寒武纪（688256.SH）

2026 年 11 月 19 日

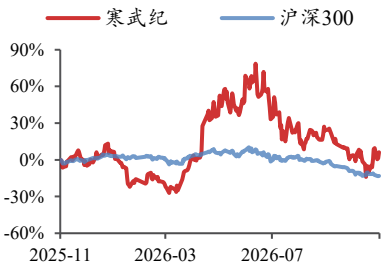
国产 AI 芯片龙头全栈布局，算力上行驱动业绩高增

——公司首次覆盖报告

投资评级：买入（首次）

日期	2026/11/18
当前股价(元)	949.00
一年最高最低(元)	1,966.00/750.01
总市值(亿元)	5,967.82
流通市值(亿元)	5,967.82
总股本(亿股)	6.29
流通股本(亿股)	6.29
近 3 个月换手率(%)	100.64

股价走势图



数据来源：聚源

张越（分析师）
zhangyue1@kysec.cn
证书编号：S0790524090003

叶进新（联系人）
yejinxin@kysec.cn
证书编号：S0790126060053

● **全栈自研巩固国产 AI 芯片龙头地位，算力上行驱动业绩拐点到来**

寒武纪专注于 AI 芯片领域十余年，能提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。公司产品广泛应用于服务器厂商和产业公司，面向互联网、金融、交通、能源、电力和制造等领域的复杂 AI 应用场景提供算力。随着 AI Agent 和多模态大模型的兴起，推理算力需求呈现出快速增长态势，公司业绩迎来拐点。2025 年公司实现营收 64.97 亿元，同比增长 453.12%，迎来营收规模的量级跨越，实现归母净利润 20.59 亿元，同比增长 555.24%，完成全年利润的扭亏为盈。考虑到公司业务战略聚焦效果显著，产品技术持续提升，叠加算力需求持续上涨，我们预计公司 2026-2028 年归母净利润分别为 71.40/155.89/254.23 亿元，对应 EPS 分别为 11.35/24.79/40.43 元，当前股价对应 PE 分别为 79.1/36.2/22.2 倍，首次覆盖给予“买入”评级。

● **软硬件深度协同构造核心壁垒，自主可控赋能国产替代加速**

寒武纪以自研 MLU 指令集为处理器生态基石，贯通云边端全系列芯片与基础软件，实现软硬件深度协同。依托自研基础软件平台，公司构建了独立于国外闭源体系的 AI 算力生态，摆脱了对国外闭源生态的依赖，筑牢国产算力安全底座。通过 Cambricon NeuWare 与 MagicMind 等核心软件栈，公司实现了从芯片硬件到系统软件的深度协同与垂直优化。这种软硬件一体化设计大幅降低了开发者的适配门槛与迁移成本，为国产 AI 算力替代提供了坚实的生态支撑。

● **持续推进软硬件适配和研发能力，全面引领国产算力新格局**

软件层面，寒武纪基于 NeuWare 生态实现 DeepSeek-V4 Day 0 首发适配并开源，通过 PyTorch/vLLM 原生支持、Triton 算子开发及 CNAgent 加速实现快速迁移；并以融合算子库、BangC、Kernel 及 5D 混合并行等技术释放极致性能。研发层面，公司 2025 年完成近 40 亿元定增，投向大模型芯片与软件平台，持续迭代新一代智能芯片及基础软件，增强大模型适应能力，巩固国产 AI 算力龙头地位。

● **风险提示：**技术迭代与生态建设不及预期；供应链与外部环境不稳定的风险。

财务摘要和估值指标

指标	2024A	2025A	2026E	2027E	2028E
营业收入(百万元)	1,174	6,497	17,008	36,007	60,007
YOY(%)	65.6	453.2	161.8	111.7	66.7
归母净利润(百万元)	-452	2,059	7,140	15,589	25,423
YOY(%)	46.7	555.2	246.7	118.3	63.1
毛利率(%)	56.7	55.2	55.9	56.9	55.0
净利率(%)	-38.5	31.7	42.0	43.3	42.4
ROE(%)	-8.4	17.4	38.9	46.3	43.2
EPS(摊薄/元)	-0.72	3.27	11.35	24.79	40.43
P/E(倍)	-1248.4	274.2	79.1	36.2	22.2
P/B(倍)	104.1	47.7	30.8	16.8	9.6

数据来源：聚源、开源证券研究所

目　录

1、 国产 AI 芯片领域龙头，全栈自研能力行业领先 4
1.1、 专注高性能通用 GPU 研发，四线产品覆盖主流算力场景 4
1.2、 公司股权结构集中稳定，高管团队产学研背景深厚 6
1.3、 受益国产 AI 算力需求快速增长，公司业绩拐点已至 8
2、 AI 算力市场供需格局重构，国产算力迎来广阔替代机遇 10
2.1、 AI 产业带动算力扩容，海外高端 GPU 供给受限打开国产替代空间 10
2.2、 推理算力需求结构性提升，国产芯片迎差异化破局机遇 12
2.3、 国内云厂商 Capex 提速，国产 AI 芯片营收高增可期 15
3、 云边端全栈算力纵深布局，寒武纪铸国产 AI 算力自主底座 16
3.1、 国产 AI 芯片竞争加剧，思元系列卡位训练卡市场前列 16
3.2、 自研架构与软硬件协同，构筑全栈自主技术底座 17
3.3、 供应链自主建设夯实长期供给基础，持续推进性能优化与算力发展 20
4、 盈利预测与投资建议 21
5、 风险提示 23
附：财务预测摘要 24

图表目录

图 1： 公司始终围绕“云边端”一体化布局与“软硬协同”的全栈技术能力展开发展 4
图 2： 公司产品线布局齐全，能够提供云边端一体、软硬件配套协同的全栈人工智能综合解决方案 5
图 3： 公司是国内少数具备从底层指令集、芯片微架构、AI 芯片到上层基础软件平台全栈自研能力的企业 6
图 4： 公司股权架构治理端权责清晰、结构稳定，重要股东具备国资背景 6
图 5： 2026 年 Q1-Q3 公司营收同比增长 106.76% 8
图 6： 2026 年 Q1-Q3 公司归母净利润同比增长 131.49% 8
图 7： 2025 年云端产品线成为公司总营收绝对核心支柱 9
图 8： 2025 年公司云端产品线营收同比增长 455.34% 9
图 9： 2021-2025 年公司费用率呈现持续优化态势 9
图 10： 公司研发投入占比始终保持在较高水平 9
图 11： 2025 年公司研发投入 11.69 亿元 9
图 12： 2025 年公司研发人员数量占总人数的 80.13% 9
图 13： 公司盈利能力持续改进 10
图 14： 公司盈利兑现弹性领先同业 10
图 15： 2026 年五大科技巨头 AI 资本开支将大幅增长 10
图 16： 从生成式人工智能到智能体的兴起，Token 调用量正呈现快速增长态势 11
图 17： 国内算力芯片企业，持续迭代云端智能计算产品，已逐渐形成完整产品体系 12
图 18： 在工程实践中，AI 算力的训练和推理几乎是两种完全不同的工作负载 12
图 19： AI 芯片产业正从通用 GPU 架构走向训练与推理双轨分化 13
图 20： 华为昇腾大 EP 集群推理方案，基于昇腾 AI 软硬件打造，适配 DeepSeek V3&R1 及业界主流 MoE 大模型 14
图 21： 端侧推理是 AI 芯片增速最快赛道，模型运行于手机、PC、汽车和 IoT 设备 14
图 22： 全球九大主流 CSP 2026 年资本支出预计达 8867 亿美元，同比增长约 90% 15
图 23： 思元 690 采用 7nm FinFET 制程 16
图 24： 690 采用 2.5D 硅中介层堆叠集成计算 Die 与 HBM 内存，实现数据高速互连 17

图 25： 寒武纪全面掌握核心芯片与系统软件的大量关键技术 17
图 26： 寒武纪 Cambricon NeuWare 是云边端训推一体的软件开发平台 19
图 27： 寒武纪推理加速引擎 MagicMind 兼顾性能、精度与易用性 19
图 28： 寒武纪在大模型训练与推理双端同步推进生态适配与性能优化 20
图 29： 公司募集资金用于面向大模型的芯片平台项目和面向大模型的软件平台项目 21

表 1： 核心管理层技术底色浓厚，产学研背景深厚、职能互补性强 7
表 2： 公司在云边端产品线的所有智能芯片和智能处理器核及基础系统软件均构建于自研的 MLU 指令集基础之上 18
表 3： 寒武纪盈利预测（亿元） 22
表 4： 可比公司盈利预测估值（截至 2026/11/16） 22

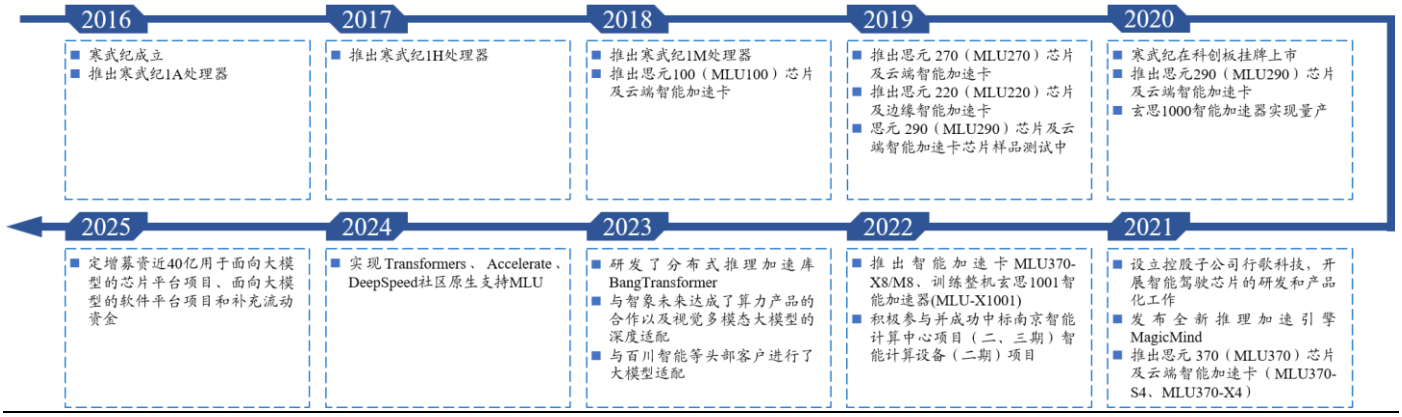
1、 国产 AI 芯片领域龙头，全栈自研能力行业领先

1.1、 专注高性能通用 GPU 研发，四线产品覆盖主流算力场景

寒武纪自成立以来一直专注于 AI 芯片产品的研发与技术创新，现已成为国内 AI 芯片领军企业。公司凭借领先的核心技术，较早实现了多项技术的产品化，能提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。产品广泛应用于服务器厂商和产业公司，面向互联网、金融、交通、能源、电力和制造等领域的复杂 AI 应用场景提供算力，推动人工智能赋能产业升级。公司掌握的智能处理器指令集、智能处理器微架构、智能芯片编程语言、智能芯片数学库等核心技术，具有壁垒高、研发难、应用广等特点，对集成电路行业与人工智能产业具有重要的技术价值、经济价值和生态价值。

公司始终围绕“云边端”一体化布局与“软硬协同”的全栈技术能力展开发展。自 2016 年 3 月成立以来，公司快速实现了技术的产业化输出，先后推出了用于终端场景的寒武纪 1A/1H/1M 系列芯片、基于思元 100/270 芯片的云端智能加速卡系列产品以及基于思元 220 芯片的边缘智能加速卡，构建了云端与边缘端产品体系；2020 年 7 月 20 日，公司在科创板成功上市，并在同年推出了思元 290 训练芯片和玄思 1000 智能加速器，补足人工智能训练产品线。至此，公司已初步建立“云边端一体、软硬件协同、训练推理融合”的新生态；2021 年公司发布了基于第四代智能处理器微架构（MLUarch03）的推训一体思元 370 智能芯片及加速卡，并投入了大量的资源优化基础系统软件平台，新一代产品及智能驾驶芯片的研发也在有序进行，公司“云边端车”协同发展的战略得到了全面推进。2022 至 2025 年，公司主要产品已与智象未来、百川智能等大模型客户完成全面适配，进入生成式 AI 算力供应链。

图1：公司始终围绕“云边端”一体化布局与“软硬协同”的全栈技术能力展开发展

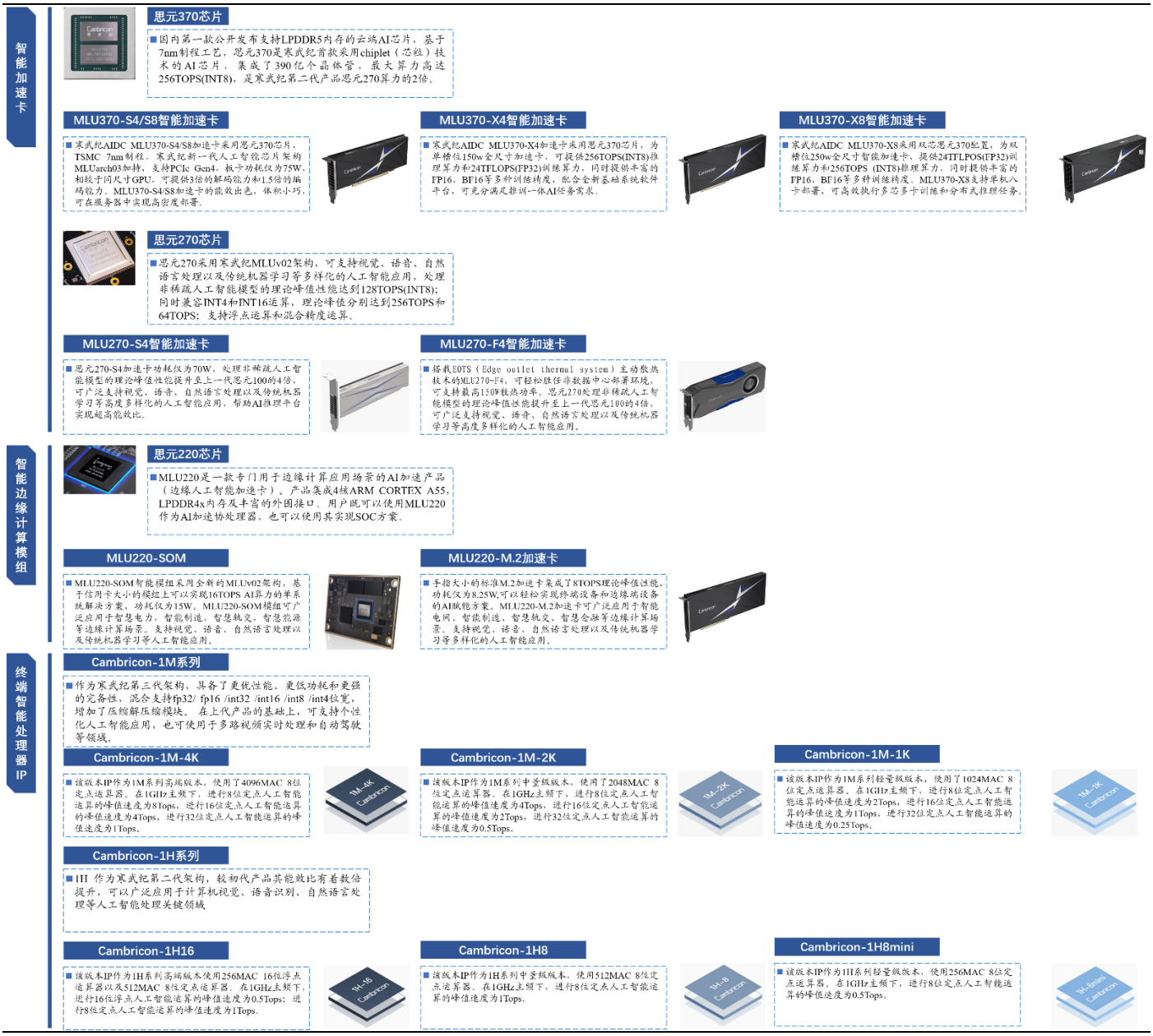


资料来源：公司公告、开源证券研究所

公司产品线布局齐全，能够提供云边端一体、软硬件配套协同的全栈人工智能综合解决方案。公司业务矩阵主要分为云端硬件、边缘端硬件、IP 授权与配套软件。云端算力产品线以思元 370 等芯片、配套加速卡与训练整机为核心，芯片具备多精度数据运算、大算力缓存与低功耗优势，可向云厂商供给高性能 AI 算力；边缘端主力产品为思元 220 边缘计算模组，可实现本地实时数据处理，大幅降低传输时延，广泛落地于物联网、智慧城市等场景。同时公司对外提供处理器 IP 授权，并搭建完整配套软件生态，硬件产品覆盖 AI 算力集群、自动驾驶、边缘计算、AIoT 终端等全类应用场景。公司是国内少数具备从底层指令集（MLU 指令集）、芯片微架构（MLUarch）、AI 芯片（思元系列）到上层基础软件平台（NeuWare）全栈自研能力

的企业。

图2：公司产品线布局齐全，能够提供云边端一体、软硬件配套协同的全栈人工智能综合解决方案



资料来源：寒武纪官网、开源证券研究所

图3：公司是国内少数具备从底层指令集、芯片微架构、AI 芯片到上层基础软件平台全栈自研能力的企业

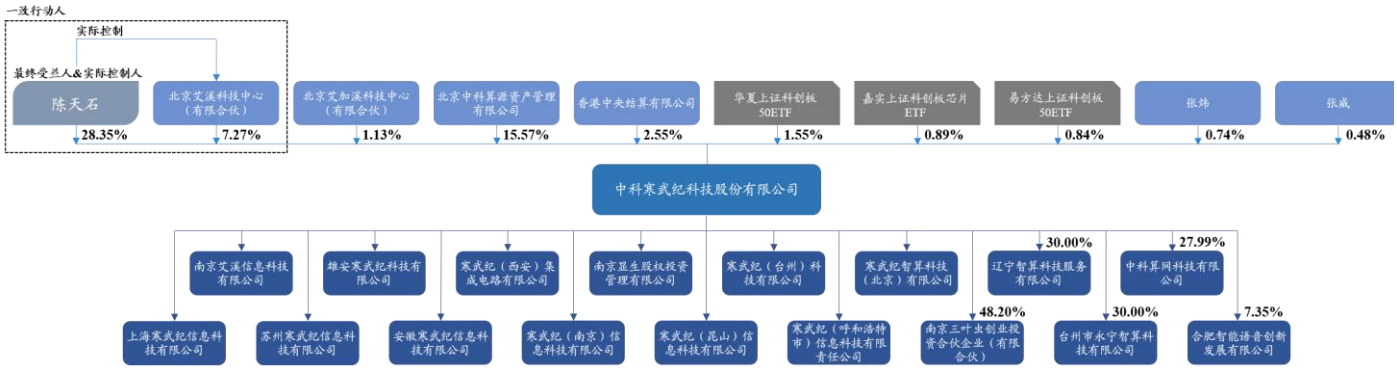


资料来源：寒武纪官网、开源证券研究所

1.2、 公司股权结构集中稳定，高管团队产学研背景深厚

公司股权架构治理端权责清晰、结构稳定，重要股东具备国资背景。陈天石为公司最终受益人、实际控制人，直接持有公司 28.35%股份，为第一大股东。北京艾溪科技中心（有限合伙）为核心员工持股平台，持股 7.27%，由陈天石担任执行事务合伙人，可完整支配该平台的股份表决权，与陈天石构成一致行动人。二者合并持有公司 35.62%的股份及对应表决权，控制权集中且稳固，能够保障公司技术路线与经营战略的长期连贯性。此外，公司第二大股东北京中科算源资产管理有限公司是由中国科学院计算技术研究所 100%全资控股，属于中科院体系下属的国有独资资产管理平台，同时也是寒武纪的创始股东，为公司提供技术研发、产业资源等层面的战略背书。

图4：公司股权架构治理端权责清晰、结构稳定，重要股东具备国资背景



资料来源：Wind、开源证券研究所

公司管理团队由创始人主导，技术底色浓厚、职能互补性强。创始人陈天石为中科大计算机博士，曾在中科院计算所担任研究员、博士生导师，深耕人工智能和处理器芯片等相关领域二十余年，积累了坚实的理论功底和丰富的研发经验。目前陈天石在公司担任董事长、总经理，同时为公司核心技术人员，集战略决策权与经营执行权于一身。核心副总经理团队几乎全部为技术出身，学历与从业背景高度硬核，适配 AI 芯片行业需要专业人员长期高强度研发投入的行业属性。整体管理团队既具备底层架构研发能力，又拥有芯片商业化落地经验，形成了“科研能力+产业能力”的高度互补。

表1：核心管理层技术底色浓厚，产学研背景深厚、职能互补性强

姓名	职位	任职经历
陈天石	董事长、总经理、核心技术人员	陈天石，男，出生于 1985 年，中国科学技术大学计算机软件与理论博士学历。中国国籍，无境外永久居留权。2010 年至 2019 年，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 年创立寒武纪，现任寒武纪第三届董事会董事长、总经理、核心技术人员。
金晓光	董事	金晓光，男，出生于 1982 年，莱斯特大学管理学硕士学历。中国国籍，无境外永久居留权。2009 年至今，就职于中国科技产业投资管理有限公司，现任董事总经理、AI 基础设施组组长、总经理助理。现任寒武纪第三届董事会董事。
刘新宇	董事	刘新宇，男，出生于 1972 年，哈尔滨工业大学博士学历。中国国籍，无境外永久居留权。2008 年至今，历任中国科学院计算技术研究所技术发展处副处长、技术发展处处长、技术发展中心主任；2016 年至 2020 年，任北京中科算源资产管理有限公司总经理；2020 年至今，任北京中科算源资产管理有限公司执行董事、总经理；2017 年至今，任北京中科图灵基金管理有限公司董事长、总经理；2019 年至今，任中科苏州智能计算技术研究院院长；2019 年至今，任北京中科晶上科技股份有限公司副董事长；2020 年至今，任龙芯中科技术股份有限公司董事；2021 年至今，任中科天玑数据科技股份有限公司董事；2023 年至今，任中科驭数（北京）科技有限公司董事。现任寒武纪第三届董事会董事。
刘少礼	董事、副总经理、核心技术人员	刘少礼，男，出生于 1987 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，无境外永久居留权。2014 年至 2019 年，就职于中国科学院计算技术研究所并任副研究员。2016 年作为寒武纪创始团队成员加入寒武纪，现任寒武纪第三届董事会董事、副总经理、核心技术人员。
王在	董事、副总经理	王在，男，出生于 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，无境外永久居留权。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，就职于中国科学院计算技术研究所从事科研工作。2016 年作为寒武纪创始团队成员加入寒武纪，现任寒武纪第三届董事会职工代表董事、副总经理。
叶淏尹	董事、副总经理、财务负责人、董事会秘书	叶淏尹，女，出生于 1988 年，北京大学西方经济学硕士学历。中国国籍，无境外永久居留权。2012 年至 2016 年，就职于中国高新投资集团公司并任投资经理、高级投资经理；2016 年至 2019 年，就职于国投创业投资管理有限公司并任投资副总裁。2019 年加入寒武纪，现任寒武纪第三届董事会董事、副总经理、财务负责人及董事会秘书。
陈帅	副总经理、核心技术人员	陈帅，男，出生于 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，有境外永久居留权（已过期）。2014 年至 2015 年，就职于中国科学院计算技术研究所并任工程师；2015 年至 2016 年，就职于多伦多大学并任电

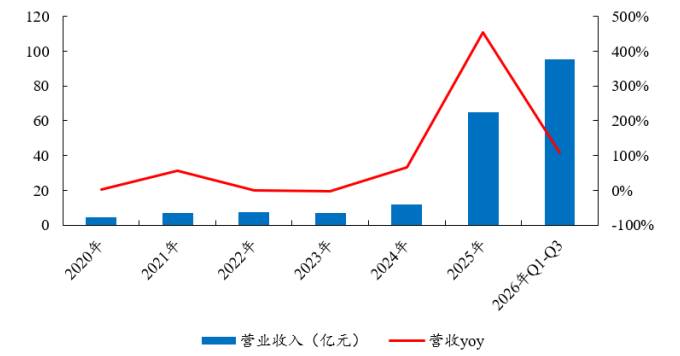
姓名	职位	任职经历
		子和计算机工程系博士后。2016 年至今，就职于寒武纪，现任寒武纪副总经理、核心技术人员。
刘毅	副总经理、核心技术人员	刘毅，男，出生于 1985 年，北京大学微电子与固体电子学硕士学历。中国国籍，无境外永久居留权。2010 年至 2012 年，就职于龙芯中科技术有限公司并任工程师；2012 年至 2016 年，就职于上海英伟达半导体（科技）有限公司并任高级工程师。2016 年至今，就职于寒武纪，现任寒武纪副总经理、核心技术人员。
张尧	副总经理、核心技术人员	张尧，男，出生于 1986 年，中国科学院计算技术研究所计算机系统结构硕士学历。中国国籍，无境外永久居留权。2012 年至 2014 年，就职于中国科学院计算技术研究所微处理器中心并任助理工程师；2014 年至 2015 年，就职于龙芯中科技术有限公司并任高级工程师；2015 年至 2016 年，就职于北京小米松果电子有限公司并任高级工程师。2016 年至今，就职于寒武纪，现任寒武纪副总经理、核心技术人员。

资料来源：公司公告、开源证券研究所

1.3、 受益国产 AI 算力需求快速增长，公司业绩拐点已至

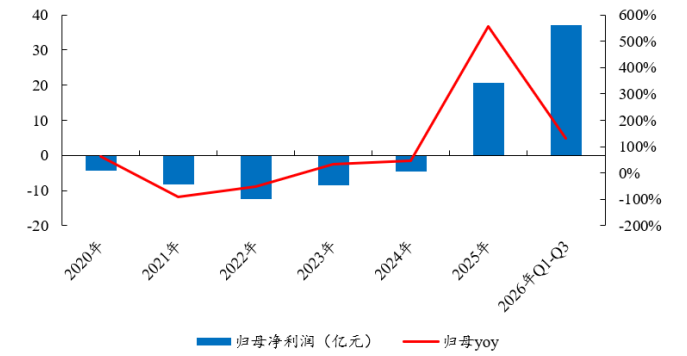
充分受益于行业需求增长和国产替代的双重红利，公司成功实现营收规模的突破和利润的扭亏为盈。随着 AIGC、大模型等技术在国内加速落地，AI 算力需求呈指数级增长，叠加国产替代进程加速推进，云端训练与推理芯片下游采购需求持续释放，公司凭借与头部互联网客户的深度绑定，依托核心产品的性能迭代与稳定交付能力，卡位核心应用场景，实现营收和利润的双升。2025 年公司实现营收 64.97 亿元，同比增长 453.21%，迎来营收规模的量级跨越，实现归母净利润 20.59 亿元，同比增长 555.24%，完成全年利润的扭亏为盈；2026 年前三季度公司实现营收 95.26 亿元，同比增长 106.76%，实现归母净利润 37.15 亿元，同比增长 131.49%。

图5：2026 年 Q1-Q3 公司营收同比增长 106.76%



数据来源：Wind、开源证券研究所

图6：2026 年 Q1-Q3 公司归母净利润同比增长 131.49%

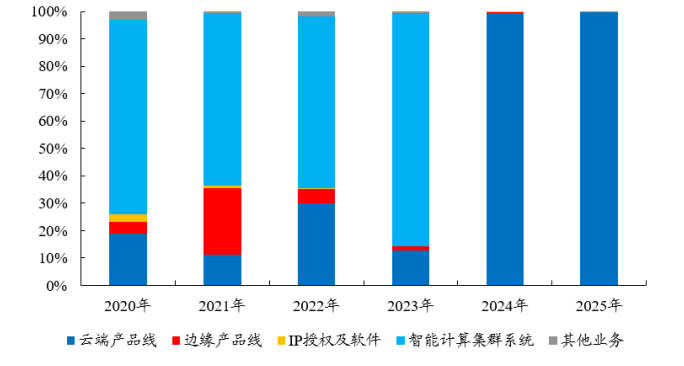


数据来源：Wind、开源证券研究所

产业结构持续优化，公司战略聚焦成效显著。公司早期核心收入来源于终端智能处理器 IP 授权，2020 年华为海思转向自研终端智能芯片、终止合作后，终端业务收入进入持续下行通道。为拓展成长空间、验证芯片落地能力并撬动地方智算中心市场，公司布局智能计算集群系统业务。2020-2023 年恰逢国内地方智算中心建设高峰期，多笔亿级大额项目集中交付并确认收入，推动该业务收入占比快速提升，成为阶段性核心营收支柱。2024 年以来，一方面地方智算整包项目建设节奏放缓，大额集中交付项目减少；另一方面公司战略逐步聚焦核心芯片主业，主动收缩低附加值的系统集成环节，智能计算集群业务收入规模随之自然回落。整体来看，公司营收结构持续优化，资源不断向高毛利的云端产品线倾斜，边缘端及终端业务占比逐

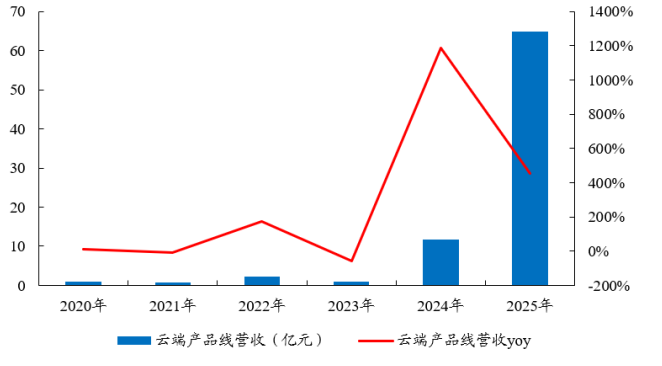
步收缩，战略聚焦与业务升级的成效持续显现。2025 年公司实现云端产品线营收 64.77 亿元，同比增长 455.34%，成为公司总营收的绝对核心支柱。

图7：2025 年云端产品线成为公司总营收绝对核心支柱



数据来源：Wind、开源证券研究所

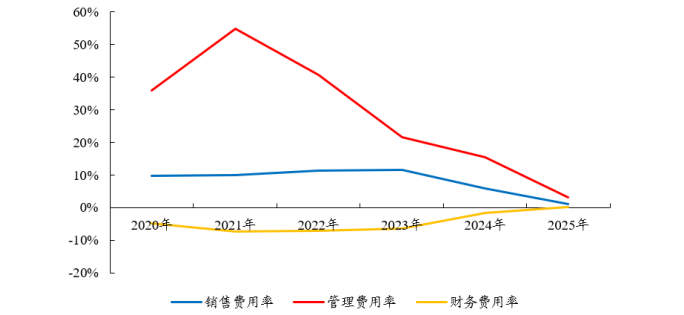
图8：2025 年公司云端产品线营收同比增长 455.34%



：Wind、开源证券研究所

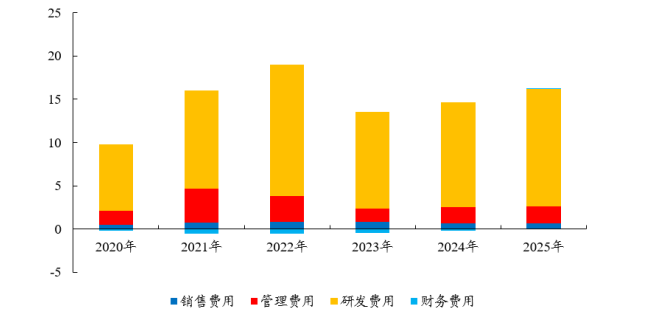
公司费用率总体持续优化，高研发投入筑牢技术壁垒。公司优化经营管控压降运营成本，强化采购管理压缩生产成本，费用率呈现持续优化态势。2025 年公司销售/管理/财务费用率分别为 1.05%、3.05%、0.17%，同比-4.92pct/-12.32pct/+1.81pct。公司始终保持高研发投入，聚焦人工智能芯片产品迭代升级，持续强化产品核心竞争力，夯实芯片技术根基。2025 年公司研发投入 11.69 亿元，占营收的比例为 17.99%；公司拥有 887 人的研发团队，占员工总人数的 80.13%，研发人才储备扎实，技术团队构成高度聚焦主业。

图9：2021-2025 年公司费用率呈现持续优化态势



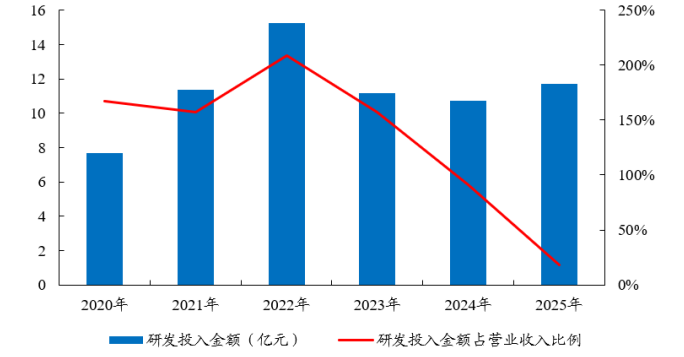
数据来源：Wind、开源证券研究所

图10：公司研发投入占比始终保持在较高水平



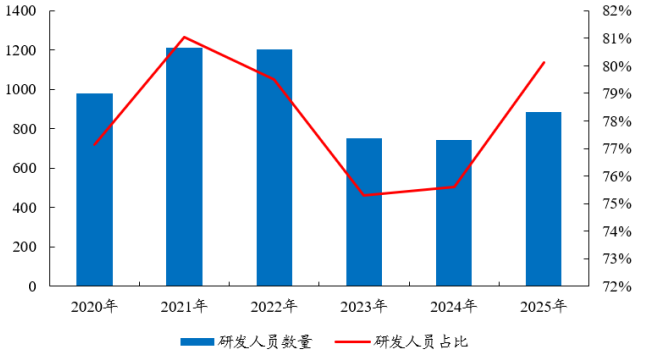
数据来源：Wind、开源证券研究所

图11：2025 年公司研发投入 11.69 亿元



数据来源：Wind、开源证券研究所

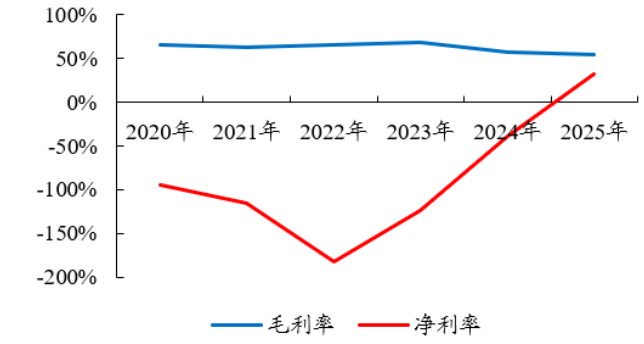
图12：2025 年公司研发人员数量占总人数的 80.13%



数据来源：Wind、开源证券研究所

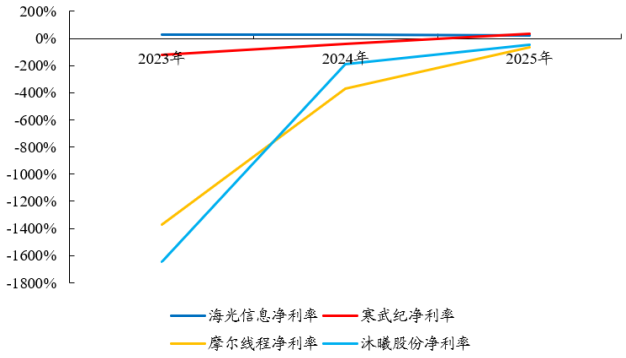
公司盈利能力持续改进，盈利兑现弹性领先同业。凭借核心技术构筑的深厚壁垒，公司毛利率长期稳定在高位，净利率随营收规模快速放量持续修复，实现扭亏为盈的关键业绩拐点。2025 年公司毛利率为 55.15%，同比-1.56pct，净利率为 31.68%，同比+70.59%。在国产 AI 算力芯片行业中，公司盈利兑现的节奏与质量均位居第一梯队，依托核心产品的规模化放量率先跑通完整商业化闭环，具备较为突出的业绩确定性与充足的长期成长弹性。

图13：公司盈利能力持续改进



数据来源：Wind、开源证券研究所

图14：公司盈利兑现弹性领先同业



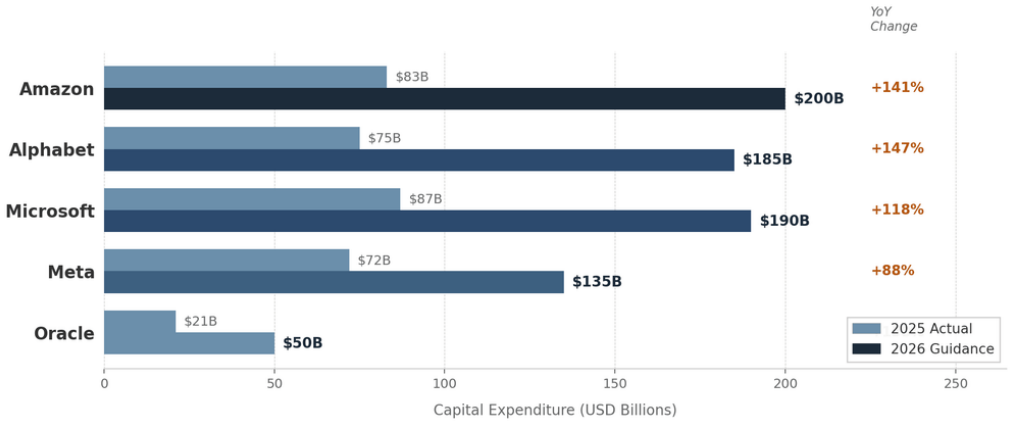
数据来源：Wind、开源证券研究所

2、 AI 算力市场供需格局重构，国产算力迎来广阔替代机遇

2.1、 AI 产业带动算力扩容，海外高端 GPU 供给受限打开国产替代空间

头部科技企业加码推理算力基建，资本开支大幅提升。随着人工智能服务进入常规部署阶段，行业开始从训练算力建设转型为推理算力建设，各大厂积极转向推理类 AI 基础设施投入，2024 年，国内推理工作负载已经超越训练端达到 65%。2022 年，亚马逊、Alphabet、微软、Meta 及甲骨文五大科技巨头资本开支达 1620 亿美元；2025 年，五家企业资本支出总额已攀升至 4480 亿美元，预计至 2030 年，全球 AI 资本支出将达到 1.3 万亿美元。

图15：2026 年五大科技巨头 AI 资本开支将大幅增长

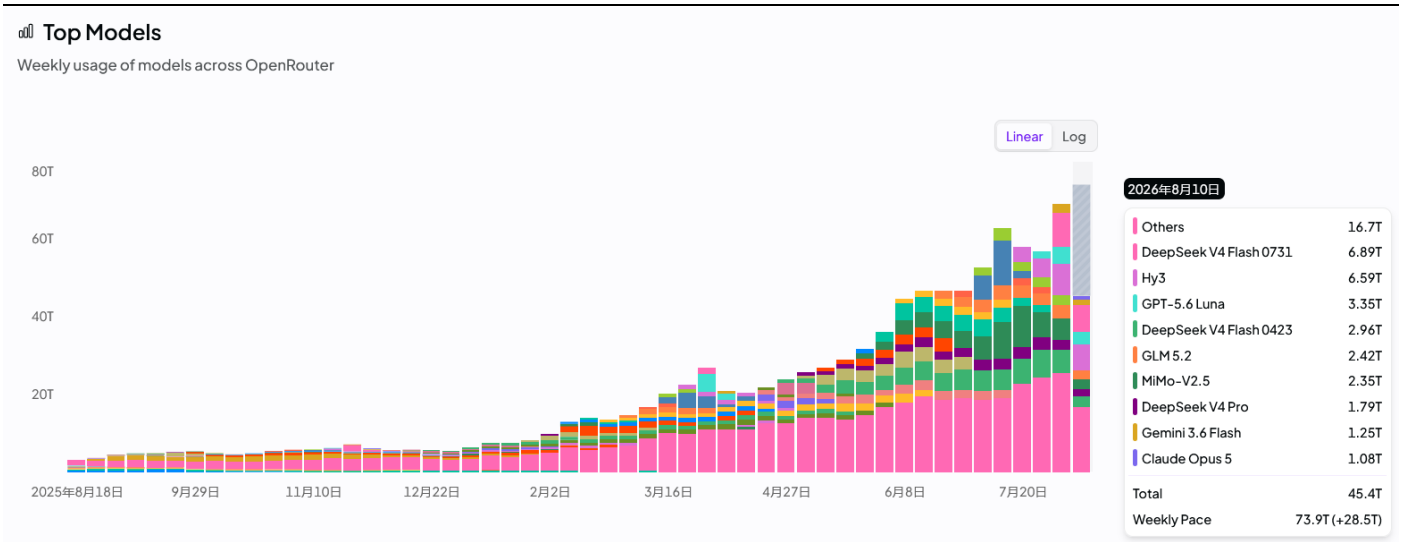


资料来源：paasa 官网

从生成式人工智能到智能体的兴起，推理算力需求正呈现快速增长态势。根据国家数据局披露的数据，2024 年初我国日均 Token 调用量约 1000 亿，2025 年底跃升至 100 万亿，2026 年 3 月进一步突破 140 万亿，两年间增长超过 1000 倍。国内

AI Infra 服务商硅基流动平台上的 Token 生成量、开发者活跃度以及企业客户数，均在 2026 年加速攀升，延续了加速增长态势。截至 2026 年 5 月，硅基流动的公有云及私有云 MaaS 平台，日均 Token 生成量已达数万亿量级，服务超过 1000 万用户和 1 万家企业客户，覆盖政企、互联网、金融、能源、交通、文娱等各行各业。

图16：从生成式人工智能到智能体的兴起，Token 调用量正呈现快速增长态势



资料来源：OpenRouter 官网

贸易政策管制叠加产能爬坡缓慢，算力供给缺口持续扩大。海外对华高端芯片出口管制持续加码，主流高性能算力卡进口渠道受限，稳定供货难以保障。叠加先进制程晶圆扩产周期漫长，产能爬坡进度偏缓，核心芯片产出增量有限。当前大模型研发及应用、智算集群落地带动算力需求稳步上行，供给增量释放速度不及需求增速，高性能算力卡整体供需缺口持续扩大。据中国信通院 2026 年 Q1 统计数据，国内高端智算算力整体缺口达 35%，仅英伟达 H100 单型号显卡市场缺口就达 43 万张，千卡级 GPU 批量采购订单的交付周期已延长至 12-15 个月，短期供给难以快速响应快速增长的 AI 训练与推理算力需求。

海外 GPU 供给受限催生了广阔的国产替代窗口期，国内自研通用算力芯片迎来发展机遇。先进工艺扩产周期长、设备与材料投入成本居高不下，代工提价直接抬升海外高端芯片的出货成本与交付周期，海外大厂产能、价格双重承压，供应链出现明显空档。以寒武纪为代表的国内算力芯片企业，持续迭代云端智能计算产品，性能逐步追赶国际一线水平。国内服务器与 AI 集群厂商出于供应链自主可控、保障算力交付稳定性的核心诉求，正加快国产算力芯片的适配验证与方案打磨，带动适配线上推理、中小模型微调的国产算力硬件落地节奏提速，持续填补国内高端算力供给缺口，推动国产算力产业生态逐步完善。

图17：国内算力芯片企业，持续迭代云端智能计算产品，已逐渐形成完整产品体系

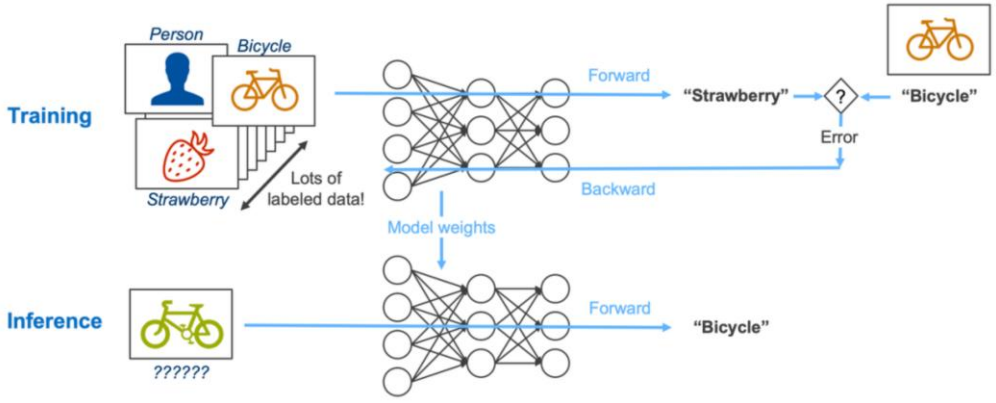


资料来源：海光信息公司公告、摩尔线程公司公告、沐曦股份公司公告、华为官网、开源证券研究所

2.2、 推理算力需求结构性提升，国产芯片迎差异化破局机遇

在工程实践中，AI 算力的训练和推理几乎是两种完全不同的工作负载。AI 训练的核心，是通过海量标注数据经算法迭代求解模型最优参数。该阶段需要持续数据输入、海量参数动态更新及数周至数月不间断运行，追求规模化运算效率。这意味着训练芯片不仅要具备强悍的算力，还需配备极高的显存带宽、高效的分布式通信能力，以及万卡级集群规模下的稳定性。推理则是模型基于已训练参数对新数据进行预测、生成响应的“运用知识”阶段，是 AI 落地的核心环节。相较训练，推理更侧重低延迟、高能效与成本控制，部署覆盖云、边、端全场景，无需迭代训练即可直接调用成熟模型，在实时响应场景中高效运行。

图18：在工程实践中，AI 算力的训练和推理几乎是两种完全不同的工作负载



资料来源：昊源诺信公众号

推理算力需求的快速增长正推动 AI 芯片产业从 GPU 主导走向训练和推理的分化路线。训练芯片需求解万亿维非凸优化问题，依赖 TB/s 级 HBM 带宽搬运海量权重与激活值，并通过 NVLink/NVSwitch 实现数百 GPU 域内全互联同步，跨域扩展则需 InfiniBand 等网络支撑，显存容量与互联带宽是核心瓶颈；推理芯片仅需执行单次前向传播，无需梯度计算与迭代训练，其优化重心在于以最低成本满足实时延迟，显存带宽往往比算力更关键，正如 Groq 的 LPU 芯片以 230MB 超大 SRAM 替代 HBM，将模型权重切分至数百张芯片的片上存储中使数据搬运延迟趋近于零，从而大幅提升推理速度。两者从架构需求到优化方向已呈现根本性差异，训练追求规模化算力与互联效率，推理则聚焦单位成本下的吞吐与延迟。

图19：AI 芯片产业正从通用 GPU 架构走向训练与推理双轨分化

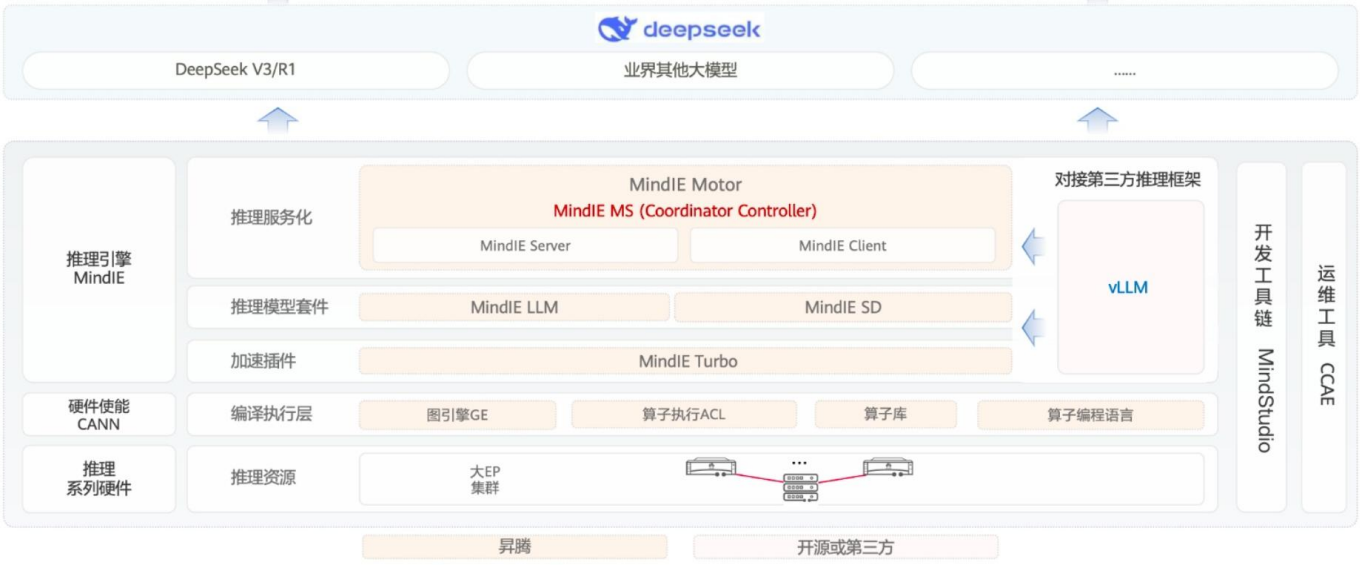


资料来源：驭势资本公众号、开源证券研究所

与训练端的高度集中不同，推理芯片市场正步入技术路线分化的新阶段。NVIDIA 训练芯片凭借 CUDA 生态、NCCL 通信库与主流 AI 框架的深度绑定所形成的软件粘性使其牢守 90%以上训练市场份额，AI 训练芯片市场呈现 NVIDIA 一家独大的格局。而推理芯片市场正在走向多元分化，NVIDIA L40S/GB200 通过精度降级进入推理，Groq LPU 以超大规模 SRAM 挑战 HBM 范式，AWS Inferentia 以每美元吞吐量最大化重塑 TCO，高通和苹果在端侧推理上通过 NPU 将 AI 推至移动设备，国内的寒武纪、地平线和华为昇腾在云端和边缘推理上构建国产替代。

云端推理芯片注重效率与成本，核心竞争维度是每美元吞吐量。NVIDIA L40S 基于 Ada Lovelace 架构，FP8 算力约 360 TFLOPS，350W 功耗，单卡价格约 1-1.5 万美元，两张在中小模型推理成本优于一张 H100。Groq LPU 以 230MB 片上 SRAM 替代 HBM，数据搬运延迟趋零，在 Llama 2-70B 上实现近 500 tokens/s，但单芯片容量成显存瓶颈。国产方面，华为昇腾超节点通过融合 kernel 与多流并行提升 DeepSeek V4 每美元吞吐量；寒武纪基于 NeuWare 与 vLLM 完成 DeepSeek-V4 适配，借助 PagedAttention 降低显存占用、提升并发，实现同等投入下更高每美元吞吐量。国产推理芯片正加速落地云服务商及企业级集群。

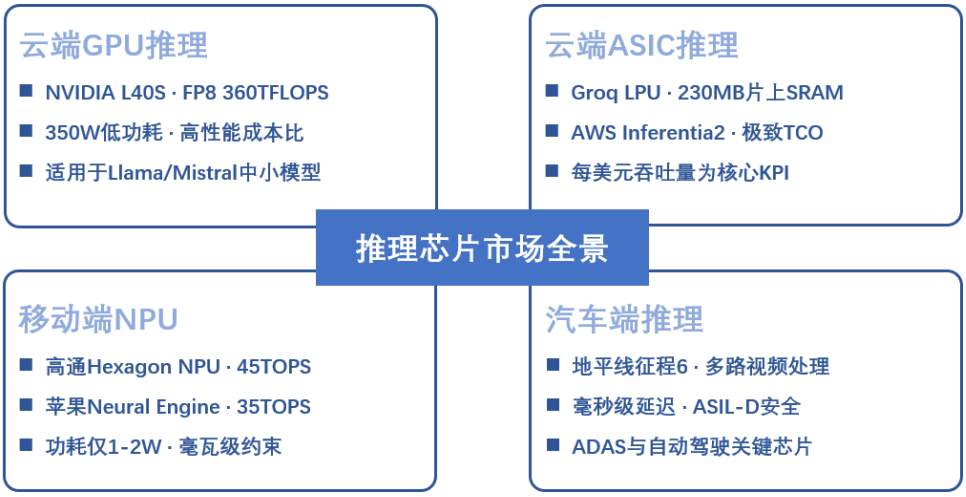
图20：华为昇腾大 EP 集群推理方案，基于昇腾 AI 软硬件打造，适配 DeepSeek V3&R1 及业界主流 MoE 大模型



资料来源：华为官网

端侧推理是 AI 芯片增速最快赛道，模型运行于手机、PC、汽车和 IoT 设备。端侧推理的核心约束与云端截然不同，其约束在于毫瓦级功耗、毫秒延迟，模型须量化压缩。NPU 是专为端侧设计的低功耗加速器——高通骁龙 8 Elite Hexagon NPU 支持 INT4、45TOPS，功耗 1-2W；苹果 A18 Pro Neural Engine 达 35TOPS，可运行 7B-13B 模型。端侧推理市场正在从手机向 PC 和汽车迅速扩展。英特尔的 Meteor Lake 首次在 CPU 中集成 NPU（Intel AI Boost），支持 Windows Copilot 的本地推理；AMD 的 Ryzen 8040 系列同样集成 Ryzen AI NPU，算力 16TOPS。国产化方面，华为昇腾、寒武纪等已切入边缘推理场景。在较为特殊的汽车端侧推理中，地平线的征程 6 芯片可以在毫秒级延迟内处理来自多路摄像头的视频流，运行目标检测、车道线识别和轨迹预测等多个 AI 模型，此类芯片对实时性和功能安全有车载级（ASIL-D）的严苛要求。隐私与低延迟为端侧推理的核心驱动力，随着模型量化技术与小模型蒸馏的发展，端侧有望成为未来五年空间最广的细分方向。

图21：端侧推理是 AI 芯片增速最快赛道，模型运行于手机、PC、汽车和 IoT 设备



资料来源：驭势资本公众号、开源证券研究所

推理芯片市场因生态依赖度低与场景碎片化而呈多元竞争格局，国产推理芯片迎来差异化破局机会。推理芯片市场的多元竞争根源于推理任务对生态锁定依赖较

低——推理软件栈相对简化，模型可通过 ONNX、OpenXLA 等标准化中间表示跨硬件迁移，无需深度绑定 CUDA。厂商只需提供高效推理引擎即可参与竞争，Groq LPU、AWS Inferentia、寒武纪思元等专用芯片得以在特定模型上实现更优性价比。推理场景极其多样（云端大模型、端侧 AI、自动驾驶、工业视觉），不同场景对延迟、功耗、成本要求各异，没有任何芯片能全维度最优。生态壁垒的弱化与垂直场景的差异化需求，为国产芯片企业绕过 CUDA 护城河、在边缘及云端推理细分领域实现国产替代提供了重要战略窗口。

2.3、 国内云厂商 Capex 提速，国产 AI 芯片营收高增可期

国内云厂商资本开支进入新一轮扩张周期，为国产 AI 芯片创造广阔增量空间。2026 年字节跳动大幅上调 AI 算力投入，资本开支计划突破 2000 亿元，较原计划增长超 25%，持续加码大模型训练与推理基础设施建设；阿里巴巴 2025 年资本支出达 859.72 亿元，未来三年云与 AI 基础设施投入将达 3800 亿至 4800 亿元，算力需求快速扩张。国内云厂商合计投入规模达数千亿级别，算力基础设施建设成为核心投向，叠加海外高端算力进口限制，直接拉动国产 AI 芯片采购需求，国产替代进程加速推进。

云厂商资本开支结构性向 AI 基础设施集中，GPU 与 AI 加速芯片为核心投向。中国云服务提供商正进入人工智能基础设施的新一轮投资周期。TrendForce 预测，字节跳动、腾讯、阿里巴巴和百度的资本支出总额在 2026 年将同比增长 80%以上。其中，字节跳动预计增幅最大，投资主要集中在大型人工智能数据中心、专用集成电路研发以及 GPU 集群部署上。全球九大主流 CSP 2026 年合计资本支出预计达 8867 亿美元，同比增长约 90%，投入主要集中于高性能计算硬件、数据中心扩建升级及网络基础设施优化三大领域。国内 AI 芯片厂商有望受益于本土算力建设进程，逐步提升市场份额。

图22：全球九大主流 CSP 2026 年资本支出预计达 8867 亿美元，同比增长约 90%



资料来源：TrendForce 官网

云厂商算力需求结构升级，驱动国产 AI 芯片价值量与渗透率双升。大模型参数规模持续扩张，训练集群从万卡向更大规模扩展，对芯片互联带宽与内存容量提出更高要求，带动单芯片价值量提升。同时，推理场景占比提升，部分国产芯片在特定精度下展现出能效比优势。云厂商 Capex 提速与算力结构升级形成共振，国产 AI 芯片厂商凭借客户绑定与产品迭代，营收规模与盈利能力有望实现跨越式增长。

3、 云边端全栈算力纵深布局，寒武纪铸国产 AI 算力自主底座

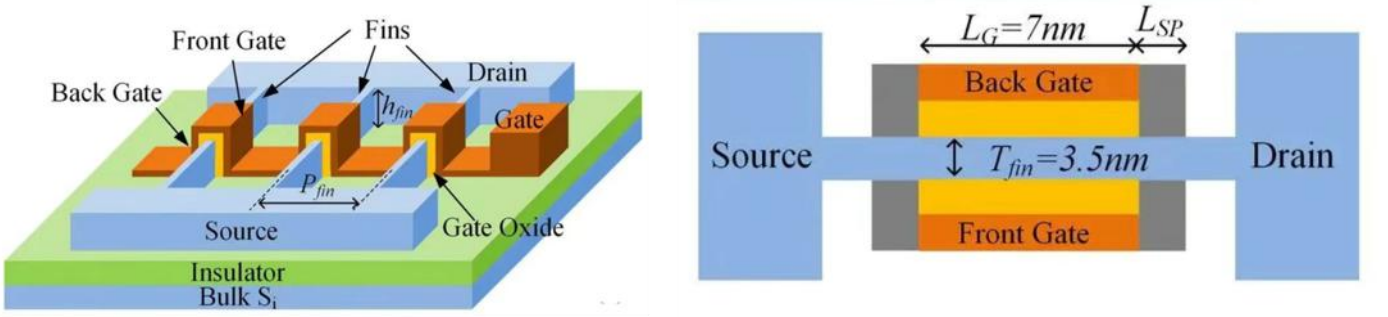
3.1、 国产 AI 芯片竞争加剧，思元系列卡位训练卡市场前列

思元 590 依托寒武纪自研的第五代 MLUarch 架构，构建起面向 AI 计算的差异化技术优势。该专用 ASIC 架构深度适配 AI 场景，集成大量矩阵乘法与向量运算单元，支持动态稀疏计算，稀疏模型推理能效比较传统架构提升 3 倍以上；支持 FP32、FP16、BF16 及 FP8 多精度计算，其中 UE8M0 FP8 格式可显著降低内存与带宽需求；动态推理引擎可随任务负载智能调节功率与算力分配，优化功耗管理；同时配备 1TB/s 带宽的 HBM2e 内存，并支持高速互联接口，为多芯片协同与大规模集群扩展奠定基础。

思元 590 算力性能已接近国际领先水平，卡位国产训练卡市场前列。其 FP16 精度算力达 256TFLOPS，部分场景达 512TOPS，整体算力达 A100 的 80%，图像分类、自然语言处理等任务可逼近 90%；FP8 推理效率较传统格式提升 300%，专家模块利用率由 30%升至 85%。能效方面，与 H100 同环境运行 BERT 模型，能效比超出预期 17%，训练时间缩短 12%，功耗降低 23%。横向对比，其 FP16 算力领先摩尔线程 MTT S4000（100TFLOPS），稳居国产训练卡前列。

思元 690 采用 7nm FinFET 制程与 Chiplet 异构封装方案，以架构革新告别单一大尺寸 Die 设计。随着 AI 芯片面积增大，传统单一大尺寸 Die 设计面临良率低、成本高、散热难三重制约，Chiplet 已成为全球高端 AI 芯片的主流路线。寒武纪将双计算 Die、单控制 Die 与单 IO Die 整合于同一芯片框架：两颗 7nm 同构计算 Die 精简自思元 590 规格、内置 64 个 MLU 核心，专职大模型训练与推理；低功耗控制 Die 统筹任务调度、时序同步与缓存一致性，跨 Die 数据延迟控制在 200 纳秒以内。

图23：思元 690 采用 7nm FinFET 制程



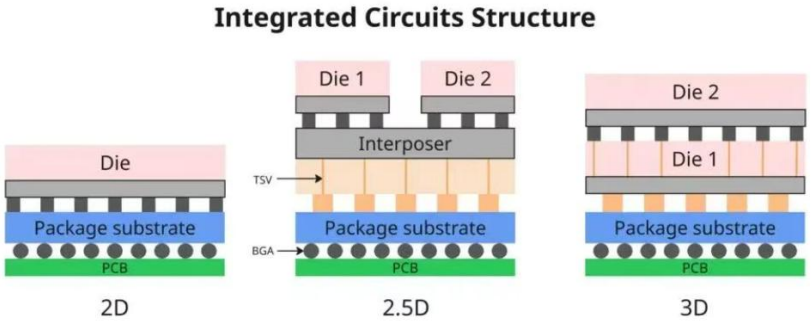
资料来源：Tiansong Cui、Bowen Chen《Layout Characterization and Power Density Analysis for Shorted-Gate and Independent-Gate 7nm FinFET Standard Cells》

在三 Die 分工协同下，思元 690 算力、带宽与集群扩展能力全面提升，综合竞争力实现代际跨越。算力端，双计算 Die 合计 128 个 MLU 核心，FP16 总算力突破 700 TFLOPS，性能直逼英伟达 H20；存储端，IO Die 外接 196GB HBM3 高带宽内存，带宽超 890GB/s；互联端，芯片支持 MLU-Link 对外互联，可轻松搭建 128 卡 AI 集群。2026 年初思元 690 正式量产，有望成为驱动寒武纪新一轮增长的核心引擎。

思元 690 依托 2.5D 先进封装与 196GB HBM3 显存突破存储瓶颈，单卡即可独立部署轻量化万亿参数大模型。芯片采用 2.5D 硅中介层堆叠集成计算 Die 与 HBM 内存，实现数据高速互连；HBM3 以 TSV 硅通孔垂直堆叠 DRAM，数据传输距离由

毫米级缩至微米级，配合 16 个独立通道实现单块最高 819GB/s、多块联用 4.8TB/s 的带宽。196GB 大显存使单卡无需多卡拆分即可承载万亿参数模型，大幅简化推理架构、降低通信延迟；超高带宽还可支撑百万 Token 超长上下文、批量并发及视频多模态并行计算，有效解决注意力机制的张量搬运拥堵。

图24：690 采用 2.5D 硅中介层堆叠集成计算 Die 与 HBM 内存，实现数据高速互连



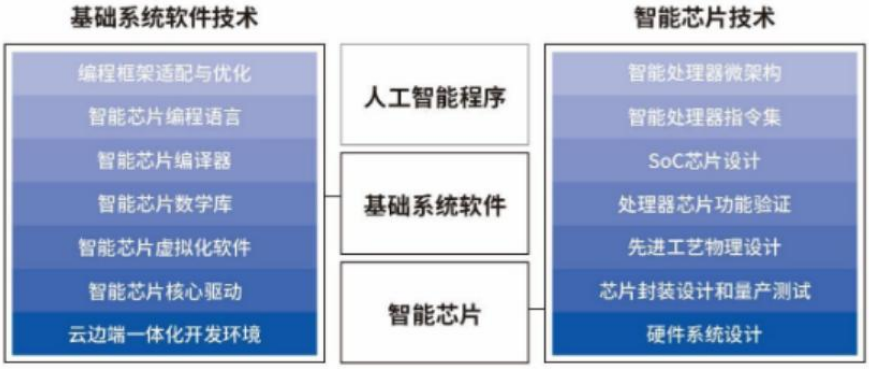
资料来源：HandWiki 官网

MLU-Link 高速互联与 NeuWare 软件平台协同构筑起寒武纪软硬一体的自研生态体系。互联端，MLU-Link 为寒武纪自研芯片间高速互联技术，经迭代升级已支持多卡协同训练，具备低延迟、高稳定、灵活组网特性，支撑千卡级集群搭建；生态端，MLU-Link 与思元算力芯片、NeuWare 全栈软件平台构成自研生态“铁三角”，NeuWare 兼容 PyTorch、TensorFlow 等主流框架并提供模型迁移工具，可实现 GPU 生态模型向 MLU 平台平滑迁移，降低客户迁移成本，强化生态护城河。

3.2、 自研架构与软硬件协同，构筑全栈自主技术底座

寒武纪坚持全栈自主研发，全面掌握核心芯片与系统软件的大量关键技术。寒武纪坚持训练推理融合、统一的基础软件平台研发策略，构建从自研芯片架构到高性能软件平台的完整体系，实现了计算架构、编译优化与算法调度的深度融合。公司在智能芯片领域掌握了智能处理器微架构、智能处理器指令集、SoC 芯片设计、处理器芯片功能验证、先进工艺物理设计、芯片封装设计与量产测试、硬件系统设计等七大类核心技术；在基础系统软件技术领域掌握了编程框架适配与优化、智能芯片编程语言、智能芯片编译器、智能芯片数学库、智能芯片虚拟化软件、智能芯片核心驱动、云边端一体化开发环境等七大类核心技术。硬件奠定算法并行计算与能效的物理基础，软件则依托编译优化、调度策略与模型适配，持续提升算力利用效率。

图25：寒武纪全面掌握核心芯片与系统软件的大量关键技术



资料来源：公司公告

寒武纪以自研 MLU 指令集为处理器生态基石，贯通云边端全系列芯片与基础软件，实现软硬件深度协同。公司能为云端、边缘端、终端提供全品类系列化智能芯片和处理器产品，是国内少数具备先进集成电路工艺下复杂芯片设计经验的企业之一，产品矩阵同时覆盖人工智能推理与训练全场景。公司也是国际上最早开展智能处理器指令集研发的少数几家企业之一。新一代商用智能处理器指令集正在研发中。公司已形成了体系完整、功能完备、高度灵活的智能芯片指令集专利群。公司在云端、边缘端、终端三条产品线的所有智能芯片和智能处理器核以及基础系统软件均构建于自研的 MLU 指令集基础之上。

表2：公司在云边端产品线的所有智能芯片和智能处理器核及基础系统软件均构建于自研的 MLU 指令集基础之上

技术大类名称	在主营业务及主要产品中的应用和贡献情况	专利或其他技术保护措施	成熟程度	技术来源
智能处理器微架构	新一代智能处理器微架构正在研发中。公司是国内外在该技术方向积累最深厚的企业之一。公司在云端、边缘端、终端三条产品线的所有智能芯片和智能处理器核均基于自研处理器架构研制。	已取得专利 693 项（其中境外专利 239 项），报告期内增加 84 项。	成熟稳定	自主研发
智能处理器指令集	指令集是处理器芯片生态的基石。公司是国际上最早开展智能处理器指令集研发的少数几家企业之一。新一代商用智能处理器指令集正在研发中。公司已形成了体系完整、功能完备、高度灵活的智能芯片指令集专利群。公司在云端、边缘端、终端三条产品线的所有智能芯片和智能处理器核以及基础系统软件均构建于自研的 MLU 指令集基础之上。	已取得专利 233 项（其中境外专利 61 项），报告期内增加 8 项。	成熟稳定	自主研发
SoC 芯片设计	公司已掌握复杂 SoC 设计的一系列关键技术，有力支撑了云端大型 SoC 芯片（思元 100、思元 270、思元 370 和思元 290）和边缘端中型 SoC 芯片（思元 220）的研发。	已取得专利 93 项（其中境外专利 32 项），报告期内增加 13 项。	成熟稳定	自主研发
处理器芯片功能验证	公司拥有成熟先进的处理器和 SoC 芯片功能验证平台，确保了智能处理器和 SoC 芯片逻辑设计按时高质量交付，有效保障了多款芯片产品的一次性流片成功。	已取得专利 15 项，报告期内增加 1 项。	成熟稳定	自主研发
先进工艺物理设计	公司已掌握 7nm 等先进工艺下开展复杂芯片物理设计的一系列关键技术，已将其成功应用于思元 100、思元 220、思元 270、思元 290、思元 370 等多款芯片的物理设计中。	非专利技术。	成熟稳定	自主研发
芯片封装设计与量产测试	应用于公司云端、边缘端不同品类芯片产品的封装设计与量产测试过程，有效支撑了公司处理器芯片的研发。	已取得专利 19 项（其中境外专利 7 项），报告期内增加 9 项。	成熟稳定	自主研发
硬件系统设计	有效解决了高速信号完整性、大功率供电下的电源完整性、大型芯片散热、机箱模块化等关键问题，支撑公司基于自研芯片研发模组/智能板卡、整机、集群等多样化的产品形态。	已取得专利 67 项（其中境外专利 13 项），报告期内增加 8 项。	成熟稳定	自主研发

资料来源：公司公告、开源证券研究所

公司依托自研基础系统软件平台，为云端、边缘端、终端全系列智能芯片提供统一编程接口与开发环境，彻底打破云边端之间的开发壁垒。Cambricon NeuWare 是寒武纪面向云边端全系列智能处理器打造的软件开发平台，采用云边端一体、训推一体架构，以统一的软件接口与完备生态支撑智能应用在全系列产品间的便捷开发、无缝迁移与高效调优；寒武纪推理加速引擎 MagicMind 是业界首个基于 MLIR 图编译技术达到商业化部署能力的推理引擎，支持跨框架的模型解析、自动后端代

码生成及优化。MagicMind 兼顾性能、精度与易用性，以简洁接口降低底层开发门槛，支持模型快速部署，并以插件化设计满足差异化需求。基于公司的基础系统软件平台，同一人工智能应用仅需简单移植即可在不同芯片上便捷高效运行，兼具灵活性与可扩展性，形成了从芯片硬件到系统软件的深度协同优势。

图26：寒武纪 Cambricon NeuWare 是云边端训推一体的软件开发平台



资料来源：寒武纪官网

图27：寒武纪推理加速引擎 MagicMind 兼顾性能、精度与易用性



资料来源：寒武纪官网

3.3、 供应链自主建设夯实长期供给基础，持续推进性能优化与算力发展

寒武纪采用 Fabless 轻资产运营模式，正逐步构建覆盖晶圆制造、先进封装、核心基材的全链条国产化供应链体系，高端 AI 芯片供应链自主可控能力持续强化。公司专注于智能芯片研发设计，将晶圆制造、封装测试等生产环节委托外部专业厂商代工，受美国实体清单管制约束，我们预计 MLU590、MLU690 等高端云端训练芯片正逐步切换至国内代工体系，依托中芯国际 N+2 等效 7nm 制程流片量产。封测环节以长电科技等公司为核心供应商，其生产基地 2.5D/Chiplet 先进封装产线可匹配 HBM 高带宽内存合封需求。

受国内先进制程产能约束与外部管制持续影响，寒武纪高端芯片供应链仍面临产能爬坡、良率提升与核心物料供给三重压力，供应链弹性建设为后续经营核心主线。当前中芯国际 N+2 产线整体处于高负荷运转状态，高端大算力芯片良率仍处于逐步爬升阶段，叠加 HBM 高带宽内存国产替代方案正处于发展状态，MLU590、MLU690 的交付节奏与产能规模显著受上游供给约束。公司持续通过加大战略备货、拓展多供应商体系、推进工艺适配优化等方式对冲供给风险。

寒武纪在大模型训练与推理双端同步推进生态适配与性能优化。训练侧重点支持 DeepSeek V3/V3.1、Qwen 系列等 MoE 模型及 GLM4.5、Wan、Hunyuan-Video 等主流架构，并基于原生 FP8 算力实现 Qwen/DeepSeek 等网络的高精度训练；推理侧则布局 W4A4、MX-FP8/MX-FP4 等新型数据类型，引入 Sparse Attention 与 Linear Attention 等高效注意力机制，同时覆盖 Qwen-Omni、Hunyuan3D、CosyVoice 及 DLM/VLM 等多模态与新兴架构。生态响应上，寒武纪实现 DeepSeek V3.2-Exp 发布即适配并同步开源代码；引擎层面持续优化 vLLM，完善混合精度低比特量化，支持通算并行、PD 分离、类 IBGDA 低时延大规模专家并行及 Torch.compile 主机侧优化，并专项攻坚长序列与超低解码延时场景，确保对主流开源模型的性能领先与全方位加速。

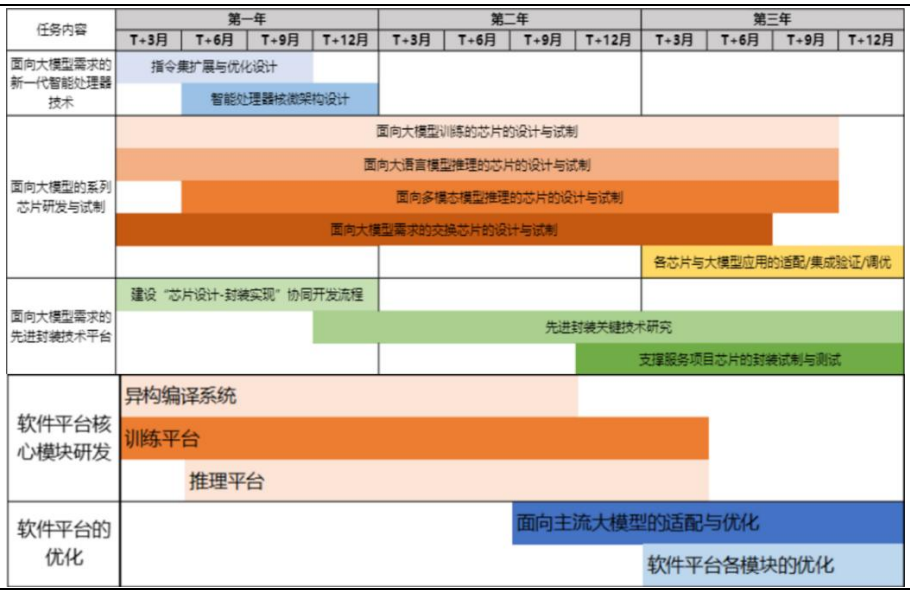
图28：寒武纪在大模型训练与推理双端同步推进生态适配与性能优化



资料来源：寒武纪开发者社区

公司持续提升芯片研发设计能力、技术储备和产品能力。2025 年 9 月公司完成定向增发，计划募集资金总额约 39.85 亿元，用于面向大模型的芯片平台项目、面向大模型的软件平台项目和补充流动资金。投资项目围绕大模型需求的多样化，研发新一代的智能芯片技术和相关产品，或将提升公司在大模型演进趋势下的技术和产品综合实力。在软件方面，开展相应的优化策略、软件算法以及软件工具的创新研究，并建设面向大模型的软件平台，支撑公司智能芯片算力性能的充分发挥，增强公司智能芯片对大模型新技术趋势和新应用拓展的灵活适应能力，或将有效提升面向大模型算法开发和应用部署的高效支撑与服务能力。

图29：公司募集资金用于面向大模型的芯片平台项目和面向大模型的软件平台项目



资料来源：公司公告

4、 盈利预测与投资建议

寒武纪是智能芯片领域全球知名的新兴公司，能提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。公司掌握的智能处理器指令集、智能处理器微架构、智能芯片编程语言、智能芯片数学库等核心技术，具有壁垒高、研发难、应用广等特点，对集成电路行业与人工智能产业具有重要的技术价值、经济价值和生态价值。依托于公司在人工智能芯片产品、基础软件平台、集群软件工具链方面取得的长足进步，公司产品在运营商、金融、互联网等多个重点行业规模化部署，通过了客户严苛环境的验证，产品普适性、稳定性、易用性获得了客户的广泛认可。

根据我们的核心假设：（1）公司营业收入主要来自于云端 AI 芯片业务，国内 AI 资本开支持续扩张带动大模型训练、推理算力需求高企，进一步拉动国产高端 AI 加速卡采购需求，公司作为国内云端算力芯片核心供给厂商，深度受益于算力国产化与智算基础设施建设浪潮，对此我们预期 2026 年将会成为公司云端芯片业务的放量阶段，收入增速迎来大幅提升；（2）公司 AI 芯片客户覆盖国内外智算中心、互联网科技企业等多元主体，综合来看，公司高端 AI 芯片产品毛利率显著高于中低端产品及通用算力产品。公司毛利率稳步提升主要源于核心产品持续迭代升级，新一代高端 AI 芯片性能、能效优势突出，产品附加值大幅提高，同时核心产品规模化放量摊薄单位研发及生产成本，叠加高毛利产品收入占比持续提升，带动公司整体毛利率水平稳步上行，我们预计公司 2026-2028 年营业收入分别为 170.08/360.07/600.07

亿元；总营收增长率分别为 161.77%/111.71%/66.65%，综合毛利率分别为 55.88%/56.89%/55.00%。

表3：寒武纪盈利预测（亿元）

	2025	2026E	2027E	2028E
营业收入	64.97	170.08	360.07	600.07
同比增速	453.21%	161.77%	111.71%	66.65%
毛利率	55.15%	55.88%	56.89%	55.00%
云端产品线收入	64.77	170.00	360.00	600.00
毛利率	55.22%	55.88%	56.89%	55.00%
边缘产品线收入	0.03	0.07	0.06	0.05
毛利率	44.52%	44.52%	44.52%	44.52%
IP 授权及软件产品线收入	0.02	0.01	0.01	0.01
毛利率	100.00%	100.00%	100.00%	100.00%

数据来源：Wind、开源证券研究所

寒武纪的主营收入主要来自于半导体产品智能芯片板块，因此可比公司主要选取海光信息、澜起科技、瑞芯微。理由如下：（1）海光信息。海光信息专注高端处理器、加速器等计算芯片及系统研发，产品包括海光通用处理器（CPU）和海光协处理器（DCU）。公司处理器兼容主流 x86 指令集，应用生态成熟丰富，搭载专用安全硬件，具备完备安全防护能力，通过权威机构安全测试。产品广泛覆盖企业计算、云计算、AI、边缘计算等场景，服务互联网、金融、能源等多行业。（2）澜起科技。澜起科技是一家全球领先的无晶圆厂集成电路设计公司，致力于为云计算及 AI 基础设施提供创新、可靠及高能效的互连解决方案。公司拥有互连类芯片和津逮服务器平台两大产品线，（3）瑞芯微。瑞芯微专注于集成电路设计与研发，已发展为领先的物联网（IoT）及人工智能物联网（AIoT）处理器芯片企业。公司以市场为导向，技术创新为核心，致力于为客户提供多层次、多平台、多场景的专业解决方案，赋能汽车电子、机器视觉、工业应用、教育办公、商业金融、智能家居以及消费电子等多元领域。

寒武纪 AI 算力芯片环节受益于大模型资本开支景气度高企及国产替代加速，公司作为国内云端 AI 芯片龙头有望充分受益于 AI 产业算力需求增加。从可比公司估值对比看，公司当期估值与可比公司基本相当，在盈利释放后则转为明显折价。折价源于盈利增速差，公司 EPS 复合增速明显高于可比公司，估值溢价具备业绩支撑，看好公司长期发展潜力。

表4：可比公司盈利预测估值（截至 2026/11/16）

证券代码	证券简称	收盘价（元/股）	EPS（元）			PE		
			2026E	2027E	2028E	2026E	2027E	2028E
688041.SH	海光信息	190.00	1.91	2.97	4.42	99.5	63.9	43.0
688008.SH	澜起科技	139.79	3.29	4.46	5.97	42.5	31.4	23.4
603893.SH	瑞芯微	208.60	4.06	5.27	6.67	51.4	39.6	31.3
	平均					64.5	44.9	32.6
688256.SH	寒武纪	897.97	11.35	24.79	40.43	79.1	36.2	22.2

数据来源：Wind、开源证券研究所（注：海光信息、澜起科技、瑞芯微盈利预测来自于 Wind 一致预期，寒武纪盈利预测来自于开源证券研究所）

我们预计 2026-2028 年归母净利润分别为 71.40/155.89/254.23 亿元，对应 EPS 分别为 11.35/24.79/40.43 元/股，当前股价对应 2026-2028 年的 PE 分别为 79.1/36.2/22.2 倍，基于公司先进的 AI 芯片技术，优质的算力客户资源，首次覆盖给予“买入”评级。

5、 风险提示

技术迭代与生态建设不及预期的风险。AI 芯片属于技术密集型赛道，制程演进与架构创新节奏持续加快，行业整体迭代周期较短。公司虽在国产智能芯片领域已形成差异化技术积累，但与国际龙头相比，在先进制程、单卡峰值算力及软件生态成熟度等维度仍存在追赶空间。若后续研发投入未能匹配技术演进速度，或新一代产品迭代进度不及预期，将削弱公司产品的市场渗透能力与长期竞争壁垒。

供应链与外部环境不稳定的风险。公司采用 Fabless 模式运营，在晶圆代工、封装测试等环节对特定供应商存在一定依赖。目前，公司及部分子公司已受到相关贸易管制政策波及；在全球半导体供应链深度重构的背景下，若外部限制措施进一步收紧，公司产能排期、工艺升级及物料交付的稳定性将面临不确定性。

附：财务预测摘要

资产负债表(百万元)	2024A	2025A	2026E	2027E	2028E
流动资产	5800	12080	21522	36840	66946
现金	1986	1317	2332	1800	5468
应收票据及应收账款	314	671	690	510	902
其他应收款	19	16	26	54	88
预付账款	774	745	2657	3890	8112
存货	1774	4944	13816	27579	49572
其他流动资产	932	4388	2000	3007	2805
非流动资产	917	1357	1839	2895	4073
长期投资	247	300	290	279	265
固定资产	231	373	844	1638	2513
无形资产	183	189	199	213	228
其他非流动资产	257	496	505	765	1068
资产总计	6718	13438	23361	39735	71019
流动负债	818	1333	4747	5770	11924
短期借款	100	0	0	51	0
应付票据及应付账款	515	1116	4512	5467	11635
其他流动负债	203	217	235	252	290
非流动负债	469	261	262	262	262
长期借款	0	0	0	0	0
其他非流动负债	469	261	262	262	262
负债合计	1287	1594	5010	6032	12186
少数股东权益	8	7	7	7	7
股本	417	422	628	628	628
资本公积	9809	11270	11064	11064	11064
留存收益	-4673	164	6756	20848	43221
归属母公司股东权益	5423	11836	18344	33697	58826
负债和股东权益	6718	13438	23361	39735	71019

现金流量表(百万元)	2024A	2025A	2026E	2027E	2028E
经营活动现金流	-1618	-498	-5	1987	5483
净利润	-457	2058	7140	15589	25423
折旧摊销	178	262	202	360	619
财务费用	-19	11	-55	-61	-108
投资损失	-23	-9	-32	-20	-15
营运资金变动	-1443	-3258	-7259	-13878	-20434
其他经营现金流	146	436	-1	-2	-2
投资活动现金流	-412	-4530	1611	-2395	-1578
资本支出	366	569	799	1415	1784
长期投资	-92	-3963	9	12	14
其他投资现金流	46	2	2400	-992	191
筹资活动现金流	48	3986	-591	-175	-186
短期借款	100	-100	0	51	-51
长期借款	0	0	0	0	0
普通股增加	1	4	207	0	0
资本公积增加	232	1461	-207	0	0
其他筹资现金流	-285	2620	-591	-226	-135
现金净增加额	-1982	-1043	1015	-583	3718

利润表(百万元)	2024A	2025A	2026E	2027E	2028E
营业收入	1174	6497	17008	36007	60007
营业成本	508	2914	7504	15523	27003
营业税金及附加	3	26	136	288	480
营业费用	70	68	85	162	270
管理费用	181	198	255	432	600
研发费用	1216	1351	2211	4321	6601
财务费用	-19	11	-55	-61	-108
资产减值损失	-53	-142	0	0	0
其他收益	220	307	236	227	248
公允价值变动收益	-0	5	1	2	2
投资净收益	23	9	32	20	15
资产处置收益	1	0	0	0	0
营业利润	-456	2061	7141	15590	25426
营业外收入	1	0	1	1	1
营业外支出	1	2	1	1	1
利润总额	-456	2059	7141	15590	25426
所得税	1	1	1	2	3
净利润	-457	2058	7140	15589	25423
少数股东损益	-5	-1	0	0	0
归属母公司净利润	-452	2059	7140	15589	25423
EBITDA	-334	2282	7288	15889	25937
EPS(元)	-0.72	3.27	11.35	24.79	40.43

主要财务比率	2024A	2025A	2026E	2027E	2028E
成长能力					
营业收入(%)	65.6	453.2	161.8	111.7	66.7
营业利润(%)	48.0	552.3	246.4	118.3	63.1
归属于母公司净利润(%)	46.7	555.2	246.7	118.3	63.1
获利能力					
毛利率(%)	56.7	55.2	55.9	56.9	55.0
净利率(%)	-38.5	31.7	42.0	43.3	42.4
ROE(%)	-8.4	17.4	38.9	46.3	43.2
ROIC(%)	-8.5	16.7	38.1	45.7	42.8
偿债能力					
资产负债率(%)	19.2	11.9	21.4	15.2	17.2
净负债比率(%)	-25.9	-8.8	-11.3	-4.4	-8.9
流动比率	7.1	9.1	4.5	6.4	5.6
速动比率	3.8	4.7	1.1	0.9	0.8
营运能力					
总资产周转率	0.2	0.6	0.9	1.1	1.1
应收账款周转率	2.5	13.3	25.0	60.0	85.0
应付账款周转率	1.4	4.6	8.0	14.0	15.0
每股指标（元）					
每股收益(最新摊薄)	-0.72	3.27	11.35	24.79	40.43
每股经营现金流(最新摊薄)	-2.57	-0.79	-0.01	3.16	8.72
每股净资产(最新摊薄)	8.62	18.82	29.17	53.58	93.54
估值比率					
P/E	-1248.4	274.2	79.1	36.2	22.2
P/B	104.1	47.7	30.8	16.8	9.6
EV/EBITDA	-1681.7	244.9	76.9	35.2	21.4

数据来源：聚源、开源证券研究所

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已正式实施。根据上述规定，开源证券评定此研报的风险等级为R4（中高风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为专业投资者及风险承受能力为C4、C5的普通投资者。若您并非专业投资者及风险承受能力为C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师声明

本研究报告的署名人员具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（outperform）	预计相对强于市场表现 5%～20%；
	中性（Neutral）	预计相对市场表现在－5%～＋5%之间波动；
	减持（underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（underperform）	预计行业弱于整体市场表现。
备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数（北交所基准指数为北证 50 指数）、港股基准指数为恒生指数、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）、美股基准指数为标普 500 或纳斯达克综合指数。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取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。		

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（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。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，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、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。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。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。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。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所引致的任何损失负任何责任。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，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。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

开源证券研究所

上海
地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

深圳
地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

北京
地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

西安
地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn