出货压力，在公司总营收占比中或将

维持较小比例。我们预计公司边缘产品线业务 2024-2026 年营收有望实现营收 0.08/0.06/0.05 亿元，毛利率达 53%/51%/48%。

图20：寒武纪盈利预测

688256.SH	单位	2019	2020	2021	2022	2023	2024E	2025E	2026E
营收	亿元	4.44	4.59	7.21	7.29	7.09	11.16	35.88	54.12
云端产品线		0.79	0.86	0.80	2.19	0.91	1.36	14.94	37.36
边缘产品线			0.21	1.75	0.38	0.11	0.08	0.06	0.05
智能计算集群系统		2.96	3.26	4.56	4.59	6.05	9.69	20.84	16.68
IP授权及软件		0.69	0.22	0.07	0.01	0.00	0.00	0.00	0.00
其它业务		0.00	0.04	0.03	0.12	0.03	0.03	0.03	0.03
营收同比		279%	3%	57%	1%	-3%	57%	222%	51%
云端产品线			9%	-7%	174%	-59%	50%	1000%	150%
边缘产品线				741%	-78%	-71%	-30%	-20%	-20%
智能计算集群系统			10%	40%	1%	32%	60%	115%	-20%
IP授权及软件		-41%	-68%	-68%	-83%	-79%	0%	0%	0%
其它业务		-72%	4226%	-29%	280%	-73%	0%	0%	0%
毛利率		68%	65%	62%	66%	69%	70%	65%	61%
云端产品线		78%	76%	59%	63%	61%	61%	58%	58%
边缘产品线			49%	41%	31%	56%	53%	51%	48%
智能计算集群系统		58%	62%	71%	70%	71%	72%	70%	68%
IP授权及软件		100%	100%	100%		100%	100%	100%	100%
其它业务		77%	53%	75%	14%	93%	60%	56%	70%

数据来源：公司公告，东吴证券研究所预测

4.2. 投资建议

结合寒武纪的主营业务，我们选取英伟达、海光信息、龙芯中科和景嘉微作为可比公司。英伟达是“图形处理芯片 GPU 的发明者”和全球最大的 GPU 供应商，在人工智能领域，英伟达的 GPU 产品可覆盖云端训练、云端推理、终端推理等各类应用场景，尤其是在云端（数据中心）的泛人工智能类芯片市场占据优势地位。海光信息主营业务为应用于服务器、工作站的高端处理器，DCU 深算三号研发进展顺利；龙芯中科专注于面向信息和工控系统的处理器及配套芯片研发，将基于 2K3000 的 GPGPU 技术及 3C6000 的龙链技术研制专用 GPGPU 芯片；景嘉微为国内 GPU 厂商，2024 年，公司成功研发了景宏系列高性能智算模块与整机产品，可应用于 AI 训练、AI 推理和科学计算等领域。截至目前，可比公司 2024-2026 年 PS 估值为 52/36/28 倍。

AI 算力作为国产替代的重中之重，未来发展空间广阔，公司作为国产 AI 芯片领军者，凭借产品优势的逐步构建，在 AI 蓬勃发展的趋势下有望持续受益。另外，国内政策端持续给予政策支持，公司有望在国内智算建设中拿到更多订单，支撑公司快速成长。我们认为公司作为国产 AI 芯片稀缺标的，有望同时受益于 AI 行业的蓬勃发展以及算力国产替代的双重逻辑，故可享受一定估值溢价，我们预计公司将在 2024-2026 年实现收入 11.16 / 35.88 /54.12 亿元，对应当前 PS 估值 292/91/60 倍。首次覆盖，给予“买入”评级。

请务必阅读正文之后的免责声明部分

图21：可比公司估值表

单位：亿元/亿美元		总市值	营业收入			PS(X)		
			2024	2025	2026	2024	2025	2026
NVDA.O	英伟达	31,906	1,293	1,992	2,419	25	16	13
688041.SH	海光信息	3,581	89	126	167	40	28	21
688047.SH	龙芯中科	557	6	9	12	91	64	47
300474.SZ	景嘉微	503	9	14	18	54	37	28
	平均值					52	36	28
688256.SH	寒武纪	3,258	11.2	35.9	54.1	292	91	60

数据来源：各公司公告，Wind，iFinD，东吴证券研究所预测

注 1：收盘价信息截至 2025 年 2 月 25 日，除寒武纪采用东吴预测外，其他 A 股上市公司均采用 iFinD 一致预期

注 2：英伟达采用 Wind 一致预期，年份对应为 FY2025/2026/2027

5. 风险提示

AI 需求不及预期风险。总体来看，人工智能芯片技术仍处于发展阶段，技术迭代速度较快，技术发展路径尚在探索中，尚未形成具有核心优势的架构和系统生态。若 AI 产业的需求不及预期，可能影响公司产品的价格与需求量。

公司持续稳定经营风险。公司目前处于持续高强度研发投入阶段，通过技术创新，保持技术的领先性、提升产品的市场竞争力。受到行业政策、国际政治经济环境、市场竞争、市场需求及研发技术产品化等综合因素的影响，公司核心技术优势转化为业绩收入存在一定的滞后性。

客户集中度较高风险。2021 年、2022 年和 2023 年，公司前五大客户的销售金额合计占营业收入比例分别为 88.60%、84.94%和 92.36%，客户集中度较高。若公司主要客户对公司产品的采购量大幅降低或者公司未能继续维持与主要客户的合作关系，将给公司业绩带来显著不利影响。

供应链稳定相关风险。公司采用 Fabless 模式经营，供应商包括 IP 授权厂商、服务器厂商、晶圆制造厂和封装测试厂等。由于集成电路整个行业链是专业化分工且技术门槛较高，加之公司及部分子公司已被列入“实体清单”，将对公司供应链的稳定造成一定风险。切换新供应商将产生一定成本，将可能对公司经营业绩产生不利影响。

请务必阅读正文之后的免责声明部分

寒武纪-U 三大财务预测表

资产负债表（百万元）	2023A	2024E	2025E	2026E
流动资产	**5,648**	**5,820**	**7,070**	**8,034**
货币资金及交易性金融资产	4,654	3,357	1,991	1,285
经营性应收款项	792	955	2,444	3,323
存货	99	1,396	2,448	3,221
合同资产	40	45	90	108
其他流动资产	63	68	96	97
非流动资产	**771**	**816**	**811**	**810**
长期股权投资	230	230	230	230
固定资产及使用权资产	183	121	86	65
在建工程	109	109	109	109
无形资产	150	200	230	250
商誉	0	0	0	0
长期待摊费用	6	6	6	6
其他非流动资产	94	151	151	151
资产总计	**6,418**	**6,636**	**7,881**	**8,844**
流动负债	**463**	**601**	**1,382**	**1,724**
短期借款及一年内到期的非流动负债	35	19	19	19
经营性应付款项	237	335	874	1,171
合同负债	0	0	0	0
其他流动负债	191	247	489	534
非流动负债	225	563	805	1,005
长期借款	0	0	0	0
应付债券	0	0	0	0
租赁负债	5	5	5	5
其他非流动负债	220	558	800	1,000
负债合计	**689**	**1,164**	**2,188**	**2,730**
归属母公司股东权益	5,650	5,409	5,622	6,029
少数股东权益	80	64	71	85
所有者权益合计	**5,730**	**5,472**	**5,693**	**6,114**
负债和股东权益	**6,418**	**6,636**	**7,881**	**8,844**

利润表（百万元）	2023A	2024E	2025E	2026E
营业总收入	**709**	**1,116**	**3,588**	**5,412**
营业成本(含金融类)	219	335	1,259	2,108
税金及附加	4	5	16	27
销售费用	82	89	179	216
管理费用	154	179	269	298
研发费用	1,118	1,172	1,794	2,435
财务费用	(45)	(99)	(66)	(32)
加:其他收益	144	0	0	0
投资净收益	74	89	100	108
公允价值变动	0	0	0	0
减值损失	(272)	0	0	0
资产处置收益	0	0	0	0
营业利润	**(876)**	**(475)**	**237**	**468**
营业外净收支	1	0	0	0
利润总额	**(875)**	**(475)**	**237**	**468**
减:所得税	3	2	17	47
净利润	**(878)**	**(477)**	**221**	**421**
减:少数股东损益	(30)	(16)	7	14
归属母公司净利润	**(848)**	**(461)**	**213**	**407**
每股收益-最新股本摊薄(元)	(2.03)	(1.10)	0.51	0.97
EBIT	(982)	(574)	171	435
EBITDA	(652)	(348)	371	622
毛利率(%)	69.16	69.98	64.91	61.04
归母净利率(%)	(119.60)	(41.33)	5.94	7.52
收入增长率(%)	(2.70)	57.30	221.51	50.84
归母净利润增长率(%)	32.47	45.65	146.21	90.88

现金流量表（百万元）	2023A	2024E	2025E	2026E
经营活动现金流	(596)	(1,655)	(1,514)	(829)
投资活动现金流	425	(183)	(95)	(77)
筹资活动现金流	1,657	541	243	200
现金净增加额	1,486	(1,297)	(1,366)	(706)
折旧和摊销	330	227	200	186
资本开支	(100)	(215)	(195)	(185)
营运资本变动	(249)	(1,315)	(1,834)	(1,328)

重要财务与估值指标	2023A	2024E	2025E	2026E
每股净资产(元)	13.56	12.96	13.47	14.44
最新发行在外股份（百万股）	417	417	417	417
ROIC(%)	(18.24)	(10.23)	2.83	6.61
ROE-摊薄(%)	(15.02)	(8.53)	3.79	6.75
资产负债率(%)	10.73	17.54	27.76	30.87
P/E（现价&最新股本摊薄）	(388.89)	(715.46)	1,548.24	811.11
P/B（现价）	58.28	61.00	58.69	54.73

数据来源:Wind,东吴证券研究所，全文如无特殊注明，相关数据的货币单位均为人民币，预测均为东吴证券研究所预测。

请务必阅读正文之后的免责声明部分

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（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。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享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。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn