



联合研究 | 公司深度 | 寒武纪-U (688256.SH)

寒武纪深度报告：云程发轫，厚积薄发

报告要点

Transformer 架构大力出奇迹的成功是本轮 AI 革命的起源，Scaling Laws（尺度定律）支撑了 AI 算力需求的爆发式增长，中国 AI 芯片市场也保持了快速的增长，另一方面中美科技摩擦加剧，科技自立紧迫性和重要性进一步提升，国产算力将迎来历史性机遇，国内政策频出，各类主要下游客户采购国产 AI 芯片意愿和占比持续提升，国产 AI 芯片份额提升空间极大；寒武纪是国产算力的领军者，具备定制化开发能力和丰富的项目经验，智能计算集群系统优势显著，有望在 AI 芯片国产化大潮中充分受益。

分析师及联系人



宗建树

SAC: S0490520030004

SFC: BUX668



杨洋

SAC: S0490517070012

SFC: BUW100



余庚宗

SAC: S0490516030002

寒武纪-U (688256.SH)

2024-09-01

联合研究 | 公司深度

投资评级 买入 | 首次

寒武纪深度报告：云程发轫，厚积薄发

引言——AIGC 时代，全球智算军备竞赛

Transformer 架构大力出奇迹的成功是本轮 AI 革命的起源，Scaling Laws（尺度定律）支撑了 AI 算力需求的爆发式增长。随着 AIGC 的兴起，未来算力、智算产业将成为 AI 浪潮中的重要支撑。随着模型性能不断提升，Transformer 架构已在文本方面（语言类）跑通，后续会在图像、视频甚至执行侧复现从而逐渐通往 AGI。当前海内外大模型迭代如火如荼、文生视频迎来突破，带动算力尤其是智算产业的需求持续旺盛。

寒武纪：以技术立身的纯正国产算力领军

寒武纪是国产算力的领军者。自成立以来，寒武纪公司深耕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。公司核心人员在处理器芯片和人工智能领域钻研多年，带领公司研发了智能处理器指令集与微架构等一系列自主创新关键技术，围绕云、边、端三种场景，打造 AI 芯片产品矩阵，还提供智能计算集群系统服务，该业务是公司营收的主要来源。公司长期的研发积累使得其产品行业内赢得高度认可，广泛应用于消费电子、数据中心、云计算等诸多场景。公司营收保持稳定，净利润实现亏损收窄，坚持投入研发，股权激励高营收增长目标彰显未来成长信心。

总量份额双升共振，国产算力迎历史机遇

AIGC 强劲需求带动下，全球 AI 芯片市场规模呈现持续高速增长态势。近年来，中国 AI 芯片市场也保持了快速的增长，根据亿欧智库数据，2023 年中国人工智能芯片市场规模达 1039 亿元，2025 年将达到 1780 亿元。以英伟达为代表的 AI 芯片厂商业绩持续高速增长兑现，印证算力需求高景气。除英伟达外，国内外大厂自研芯片以及其他类型 AI 芯片也在与日俱增，且着眼中长期结构，AI 算力需求重心有望逐步由训练向推理转移，中长期成长空间广阔。

中美科技摩擦加剧，科技自立紧迫性和重要性进一步提升，国产算力将迎来历史性机遇，国内政策频出，各类主要下游客户采购国产 AI 芯片意愿和占比持续提升，国产 AI 芯片份额提升空间极大。

云边端产品升级迭代，软硬件协同筑生态

寒武纪公司掌握 AI 芯片及系统软件领域关键核心技术，云边端产品不断迭代，产品矩阵逐步完善，重视软件平台的生态发展，推出软件开发平台 Cambricon NeuWare 作为打造云边端统一的 AI 开发生态的核心部件。整体上，公司提供云边端一体、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，受益于软硬件协同，公司具备定制化开发能力和丰富的项目经验，智能计算集群系统优势显著，有望在 AI 芯片国产化大潮中充分受益。

风险提示

- 1、需求不及预期的风险；
- 2、产能供应不足的风险；
- 3、市场竞争加剧的风险；
- 4、盈利预测假设不成立或不及预期的风险。

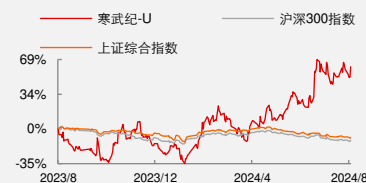
请阅读最后评级说明和重要声明

公司基础数据

当前股价(元)	256.69
总股本(万股)	41,746
流通A股/B股(万股)	41,746/0
每股净资产(元)	12.54
近12月最高/最低价(元)	278.59/95.85

注：股价为 2024 年 8 月 30 日收盘价

市场表现对比图(近 12 个月)



资料来源：Wind



更多研报请访问
长江研究小程序

目录

引言——AIGC 时代，全球智算军备竞赛	6
寒武纪：以技术立身的纯正国产算力领军	9
筚路蓝缕，国产人工智能芯片领军享先发	9
营收相对稳健，持续投入或迎来厚积薄发	14
总量份额双升共振，国产算力迎历史机遇	17
技术变革驱动，全球算力行业持续高增长	17
国内外因素共振，国产算力份额有望提升	19
云边端产品升级迭代，软硬件协同筑生态	23
兼具软硬件核心技术，云边端产品频迭代	23
软硬件协同完善生态，智算集群优势尽显	25
风险提示	29

图表目录

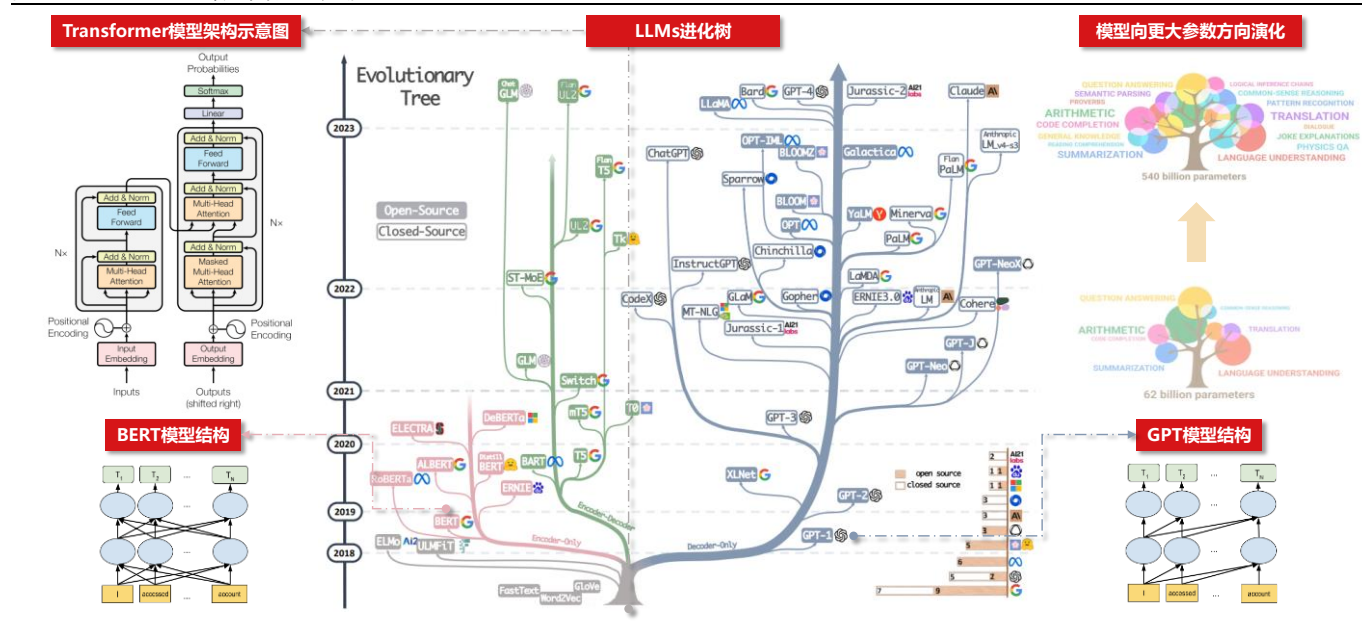
图 1：Transformer 引发本轮 AI 浪潮	6
图 2：参数规模达到一定量级时产生涌现效应	6
图 3：模型参数规模、训练数据集规模、计算资源投入与模型性能间存在 Scaling Law	7
图 4：三要素同步提升以获得最佳性能	7
图 5：Scaling Law 将不断在其他领域复现，带动算力产业需求持续旺盛	8
图 6：寒武纪发展历程	9
图 7：寒武纪股权结构	10
图 8：终端智能处理器 IP 架构	13
图 9：寒武纪智能计算集群系统软硬件总体架构	14
图 10：寒武纪营收及增速	14
图 11：寒武纪归母净利润	14
图 12：寒武纪业务结构	15
图 13：寒武纪各业务占比	15
图 14：寒武纪近五年毛利率变化	15
图 15：寒武纪细分子业务毛利率变化	15
图 16：公司期间费用	16
图 17：全球 GPU 市场空间预测（单位：亿美元）	17
图 18：全球数据中心 GPU 市场空间及预测（单位：亿美元）	17
图 19：2019-2025 年中国人工智能芯片市场规模（单位：亿元）	17
图 20：中国人工智能芯片市场规模占比	17
图 21：英伟达营业收入（单位：亿美元）	18
图 22：英伟达数据中心业务营业收入（单位：亿美元）	18
图 23：2019 年 1 月至今英伟达股价走势（单位：美元）	18
图 24：国内外代表性互联网厂商自研芯片情况	19
图 25：2022-2027 年中国人工智能服务器工作负载结构及预测	19

图 26：寒武纪核心技术框架结构	23
图 27：云边端产品布局	24
图 28：寒武纪基础软件系统平台 Cambricon NeuWare	26
图 29：Cambricon NeuWare 训练软件平台	26
图 30：推理加速引擎 MagicMind	26
图 31：智能计算集群系统软硬件总体架构	27
图 32：智能计算集群系统整体业务流程	27
表 1：寒武纪管理层简介	10
表 2：云、边、端三种场景对芯片特性要求不同	11
表 3：公司云端产品	11
表 4：公司边缘端产品	12
表 5：公司终端智能处理器 IP	12
表 6：首次授予部分分公司层面业绩考核要求	16
表 7：美国政府不断限制中国使用美先进算力	20
表 8：2024 年各省份智算中心相关政策梳理	20
表 9：近年来三大运营商 AI 服务器集采招标统计	21
表 10：云边端三大产品线产品迭代梳理	23
表 11：主要算力芯片参数指标对比	24
表 12：2019-2023 年寒武纪中标智算相关项目梳理	27
表 13：公司利润敏感性分析	29

引言——AIGC 时代，全球智算军备竞赛

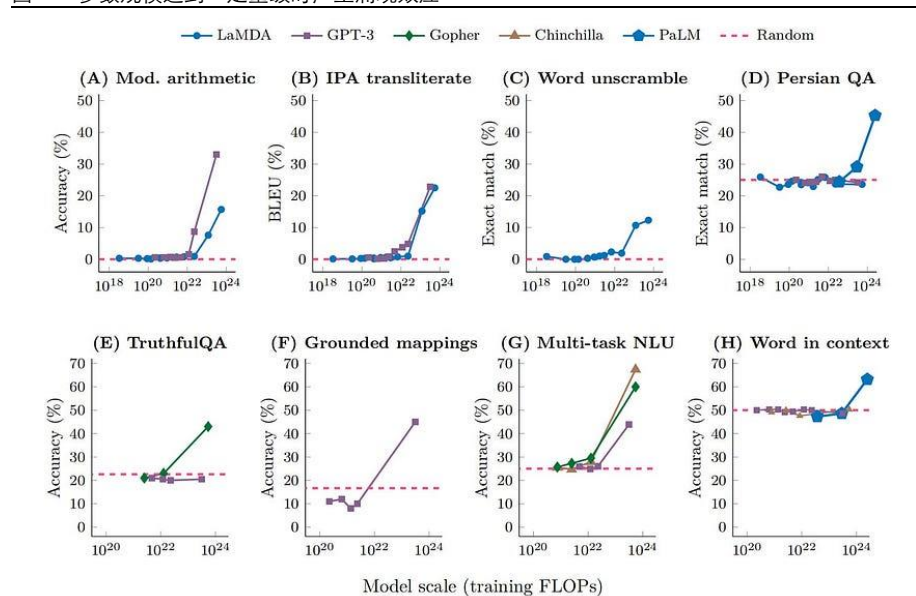
Transformer 架构大力出奇迹的成功是本轮 AI 革命的重要理论支撑。本轮生成式 AI 的技术变革由 Transformer 架构引发，在基于 Transformer 架构进行模型迭代过程中，实证产业规律显示，模型的规模越大（参数越多），模型的能力越强，并且当参数达到一定量级时模型会涌现出新的能力产生质变，这一现象被总结为“尺度定律”（Scaling Laws）。

图 1：Transformer 引发本轮 AI 浪潮



资料来源：《Attention Is All You Need》，Google Research，《Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond》，Talk AI，长江证券研究所

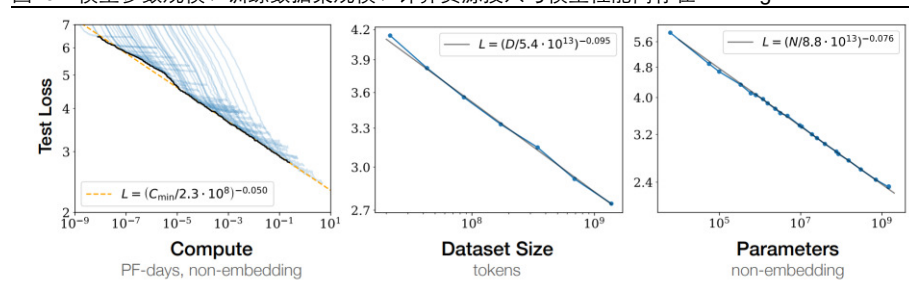
图 2：参数规模达到一定量级时产生涌现效应



资料来源：《Are Emergent Abilities of Large Language Models a Mirage?》，长江证券研究所

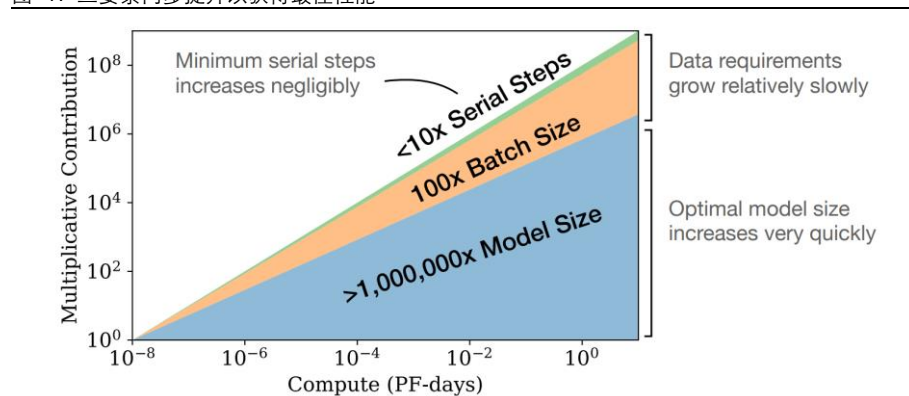
Scaling Laws（尺度定律）支撑 AI 算力需求的爆发式增长。Scaling Laws 表明，模型规模越大则能力越强。其中，模型规模包括模型的参数量、数据集大小以及计算量。因此，想要获得更强大的模型能力，就需要不断扩大模型规模，进而需要海量的计算资源来支持更大规模的模型训练，由此引发了算力需求的激增。

图 3：模型参数规模、训练数据集规模、计算资源投入与模型性能间存在 Scaling Law



资料来源：《Scaling Laws for Neural Language Models》，长江证券研究所

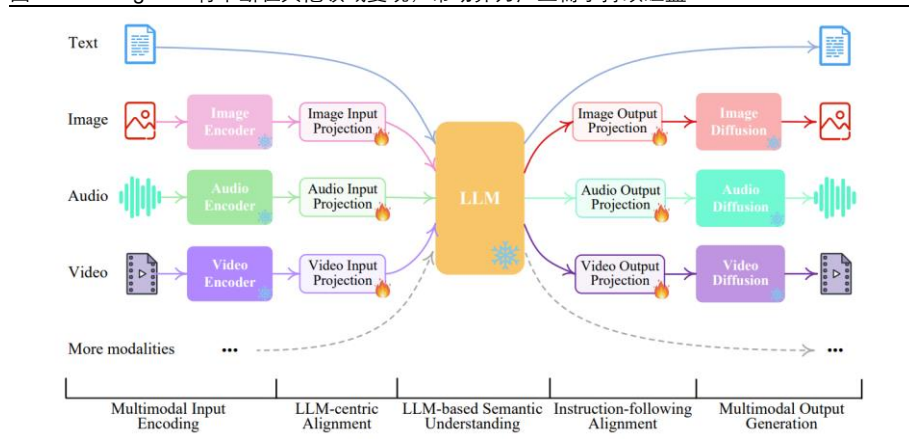
图 4：三要素同步提升以获得最佳性能



资料来源：《Scaling Laws for Neural Language Models》，长江证券研究所

随着 AIGC 的兴起，算力尤其是智算产业爆发式增长已成为 AI 浪潮中的确定性趋势。随着模型性能不断提升，Transformer 架构已在文本方面（语言类）证明其成效，后续会在图像、视频甚至执行侧复现逐渐通往 AGI。当前海内外大模型迭代如火如荼、文生视频迎来突破，带动算力产业的需求持续旺盛，算力尤其是智算产业爆发式增长已成为 AI 浪潮中的确定性趋势。

图 5: Scaling Law 将不断在其他领域复现, 带动算力产业需求持续旺盛



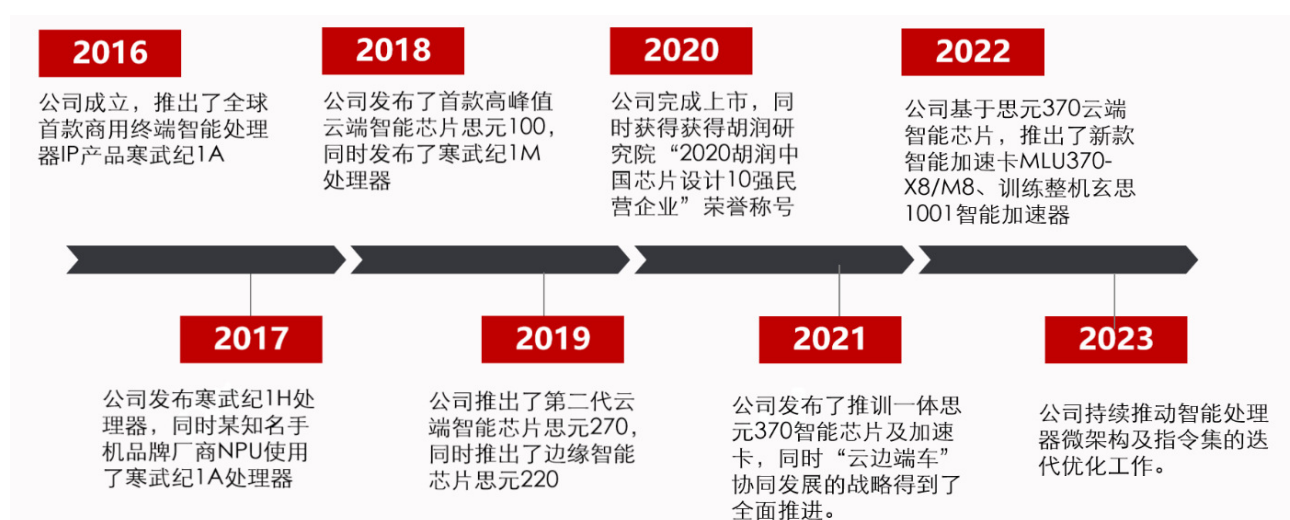
资料来源:《Any-to-Any Multimodal Large Language Model》, 长江证券研究所

寒武纪：以技术立身的纯正国产算力领军

筚路蓝缕，国产人工智能芯片领军享先发

寒武纪是国产算力的领军者。自成立以来，寒武纪公司深耕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。公司核心人员在处理器芯片和人工智能领域钻研多年，带领公司研发了智能处理器指令集与微架构等一系列自主创新关键技术。公司长期的研发积累使得其产品在行业内赢得高度认可，广泛应用于消费电子、数据中心、云计算等诸多场景。

图 6：寒武纪发展历程

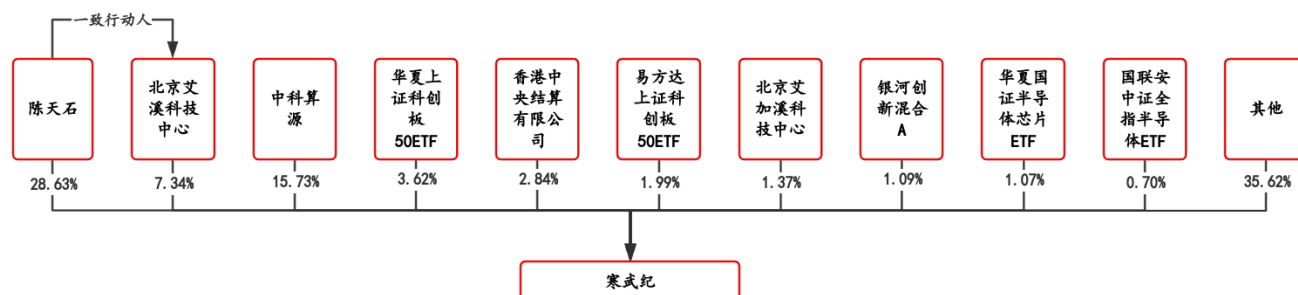


资料来源：公司官网，公司招股说明书，公司年报，长江证券研究所

公司创始人陈天石是寒武纪的实际控制人，是公司的第一大股东。公司董事长陈天石通过直接持股以及间接持股的方式，持有公司共计 35.97%的股权，对董事会实施决策有着举足轻重的影响，公司整体股权结构稳定。

中科院旗下中科算源为公司第二大股东，控股 15.73%，产学研结合为公司发展注入强劲动力。中科算源为中科院旗下子公司，其为寒武纪的发展提供了良好的背景、平台、资源等，产学研协同使得公司快速跻身于 AI 芯片领军公司行列。同时，公司核心团队多人出自中科院，具有强大的研发背景以及优越的研发技术。

图 7：寒武纪股权结构



资料来源：同花顺，截至 2024 年 7 月 26 日，长江证券研究所

寒武纪把技术创新作为公司的战略重点，持续进行研发投入，确保在中国人工智能芯片领域的领先地位。公司董事长、总经理陈天石博士曾在中科院计算所担任研究员、博士生导师，在人工智能和处理器芯片等相关领域从事基础科研工作十余年，积累了坚实的理论功底和丰富的研发经验，创办并领导公司在智能芯片方向快速跻身全球初创公司前列。

公司在技术研发方面建立了成熟团队，核心骨干均在行业中深耕多年。公司核心研发人员多毕业于著名高校或科研院所，拥有计算机、微电子等相关专业的学历背景，多名骨干成员拥有知名半导体公司多年的工作经历。公司研发队伍结构合理、技能全面，有力支撑了公司的技术创新和产品研发。

表 1：寒武纪管理层简介

姓名	职务	简介
陈天石	董事长，总经理	中国科学技术大学计算机软件与理论专业博士学位，2010 年 7 月至 2019 年 9 月就职于中科院计算所（2018 年 4 月办理离岗创业），历任助理研究员、副研究员及硕士生导师、研究员及博士生导师。2016 年 3 月创立中科寒武纪科技股份有限公司，现任中科寒武纪科技股份有限公司董事长、总经理。
叶淦尹	董秘，财务负责人	北京大学西方经济学硕士学位，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 年加入中科寒武纪科技股份有限公司，现任中科寒武纪科技股份有限公司董事、副总经理、财务负责人及董事会秘书。
王在	副总经理	中国科学技术大学计算机应用技术博士学位。2011 年至 2015 年就职于郑州商品交易所并任核心交易系统工程师，2015 年至 2016 年就职于中原银行并任信息科技部电子银行系统主管，2016 年至 2018 年就职于中科院计算所从事科研工作。2016 年作为中科寒武纪科技股份有限公司创始团队成员加入中科寒武纪科技股份有限公司，现任中科寒武纪科技股份有限公司董事、副总经理。
刘少礼	副总经理	中科院计算所计算机系统结构博士学位。2014 年至 2019 年就职于中科院计算所并任副研究员（2018 年 4 月办理离岗创业）。2016 年作为中科寒武纪科技股份有限公司创始团队成员加入中科寒武纪科技股份有限公司，现任中科寒武纪科技股份有限公司董事、副总经理。
陈煜	副总经理	浙江大学通信工程本科学历。2003 年至 2009 年，就职于智多微电子（上海）有限公司，历任工程师，高级工程师，设计经理，SoC 设计部总监。2009 年至 2012 年，任宇芯科技有限公司芯片部总监。2013 年至 2016 年，任上海华力创通半导体有限公司芯片部总监。2017 年至今，就职于中科寒武纪科技股份有限公司，历任芯片部副总监，现任中科寒武纪科技股份有限公司副总经理。
刘道福	核心技术人员	中科院计算所计算机系统结构博士学位。2015 年至 2019 年就职于中科院计算所，历任助理研究员、高级工程师，2016 年作为寒武纪创始团队成员加入寒武纪。2019 年 11 月至 2022 年 11 月任中科寒武纪科技股份有限公司副总经理，现为寒武

纪核心技术人员。

资料来源：同花顺，长江证券研究所

公司围绕云、边、端三种场景，打造 AI 芯片产品矩阵。经过多年的研发创新，公司核心产品不断升级迭代，目前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。人工智能技术在云端、边缘端和终端设备的应用中都需要由核心芯片提供计算能力支撑，但云、边、端三种场景对于芯片的运算能力和功耗等特性要求不同，单一品类的智能芯片难以满足实际应用的需求。

表 2：云、边、端三种场景对芯片特性要求不同

应用场景	芯片需求	典型计算能力	典型功耗	典型应用领域
终端	低功耗、高能效、推理任务为主、成本敏感、硬件产品形态众多	< 8TOPS	< 5 瓦	各类消费类电子、物联网产品等
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	> 30TOPS	> 50 瓦	云计算数据中心、企业私有云等
边缘端	对功耗、性能、尺寸的要求常介于终端和云端之间、推理任务为主、多用于插电设备、硬件产品形态较少	5TOPS 至 30TOPS	4 瓦至 15 瓦	智能制造、智能家居、智能零售、智慧交通、智慧金融、智慧医疗、智能驾驶等众多应用领域

资料来源：公司招股说明书，长江证券研究所

公司云端产品线目前包括云端智能芯片、加速卡及训练整机。其中，云端智能芯片及加速卡是云服务器、数据中心等进行人工智能处理的核心器件，其主要作用是为云计算和数据中心场景下的人工智能应用程序提供高计算密度、高能效的硬件计算资源，支撑该类场景下复杂度和数据吞吐量高速增长的人工智能处理任务。公司的训练整机是由公司自研云端智能芯片及加速卡提供核心计算能力，且整机亦由公司自研的训练服务器产品。公司的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。

表 3：公司云端产品

产品	系列	型号	产品概况
智能加速卡	思元 270	MLU270-S4	功耗仅为 70W，处理非稀疏人工智能模型的理论峰值性能提升至上一代思元 100 的 4 倍，可广泛支持视觉、语音、自然语言处理以及传统机器学习等高度多样化的人工智能应用，帮助 AI 推理平台实现超高能效比。
		MLU270-F4	采用寒武纪 MLUv02 架构，搭载 EOTS 主动散热技术的 MLU270-F4，可轻松胜任非数据中心部署环境。可支持最高 150W 散热功率，处理非稀疏人工智能模型的理论峰值性能提升至上一代思元 100 的 4 倍。
	思元 290	MLU290-M5	搭载寒武纪首颗训练芯片思元 290，采用台积电 7nm 先进制程工艺，采用 MLUv02 扩展架构。采用开放加速模块 OAM 设计，具备 64 个 MLU Core，1.23TB/s 内存带宽以及全新 MLU-Link 芯片间互联技术，同时支持单机八卡机内互联，多机多卡机间互联，全面支持 AI 训练、推理或混合型人工智能计算加速任务。
		MLU370-S4/S8	采用思元 370 芯片，TSMC 7nm 制程，寒武纪新一代人工智能芯片架构 MLUarch03 加持，支持 PCIe Gen4，板卡功耗仅为 75W，相较于同尺寸 GPU，可提供 3 倍的解码能力和 1.5 倍的编码能力。能效出色，体积小巧，可在服务器中实现高密度部署。
	思元 370	MLU370-X4	采用思元 370 芯片，为单槽位 150w 全尺寸加速卡，可提供高达 256TOPS(INT8)推理算力，和 24TFLOPS(FP32)训练算力，同时提供丰富的 FP16、BF16 等多种训练

精度，配合全新基础系统软件平台，可充分满足推理一体 AI 任务需求。		
智能加速系统	MLU370-X8	采用双芯思元 370 配置，为双槽位 250w 全尺寸智能加速卡，提供 24TFLOPS(FP32)训练算力和 256TOPS (INT8)推理算力，同时提供丰富的 FP16、BF16 等多种训练精度。基于双芯思元 370 打造的 MLU370-X8 整合了两倍于标准思元 370 加速卡的内存、编解码资源，同时 MLU370-X8 搭载 MLU-Link 多芯互联技术，每张加速卡可获得 200GB/s 的通讯吞吐性能，是 PCIe 4.0 带宽的 3.1 倍，支持单机八卡部署，可高效执行多芯多卡训练和分布式推理任务。
	玄思 1000 智能加速器整机	在 2U 机箱内集成了 4 颗思元 290 智能芯片，2 台玄思 1000 加速器与 CPU 服务器可组成一套包括 8 张加速卡的整机系统，可实现 AI 算力多向扩展，满足性能、扩展性、灵活性、鲁棒性的要求。

资料来源：公司官网，长江证券研究所

公司边缘端产品主要为思元 220 系列产品，通过在终端和云端之间的设备上配备适度的计算能力，弥补终端设备计算能力不足的劣势，同时缓解云计算场景下数据隐私、带宽与延时等潜在问题。

表 4：公司边缘端产品

系列	型号	简介
思元 220	MLU220-SOM	MLU220-SOM 智能模组采用全新的 MLUv02 架构，基于信用卡大小的模组上可以实现 16TOPS AI 算力的单系统解决方案，功耗仅为 15W。
	MLU220-M.2	思元 220 芯片基于寒武纪 MLUv02 架构，手指大小的标准 M.2 加速卡集成了 8TOPS 理论峰值性能，功耗仅为 8.25W,可以轻松实现终端设备和边缘端设备的 AI 赋能方案。

资料来源：公司官网，长江证券研究所

公司终端产品线主要包括 IP 授权和基础系统软件平台。IP 授权是将公司研发的智能处理器 IP 等知识产权授权给客户在其产品中使用。基础系统软件平台是公司云边端全系列智能芯片与处理器产品提供统一的平台级基础系统软件。

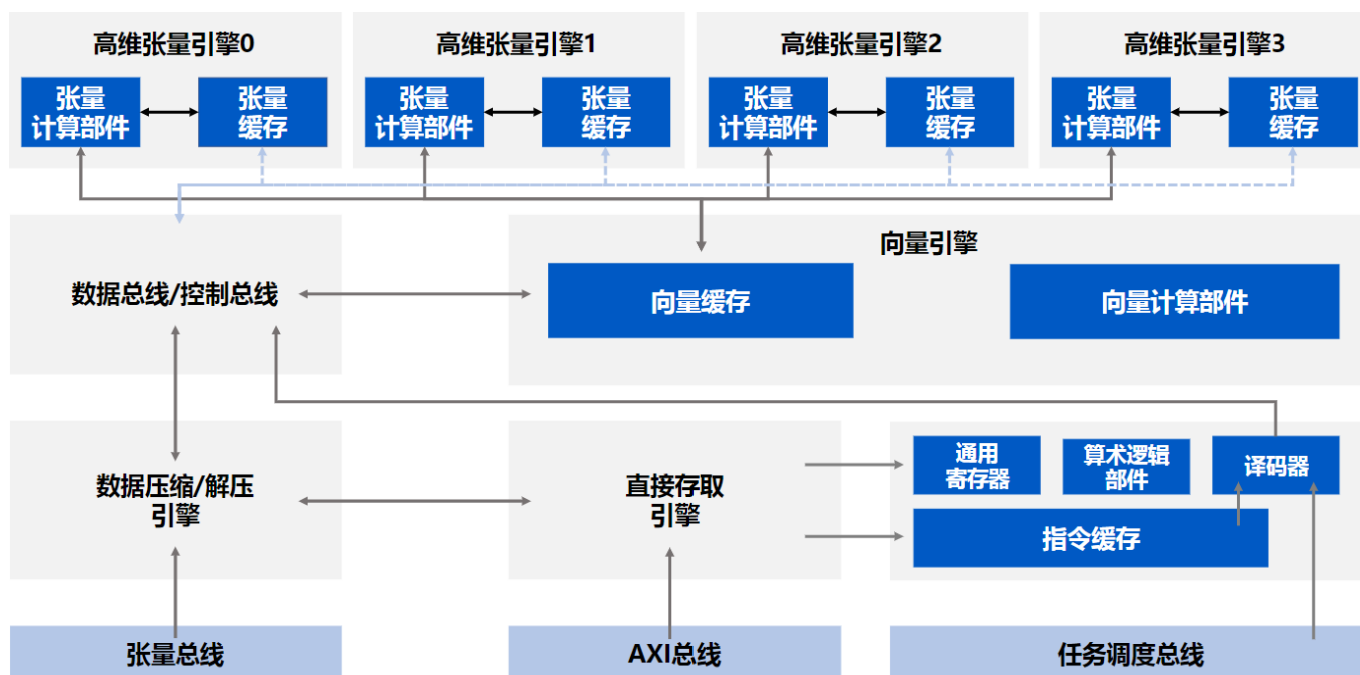
表 5：公司终端智能处理器 IP

产品	系列	型号	简介
终端智能处理器 IP	Cambricon-1H	Cambricon-1H16	使用 256MAC 16 位浮点运算器以及 512MAC 8 位定点运算器。在 1GHz 主频下，进行 16 位浮点人工智能运算的峰值速度为 0.5Tops；进行 8 位定点人工智能运算的峰值速度为 1Tops。
		Cambricon-1H8	使用 512MAC 8 位定点运算器。在 1GHz 主频下，进行 8 位定点人工智能运算的峰值速度为 1Tops。
		Cambricon-1H8mini	使用 256MAC 8 位定点运算器。在 1GHz 主频下，进行 8 位定点人工智能运算的峰值速度为 0.5Tops。
	Cambricon-1M	Cambricon-1M-4K	使用了 4096MAC 8 位定点运算器。在 1GHz 主频下，进行 8 位定点人工智能运算的峰值速度为 8Tops，进行 16 位定点人工智能运算的峰值速度为 4Tops，进行 32 位定点人工智能运算的峰值速度为 1Tops。
		Cambricon-1M-2K	使用了 2048MAC 8 位定点运算器。在 1GHz 主频下，进行 8 位定点人工智能运算的峰值速度为 4Tops，进行 16 位定点人工智能运算的峰值速度为 2Tops，进行 32 位定点人工智能运算的峰值速度为 0.5Tops。
		Cambricon-1M-1K	使用了 1024MAC 8 位定点运算器。在 1GHz 主频下，进行 8 位定点人工智能运算的峰值速度为 2Tops，进行 16 位定点人

工智能运算的峰值速度为 1Tops，进行 32 位定点人工智能运算的峰值速度为 0.25Tops。

资料来源：公司官网，长江证券研究所

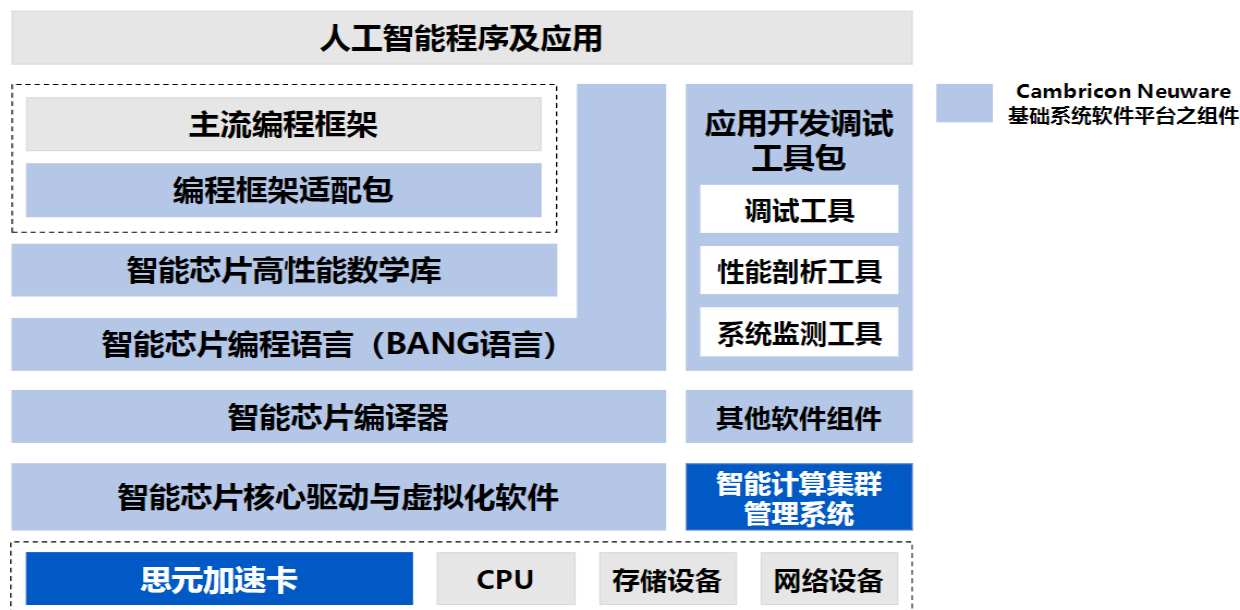
图 8：终端智能处理器 IP 架构



资料来源：寒武纪招股说明书，长江证券研究所

除在三大产品线布局之外，公司还提供智能计算集群系统服务，该业务是公司营收的主要来源。寒武纪将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，聚焦于人工智能技术在数据中心的应用，为人工智能应用部署技术能力相对较弱的客户提供软硬件整体解决方案，以科学地配置和管理集群的软硬件、提升运行效率。

图 9：寒武纪智能计算集群系统软硬件总体架构

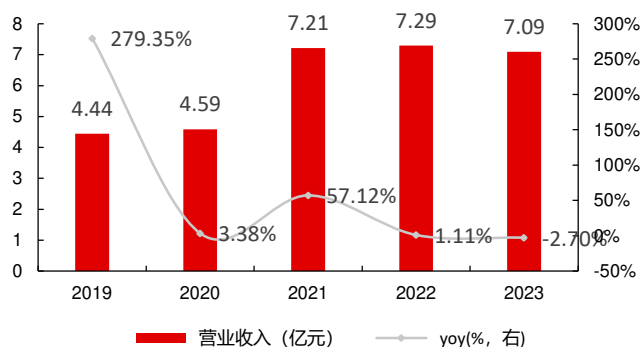


资料来源：寒武纪招股说明书，长江证券研究所

营收相对稳健，持续投入或迎来厚积薄发

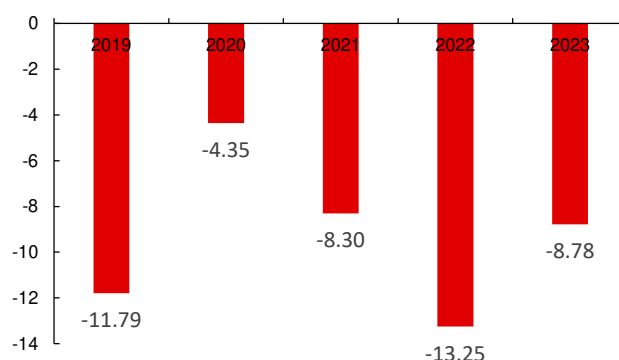
公司营收保持稳定，净利润实现亏损收窄。2021 年至 2023 年公司营收保持相对稳定，受供应链影响，公司调整销售策略，优先服务毛利较高、信用较好的客户，导致 2023 年营收同比下降 2.70%。2023 年公司归母净利润亏损收窄 4.47 亿元，主要原因是公司按归属期分摊的股份支付费用较上年同期减少、研发效率提升、资源配置优化以及本期计提的资产减值损失较上年同期有所减少。展望未来随着公司需求在 AIGC 驱动下向好，供给端随着国内技术进步与突破不断改善，有望迎来业绩大幅向好。

图 10：寒武纪营收及增速



资料来源：同花顺，长江证券研究所

图 11：寒武纪归母净利润（亿元）

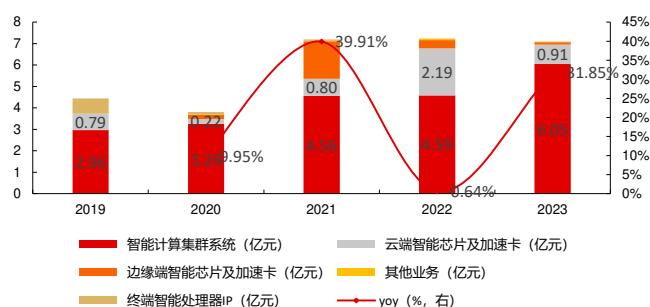


资料来源：同花顺，长江证券研究所

智能计算集群系统业务占比高，是公司营收的主要来源。基于智能计算集群系统业务的竞争优势和前期落地项目的良好经验，公司积极参与台州、沈阳两地的算力基础设施建设项目，并且公司的智能计算集群系统业务已在多个城市实现拓展，国内市场占有率目

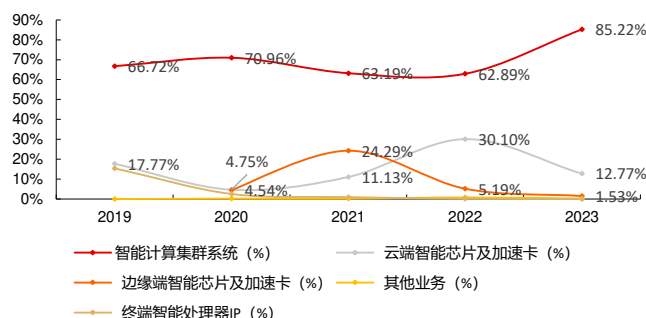
前处于第一梯队。2023 年公司智能计算集群系统业务营业收入为 6.05 亿元，同比增长 31.85%，占比由 62.89% 上升至 85.22%。公司云端以及边缘端智能芯片及加速卡业务营收及占比均有所下降，营收分别由 2.19 亿元/0.38 亿元，下降至 0.91 亿元/0.11 亿元，占比由 30.10%/5.19%，下降至 12.77%/1.53%。

图 12：寒武纪业务结构



资料来源：同花顺，长江证券研究所

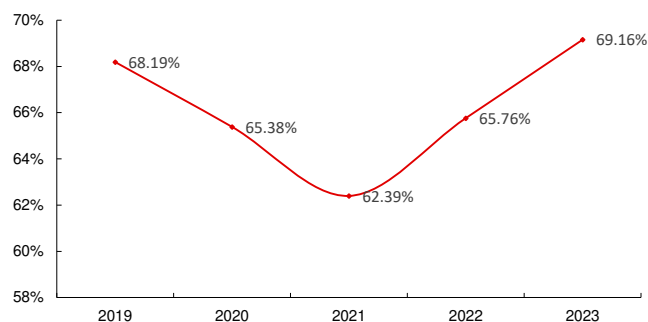
图 13：寒武纪各业务占比



资料来源：同花顺，长江证券研究所

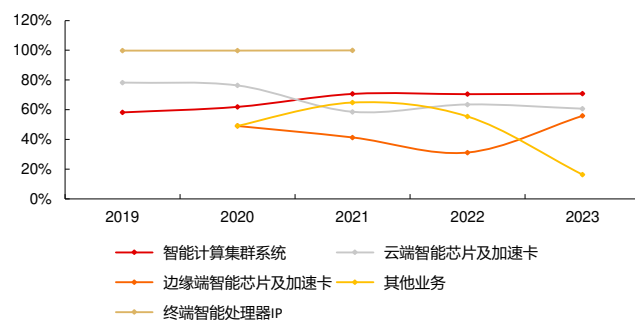
公司调整战略，陆续暂停部分预期毛利率较低的研发项目，毛利率不断上升。受 BIS 将公司列入“实体清单”影响，公司暂停部分毛利率较低的研发项目，2023 年毛利率为 69.16%，较 2022 年上升 3.40%。公司细分子业务中，2021 年终端智能处理器 IP 业务毛利率接近 100%；公司营收主要来源智能计算集群系统的毛利率稳定在 70% 左右，对公司的盈利能力起到主要的支撑作用。

图 14：寒武纪近五年毛利率变化



资料来源：同花顺，长江证券研究所

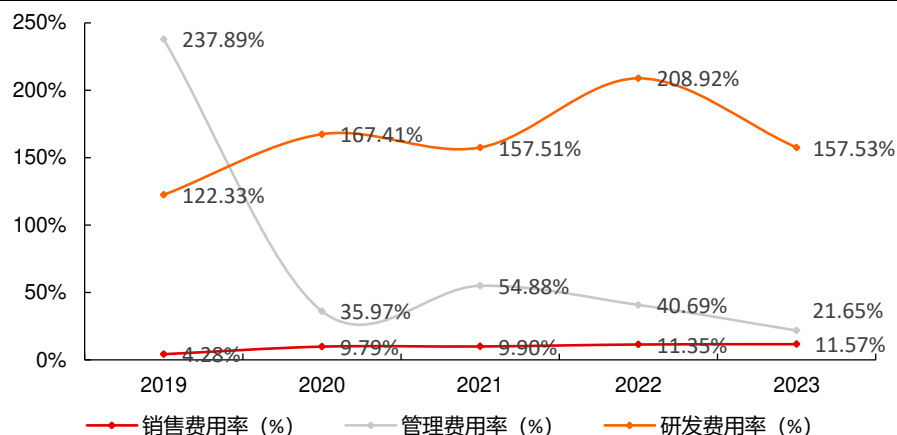
图 15：寒武纪细分子业务毛利率变化



资料来源：同花顺，长江证券研究所

公司在研发效率提升的情况下持续保持研发高投入。公司为了打造优质的产品 & 易用的软件生态平台，确保在中国人工智能芯片领域的领先地位，始终高度重视研发，自 2019 年以来，公司始终保持 100% 以上的研发费用率，2023 年，公司的研发费用率达到了 157.53%，较去年有所下降，主要由研发效率提升、资源配置优化所导致。近三年来公司的管理费用率呈现逐渐下降趋势，分别为 54.88%/40.69%/21.65%，公司按归属期分摊的股份支付费用减少导致了公司管理费用率的下降。

图 16：公司期间费用



资料来源：同花顺，长江证券研究所

股权激励高营收目标彰显成长信心。为了进一步建立、健全公司长效激励机制，吸引和留住优秀人才，公司实施股权激励计划，不仅彰显公司对于未来业绩的信心与良好预期，同时能够充分调动公司核心团队的积极性，有效地将股东利益、公司利益和核心团队个人利益结合在一起。

表 6：首次授予部分公司层面业绩考核要求

归属期	目标值	触发值
	公司层面归属系数 100%	公司层面归属系数 80%
第一个归属期	2024 年营业收入值不低于 11 亿元	2024 年营业收入值不低于 8.8 亿元
第二个归属期	2024-2025 年累计营业收入值不低于 26 亿元	2024-2025 年累计营业收入值不低于 20.8 亿元
第三个归属期	2024-2026 年累计营业收入值不低于 46 亿元	2024-2026 年累计营业收入值不低于 36.8 亿元

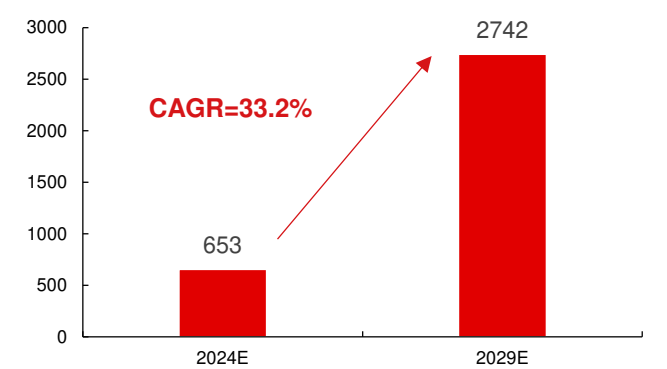
资料来源：公司公告，长江证券研究所

总量份额双升共振，国产算力迎历史机遇

技术变革驱动，全球算力行业持续高增长

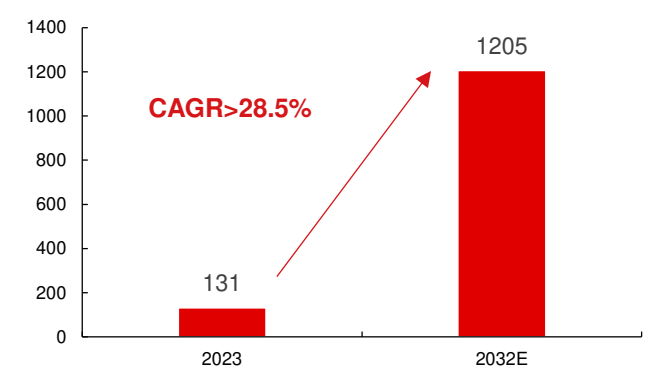
在生成式 AI 强劲的需求带动下，全球 GPU 芯片市场规模持续呈现高速增长态势。根据 Fortune Business Insights 和 Mordor Intelligence 数据，2023 年全球 GPU 芯片市场规模为 484 亿美元，预计 2024 年达到 653 亿美元，2029 年将达到 2742 亿美元规模，2024-2029 年复合增长率为 33.2%。其中，数据中心 GPU 在 2023 年的全球市场规模达到 131 亿美元，预计 2024-2032 年复合增长率将超过 28.5%。AI 处理大数据集和复杂算法的需求不断增长，推动了全球对高性能计算解决方案的需求。

图 17：全球 GPU 市场空间预测（单位：亿美元）



资料来源：Mordor Intelligence，长江证券研究所

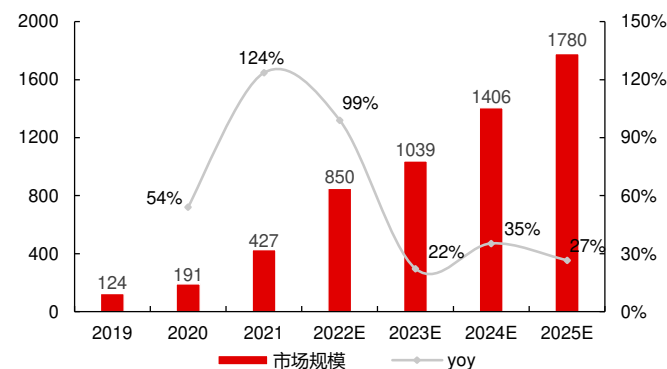
图 18：全球数据中心 GPU 市场空间及预测（单位：亿美元）



资料来源：Global Market Insights，长江证券研究所

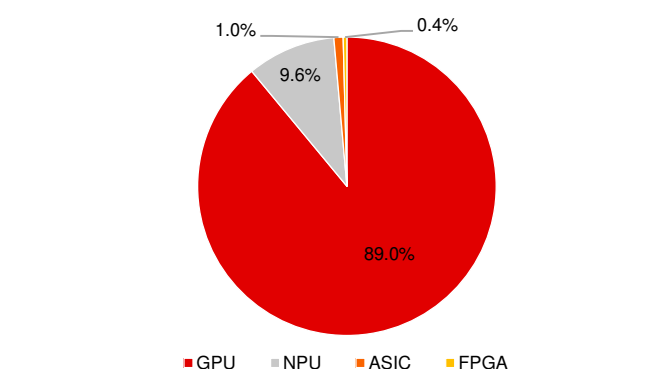
中国人工智能芯片市场规模保持高速增长。近年来，中国 AI 芯片市场保持了快速的增长趋势。根据亿欧智库数据，预计 2024 年中国人工智能芯片市场规模将达到 1406 亿元；2025 年中国人工智能芯片市场规模将达到 1780 亿元。当前国内 AI 芯片市场仍以 GPU 为主实现数据中心计算加速，市场占有率近 90%。因统计口径的差异，市场对于人工智能芯片规模预测存在差异，但普遍呈现空间较大，高速成长的特点。

图 19：2019-2025 年中国人工智能芯片市场规模（单位：亿元）



资料来源：亿欧智库，长江证券研究所

图 20：中国人工智能芯片市场规模占比

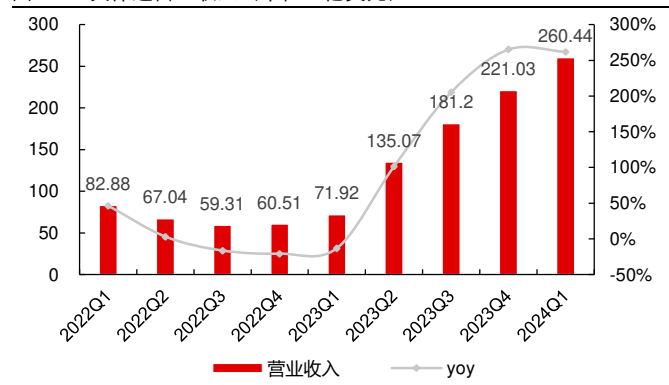


资料来源：IDC，长江证券研究所

以英伟达为代表的 AI 芯片厂商业绩高增长持续兑现，印证算力需求高景气。本轮 AI 浪潮下，英伟达作为高端 AI 芯片的垄断方，其营收规模迎来爆发式增长，且不断用超预期的季度业绩验证算力需求的高景气。英伟达 2024Q1 财报超市场预期，实现营业收入

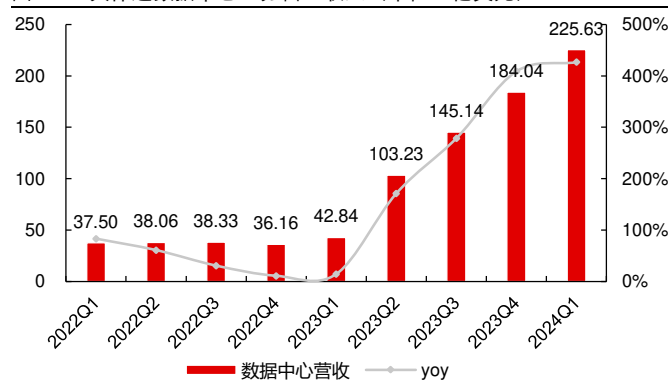
260 亿美元（彭博一致预期 247 亿美元），主要由数据中心业务驱动（2024Q1 数据中心业务营收同比增长 426.68%）。业绩高增下英伟达市值不断创新高，2022 年至 2024/8/15，市值累计增长 291%。

图 21：英伟达营业收入（单位：亿美元）



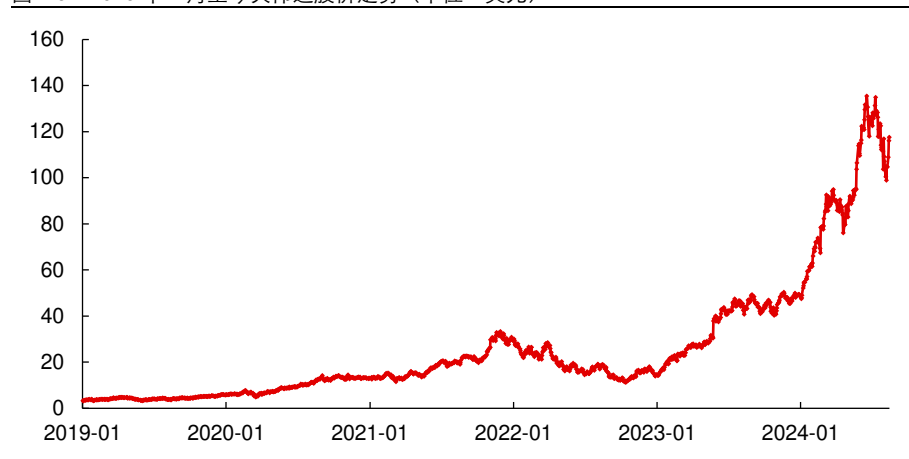
资料来源：Wind，长江证券研究所

图 22：英伟达数据中心业务营业收入（单位：亿美元）



资料来源：Wind，长江证券研究所

图 23：2019 年 1 月至今英伟达股价走势（单位：美元）



资料来源：Wind，长江证券研究所（数据截至 2024/8/15）

除英伟达外，国内外大厂自研芯片以及其他类型 AI 芯片也在与日俱增。大厂自研芯片方面，出于成本控制以及供应链安全等因素的考量，全球各大云计算厂商与头部科技公司正纷纷加速推进自研 AI 芯片，以摆脱对英伟达的依赖，算力芯片格局进一步复杂化。同时，其他类型 AI 芯片的市场规模也在不断扩张。根据 Business Research Insights 数据，2021 年全球 ASIC 市场规模为 155 亿美元，预计 2032 年达到 395 亿美元，复合增长率达到 8.9%；根据 Mordor Intelligence 的预测，预计 2024 年全球 FPGA 市场规模将达到 69 亿美元，2029 年将达到 124 亿美元，复合增长率达 12.5%，整体上其他类型智能芯片的需求也在日益增加。

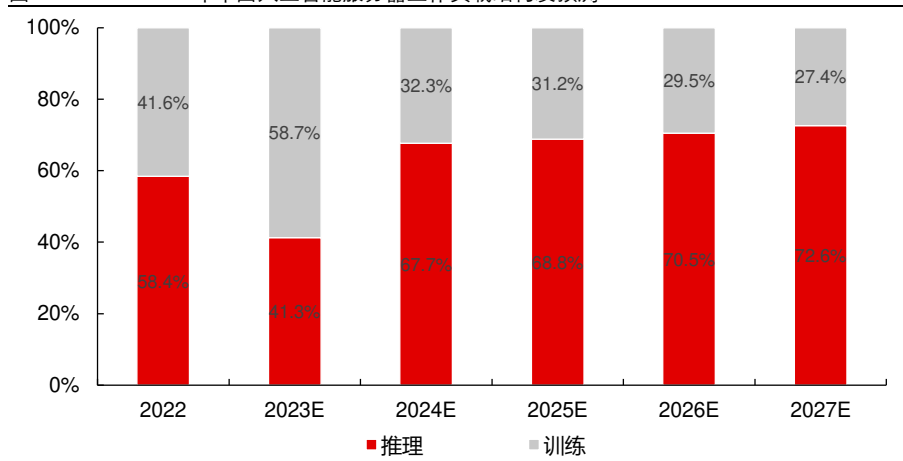
图 24：国内外代表性互联网厂商自研芯片情况

公司	公司logo	2015	2016	2017	2018	2019	2020	2021	2022	2023
谷歌	Google	TPU v1		TPU v2	TPU v3			TPU v4		TPU v5e TPU v5p
微软	Microsoft									Maia 100
AWS	amazon					Inferentia 1 (推理)	Trainium 1 (训练)		Inferentia 2 (推理)	Trainium 2 (训练)
Meta	Meta									MTIA
阿里	阿里巴巴					含光800				
百度	Baidu 百度					昆仑1		昆仑2		
腾讯	Tencent 腾讯					T10 (推理)	i10 (训练)	紫霄/T20/i20		
字节	ByteDance 字节跳动									视频编码芯片

资料来源：谷歌官网，微软官网，亚马逊官网，英伟达官网，36 氪，Anandtech，Trendforce，腾讯新闻，昆仑芯官网，燧原科技官网，鹅厂技术派，火山引擎官网，长江证券研究所

着眼中长期结构，AI 算力需求重心有望逐步由训练向推理转移。根据英伟达 FY24Q4 的电话会披露，公司 2023 年数据中心收入中约 40%来自推理业务。伴随着训练模型的完善与成熟，企业将把更多算力从模型训练转移到 AI 推理工作中，意味着人工智能模型将逐步进入广泛投产模式，处理推理工作负载的算力占比将随之攀升。根据 IDC 数据，2022 年中国数据中心用于推理的服务器的市场份额占比已经过半，达到 58.4%，预计到 2027 年，用于推理的工作负载将达到 72.6%。

图 25：2022-2027 年中国人工智能服务器工作负载结构及预测



资料来源：IDC，浪潮信息，长江证券研究所

国内外因素共振，国产算力份额有望提升

中美科技摩擦加剧，面对产品封锁与份额垄断，发展自主可控的国产算力的紧迫性和重要性得以进一步提升，政策驱动下有望实现进一步加速。近年来，美国政府为了扼制中国高新科技产业发展，采取多种手段对中国的芯片产业进行限制，包括：2022 年 10 月 7 日美国商务部工业与安全局（BIS）对出口中国的 AI 芯片施加限制；2023 年 10 月 17

日宣布计划进一步遏制向中国出售更先进的 AI 芯片；2024 年 1 月 29 日，美政府发布的监管文件中的“客户识别计划”将强制要求美国云计算公司验证外国用户的身份，进而限制中国企业通过访问美国数据中心以训练人工智能模型。在美政府不断对我国算力“卡脖子”的形势下，通过自研并自主供应发展国产算力，达成算力自主将是我国突破美算力封锁的必由之路，或将加速中国人工智能厂商向国产算力迁移。

表 7：美国政府不断限制中国使用美先进算力

时间	事件
2022 年 8 月 31 日	美国芯片厂商英伟达和 AMD 收到通知要求其停止向中国出口 2 种尖端人工智能芯片，分别为英伟达的云端加速器系列 A100 与 H100，以及 AMD 的数据中心级 GPU MI100 与 MI200。
2022 年 10 月 7 日	美国商务部公布了一系列全面出口管制措施，针对中国芯片产业的打压再度升级，包括将某些先进、高性能的计算机芯片和含有此类芯片的计算机产品加入管制清单，对最终用途在中国的超级计算机或半导体开发及生产应用项目增加新的许可证要求等。
2023 年 10 月 17 日	美国商务部工业与安全局（BIS）公布了两项临时最终规则（IFR）扩大或细化了美国对中国内地、香港和澳门的先进计算芯片及相关物项和半导体制造设备的出口管制。
2024 年 1 月 29 日	要求美国云计算公司通过“客户识别计划”来验证其外国用户的身份，进而可能限制中国企业通过美国云计算公司训练人工智能模型。
2024 年 3 月 29 日	美国再次修订防止中国取得美国 AI 芯片和芯片制造设备的规定。

资料来源：路透社，美国商务部工业与安全局官网，长江证券研究所

随着科技自立紧迫性和重要性进一步提升，国产算力将迎来历史性机遇。从国产算力的下游需求来看：一方面，政府智算中心建设持续加码，另一方面运营商招标订单需求旺盛，叠加互联网企业开始转向寻求与国产厂商的合作，国产算力份额有望实现提升。

国内政策频出，国产智算中心加速建设。政府层面多次出台政策支持算力基础设施建设，此前，工业和信息化部、中央网信办、教育部、国家卫生健康委、中国人民银行、国务院国资委等六部门联合印发的《算力基础设施高质量发展行动计划》中提出到 2023 年底建成 30 个智算中心，到 2025 年底建成 50 个，并且算力规模超过 300EFLOPS，智能算力占比达到 35%。目前全国正在建设或提出建设智算中心的城市已经超过 30 个，已投入运营的包括南京智能计算中心、长沙 5A 级智能计算中心等。此外，2024 年 3-4 月，上海市、广东省、北京市相继印发算力基础设施相关文件，均对算力的国产化率提出了相关要求，最低在 50%之上。各地的相关政策是对中央顶层设计的坚定执行，从中央到地方，产业趋势正不断强化，或将推动各地智算中心建设加速。

表 8：2024 年各省份智算中心相关政策梳理

时间	发布单位/省份	政策/事项	智算中心建设重点内容
2024.05	甘肃省通信管理局联合省发改委、省工信厅等六部门	《甘肃算力基础设施高质量发展三年行动计划（2024-2026 年）》	到 2026 年，全省算力规模超过 30EFLOPS。
2024.04	北京市经济和信息化局 北京市通信管理局	《北京市算力基础设施建设实施方案（2024—2027 年）》	1、智算资源供给集群化。改变智算建设“小、散”局面，集中建设一批智算单一大集群，到 2025 年，本市智算供给规模达到 45EFLOPS。 2、智算设施建设自主化。到 2027 年，实现智算基础设施软硬件产品全栈自主可控，整体性能达到国内领先水平，具备 100%自主可控智算中心建设能力。
2024.04	福建省数字福建建设领导小组	《2024 年数字福建工作要点》	统筹布局算力基础设施。加强全省算力资源跨区域统筹布局和统一调度，依托数字福建（长乐、安溪）产业园推进建设存算一体的新型数据中心。规模化、集约化布局建设人工智能计算中心，推动建立算力共享平台。加快布局一批云资源池、边缘云节点等基础设施，推动区域数字化协同发展。
2024.04	江苏省通信管理局会同省发	《江苏省算力基础设施发展专	到 2030 年，全省数据中心机架规模达 120 万标准机架，全省在用总算力超过

	展和改革委员会、省工业和 信息化厅、省数据局	项规划》	50EFLOPS，智能算力占比超过 45%，智算中心突破 20 个，算力算效水平显著提高。
2024.03	上海市通信管理局会同市委 网信办、市发展改革委、市 数据局等十一部门	《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）》	1、智算设施高效协同。到 2025 年，本市智能算力规模超过 30EFlops，占比达到总算力的 50%以上。 2、智算要素自主可控。到 2025 年，本市新建智算中心国产算力芯片使用占比超过 50%，国产存储使用占比超过 50%，服务具有国际影响力的通用及垂直行业大模型设计应用企业超过 10 家。
2024.03	广东省通信管理局等九部门	《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）》	1、到 2025 年，在算力方面，算力规模达到 38EFLOPS，智能算力占比达到 50%。建成智能计算中心 10 个，基本形成算力规模体量与数字化发展需求相适应、算力供给结构与业务需求相匹配的发展格局。 2、力争到 2025 年底，新增国产化算力占比达到 70%，基本形成与广东经济社会数字化发展需要相适应的算力、运力、存力资源体系和供给体系，建成国内领先、国际一流且具有全球影响力的区域级核心算力枢纽。
2024.03	新疆	拟筹建融合算力中心	新疆将构建具备“国家级”算力水平的涵盖超算、智算的融合算力中心。到 2024 年底，建成超算算力（FP64）不低于 200P Flops、智算算力（FP16）不低于 400PFlops 的融合算力中心。
2024.01	山西省通信管理局联合省工 业和信息化厅等七部门	《山西省算力基础设施高质量 发展实施方案》	到 2025 年，全省算力规模超过 9EFLOPS，智能算力占比达到 35%，建成超过 4 个智能计算中心。
2024.01	武汉市人民政府办公厅	《武汉市推进算力基础设施及 应用产业高质量发展行动方案 （2024-2025 年）》	到 2025 年，力争全市算力规模达到每秒 5×10^{18} 浮点运算次数（5EFLOPS），其中高性能算力超过 4.5EFLOPS，即智算算力超过 4.1EFLOPS，超算算力达到 0.4EFLOPS。

资料来源：各省份发改委，中国新疆，长江证券研究所

从运营商招标情况来看，国产算力需求旺盛正当时。着眼于我国 AI 服务器下游应用分布，据 IDC 统计数据显示，2022 年中国 AI 服务器市场按销售额统计，互联网厂商占据 47.5%，运营商占比 19.6%。互联网厂商此前大多采用海外 GPU 芯片的 AI 服务器，而以运营商为代表的国企则更多采用国产 AI 芯片服务器。从运营商招标情况来看，国内三大运营商算力采购以国产厂商为主。以中国移动为例，根据此前发布的《中国移动 2024 年至 2025 年新型智算中心采购_招标公告》，此次招标将采购人工智能服务器 7994 台及配套产品用于智算中心建设。此次总中标金额约为 191.04 亿元（不含税），平均每台人工智能服务器及其相关配套费用约为 238.98 万元，单台服务器价值量较高，采购量级及金额均超出此前市场预期，算力需求高景气得到产业验证。同时该项目公示的中标人全部为国产服务器厂商（服务器搭载国产 AI 芯片），验证了对国产算力芯片服务器的需求旺盛正当时。

表 9：近年来三大运营商 AI 服务器集采招标统计

招标单位	采购计划	AI 服务器 采购台数	标包	分包 采购台数	中标（候选）人	中标份额
中国移动	2024 年至 2025 年新型 智算中心采购项目	7994	特定场景 AI 训练服务器	7994	河南昆仑技术有限公司	21.05%
					四川华鲲振宇智能科技有限责任公司	17.54%
					宝德计算机系统股份有限公司	15.79%
					百信信息技术有限公司	14.04%
					武汉长江计算科技有限公司	12.28%
					神州鲲泰（厦门）信息技术有限公司	10.53%
					湖南湘江鲲鹏信息科技有限责任公司	8.77%
	2023 年至 2024 年新型 智算中心（试验网）集采 项目	1204	标包 1-3	716	-	-
			标包 4	52	新华三技术有限公司	70%
					烽火通信科技股份有限公司	30%

			标包 5	16	中兴通讯股份有限公司	100%
			标包 6	64	新华三技术有限公司	70%
					河南昆仑技术有限公司	30%
			标包 11	356	河南昆仑技术有限公司	256 台
					四川华鲲振宇智能科技有限责任公司	100 台
中国电信	AI 算力服务器（2023-2024 年）集中采购项目	4175	训练型风冷服务器（I 系列）	2073	超聚变数字技术有限公司	-
			训练型液冷服务器（I 系列）	125	超聚变数字技术有限公司	-
			训练型风冷服务器（G 系列）	1048	四川华鲲振宇智能科技有限责任公司	-
			训练型液冷服务器（G 系列）	929	四川华鲲振宇智能科技有限责任公司	-
	天翼云 2023 年服务器及相关服务采购项目	2000	标包 1	1500	浪潮电子信息产业股份有限公司	100%
			标包 2	500	超聚变数字技术有限公司	100%
中国联通	2024 年中国联通人工智能服务器集中采购项目	2503	-	2503	河南昆仑技术有限公司	100%

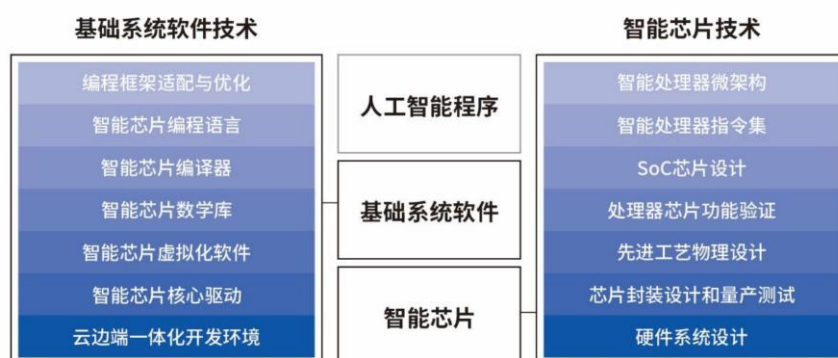
资料来源：C114 通信网，长江证券研究所

互联网厂商开始积极寻求与国产芯片和服务器厂商进行合作。据路透社报道，2023 年百度为 200 台服务器订购了 1600 片华为昇腾 910B 的 AI 算力芯片，助力大模型迭代。随着美国对华芯片封锁加剧，原来作为英伟达客户的国内互联网大厂开始寻找新的出路，未来在政策驱动下将有更多互联网企业转向与国内 AI 芯片服务器厂商合作。

云边端产品升级迭代，软硬件协同筑生态 兼具软硬件核心技术，云边端产品频迭代

寒武纪公司掌握 AI 芯片及系统软件领域关键核心技术。通用型智能芯片及其基础系统软件的研发是极端复杂的系统工程，需要全面掌握核心芯片与系统软件的大量关键技术，技术难度大、涉及方向广。其中处理器微架构与指令集两大类技术属于最底层的核心技术。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。

图 26：寒武纪核心技术框架结构



资料来源：公司公告，长江证券研究所

云边端产品不断迭代，产品矩阵逐步完善。公司以云端芯片为业务核心，延伸至边缘产品线服务客户需求，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，是同时具备人工智能推理和训练智能芯片产品的企业，也是国内少数具有先进集成电路工艺（如 7nm）下复杂芯片设计经验的企业之一。公司研发的寒武纪 1A 处理器是全球首款商用终端智能处理器 IP 产品，思元 100 (MLU100) 芯片是中国首款高峰值云端智能芯片。思元 290 (MLU290) 芯片是寒武纪首款云端训练智能芯片，采用了 7nm 工艺，在 4 位和 8 位定点运算下，理论峰值性能分别高达 1024TOPS、512TOPS。思元 370 (MLU370) 芯片是寒武纪首款采用 Chiplet（芯粒）技术的人工智能芯片，是寒武纪第二代云端推理产品思元 270 算力的 2 倍。

表 10：云边端三大产品线产品迭代梳理

产品线	产品类型	寒武纪主要产品	推出时间
云端产品线	云端智能芯片及加速卡	思元 100 (MLU100) 芯片及云端智能加速卡	2018 年
		思元 270 (MLU270) 芯片及云端智能加速卡	2019 年
		思元 290 (MLU290) 芯片及云端智能加速卡	2021 年
		思元 370 (MLU370) 芯片及云端智能加速卡	2021、2022 年
	训练整机	玄思 1000 智能加速器	2020 年
		玄思 1001 智能加速器	2022 年

边缘产品线	边缘智能芯片及加速卡	思元 220 (MLU220) 芯片及边缘智能加速卡	2019 年
IP 授权及软件	终端智能处理器 IP	寒武纪 1A 处理器	2016 年
		寒武纪 1H 处理器	2017 年
		寒武纪 1M 处理器	2018 年
	基础系统软件平台	寒武纪基础软件开发平台（适用于公司所有芯片与处理器产品）	持续研发和升级以适配新的芯片

资料来源：公司公告，长江证券研究所

图 27：云边端产品布局



资料来源：寒武纪 2023 年度社会责任报告，长江证券研究所

云端智能芯片水平国内领先

云端智能芯片业务为公司业务核心，主要以实体芯片或加速卡的形式应用于各类云服务器或数据中心。云端智能芯片及加速卡是云服务器、数据中心等进行人工智能处理的核心器件，其主要作用是云计算和数据中心场景下的人工智能应用程序提供高计算密度、高能效的硬件计算资源，支撑该类场景下复杂度和数据吞吐量高速增长的人工智能处理任务。

对比国内外主要云端智能芯片产品，公司云端智能芯片处于国内第一梯队。寒武纪、英伟达、华为最新研制的云端芯片产品都已采用 7nm 等先进工艺，在性能功耗比上较为接近。在峰值计算能力方面，英伟达凭借 A100 占据领先地位；在智能计算的基础软件生态上，英伟达的 CUDA 软件生态成熟完备，在该领域处于绝对领先，相对公司、华为等具备显著的优势；在智能计算市场份额与认知度上，英伟达的 GPU 产品也处于行业领先地位。整体上，公司云端芯片在国产芯片中处于靠前位置，作为国产云端芯片第一梯队相较于英伟达仍有一定差距，但在性能等方面正不断实现追赶。

表 11：主要算力芯片参数指标对比

厂商	芯片	可用任务	芯片类型	工艺	算力 TFLOPS @FP16	算力 TFLOPS @FP32	算力 TOPS @int8	功耗 (w)	显存带宽	互联带宽
英伟达	V100	训练+推理	GPU	12nm	130	14	-	300	900GB/s	PCIe 3.0: 64GB/s NVLINK: 300GB/s
	A100	训练+推理	GPU	7nm	312	19.5	624	400	2039GB/s	PCIe 4.0: 64GB/s NVLINK: 600GB/s

	A800	训练+推理	GPU	7nm	312	19.5	624	400	2039GB/s	PCIe 4.0: 64GB/s NVLINK: 400GB/s
	H100	训练+推理	GPU	4nm	1979	67	3958	700	3300GB/s	PCIe 5.0: 128GB/s NVLINK: 900GB/s
华为	昇腾 310	推理	ASIC	12nm	8	-	16	8	-	-
	昇腾 910	训练	ASIC	7nm	320	-	640	310	-	-
寒武纪	MLU270-F4	推理	ASIC	16nm	-	-	128	150	102GB/s	-
	MLU270-S4	推理	ASIC	16nm	-	-	128	70	102GB/s	-
	MLU290-M5	训练	ASIC	7nm	-	-	512	350	1228GB/s	600GB/s
	MLU370-X8	训练+推理	ASIC	7nm	96	24	256	250	614.4GB/s	200GB/s
	MLU370-X4	训练+推理	ASIC	7nm	96	24	256	150	307.2GB/s	-
	MLU370-S4	训练+推理	ASIC	7nm	72	18	192	75	307.2 GB/s	-
百度	昆仑一代	推理	ASIC	14nm	64	-	256	150	512GB/s	-
	昆仑二代	训练+推理	ASIC	7nm	128	-	256	120	512GB/s	-

资料来源：各公司官网、招股说明书及相关公告，长江证券研究所

边缘端、终端产品补全应用场景需求

公司陆续推出终端智能处理器 **1A、1H 和 1M**，累计出货过亿。公司终端智能处理器产品主要以 IP 授权形式应用于智能终端设备中，目前已推出寒武纪 1A、寒武纪 1H 和寒武纪 1M 三代终端智能处理器，覆盖视觉、语音、自然语言处理以及传统机器学习等高度多样化的人工智能处理任务，可集成于各芯片设计厂商的 SoC 芯片产品中，支持公司自研的 Cambricon NeuWare 基础系统软件平台，同时支持各种主流编程框架，广泛应用于包括智能手机、智能摄像头在内的各类智能终端。当前已有多家国内著名芯片设计公司获得了公司终端智能处理器的商业 IP 授权，迄今已集成于超过 1 亿台智能手机及其他智能终端设备中。

公司针对边缘计算推出思元 220 芯片，累计销量突破百万片。边缘计算是近年来兴起的一种新型计算范式，在终端和云端之间的设备上配备适度的计算能力，一方面可有效弥补终端设备计算能力不足的劣势，另一方面可缓解云计算场景下数据隐私、带宽与延时等潜在问题。边缘计算范式和人工智能技术的结合将推动智能制造、智能零售、智能教育、智能家居、智能电网等众多领域的高速发展。公司于 2019 年 11 月推出了边缘智能芯片思元 220 及相应的 M.2 加速卡，可支持边缘计算场景下的智能数据分析与建模、视觉、语音、自然语言处理等多样化的人工智能应用。思元 220 的推出，标志着公司已经具备了从终端、边缘端到云端完整的智能芯片产品线。自发布以来，思元 220 累计销量突破百万片。

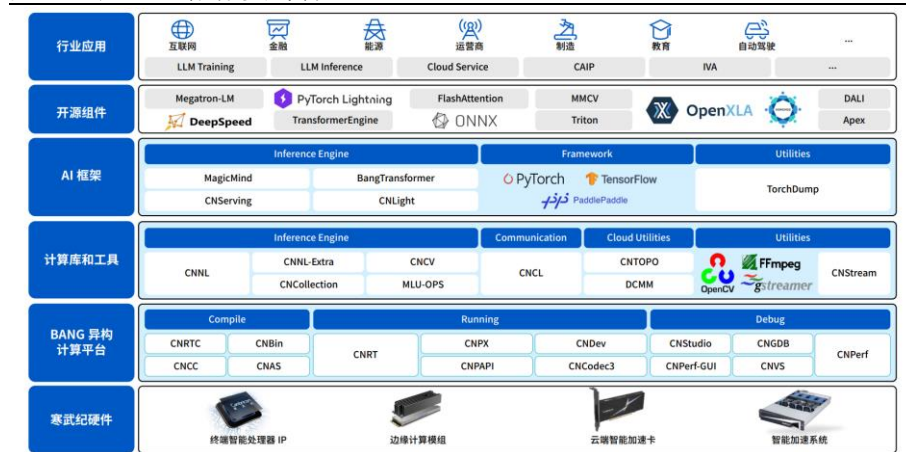
软硬件协同完善生态，智算集群优势尽显

以软件平台为核心打造云边端统一的 AI 开发生态

公司重视软件平台的生态发展，推出软件开发平台 **Cambricon NeuWare** 作为打造云边端统一的 AI 开发生态的核心部件。Cambricon NeuWare 是寒武纪专门针对其云、边、端的智能处理器产品打造的软件开发平台。其采用云边端一体、训推一体架构，可同时

支持寒武纪云、边、端的全系列产品。寒武纪终端 IP、边缘端芯片、云端芯片共享同样的软件接口和完备生态，可以方便地进行智能应用的开发，迁移和调优。

图 28：寒武纪基础软件系统平台 Cambricon NeuWare



资料来源：寒武纪 2023 年度社会责任报告，长江证券研究所

训练端，其训练软件平台支持基于主流开源框架原生分布式通信方式，同时也支持 **Horovod 开源分布式通信框架**，来实现从单卡到集群的分布式训练任务。支持多种网络拓扑组织方式，并完整支持数据并行、模型并行和混合并行的训练方法。该训练软件平台还支持丰富的图形图像、语音、推荐以及 NLP 训练任务。通过底层算子库 **CNNL** 和通信库 **CNCL**，在实际训练业务中达到业界领先的硬件计算效率和通信效率。同时提供模型快速迁移方法，帮助用户快速完成现有业务模型的迁移。

推理端，公司推出推理加速引擎 **MagicMind** 提升产品性价比。**MagicMind** 是寒武纪全新打造的推理加速引擎，也是业界首个基于 MLIR 图编译技术达到商业化部署能力的推理引擎。借助 **MagicMind**，用户仅需投入极少的开发成本，即可将推理业务部署到寒武纪全系列 products 上，并获得颇具竞争力的性能。

图 29：Cambricon NeuWare 训练软件平台



资料来源：公司官网，长江证券研究所

图 30：推理加速引擎 MagicMind



资料来源：公司官网，长江证券研究所

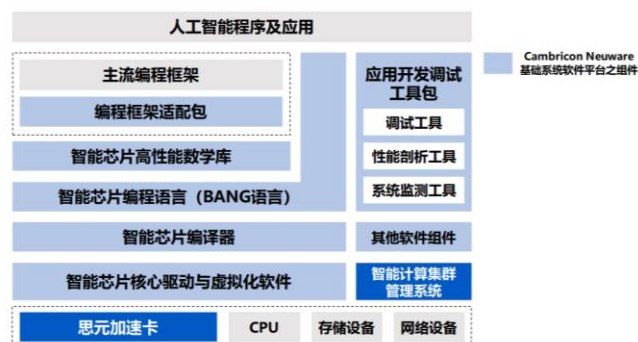
整体上，公司提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。未来有望在软硬件的协同发展下，进一步打造自身核心竞争力，完善产品生态，实现全新突破。

受益于软硬件协同，智能计算集群系统优势显著

公司智能计算集群系统的核心是自研的思元系列智能芯片加速卡、基础系统软件平台，以及智能计算集群管理系统。公司通过将自有云端智能芯片加速卡产品与产业链其他厂商提供的服务器、存储设备和网络设备等硬件设施集成，为整个计算集群提供智能计算能力支撑；以自研的 Cambricon NeuWare 基础系统软件平台为基础，结合客户需求搭建定制化的应用管理平台，提供系统运维、资源调度、应用管理等功能。

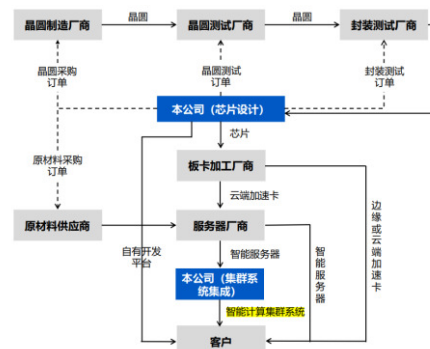
定制化开发充分发挥技术优势，软件平台助力产品提升竞争力。公司智能计算集群系统一般根据客户要求定制化开发，最大限度地发挥思元系列芯片及加速卡产品的技术优势和特点，降低了客户使用和维护复杂计算集群设备的难度和成本，将计算能力以云计算的形式输出，降低了用户开发、部署智能应用的门槛，为智能应用程序的维护、升级提供了有力支持。同时，公司智能计算集群系统具有实时监控运行状态、迅速定位系统异常、故障快速恢复等优势，具有较强的市场竞争力。公司 Cambricon NeuWare 基础软件系统平台还具备优秀的可扩展性，能有效支撑众多用户的智能计算需求，大幅提升智能硬件的利用率，同时提升公司在人工智能数据中心市场的竞争力。

图 31：智能计算集群系统软硬件总体架构



资料来源：首次公开发行股票并在科创板上市招股说明书，长江证券研究所

图 32：智能计算集群系统整体业务流程



资料来源：首次公开发行股票并在科创板上市招股说明书，长江证券研究所

项目经验助力智能计算集群系统业务进一步推进。公司自 2019 年以来屡次中标智能计算中心项目，具备充分的项目经验和技術储备，积极参与台州、沈阳两地的算力基础设施建设项目，以公司的训练整机产品作为核心计算设备，集成配套的软硬件，最终形成智能计算集群系统交付给客户，2023 年智算集群系统业务共实现收入 6.05 亿元。随着人工智能技术的不断进步，对算力基础设施的需求逐年提升。公司的智能计算集群系统业务已在多个城市实现拓展，国内市场占有率已然处于第一梯队，公司产品的竞争力和品牌影响力与日俱增。已完成的项目对公司持续推进智能计算集群系统业务奠定了良好的标杆作用。

表 12：2019-2023 年寒武纪中标智算相关项目梳理

中标日期	中标项目	采购单位	中标金额 (亿元)
2019 年 8 月 28 日	协同创新港 2 号研发中试楼数据中心人工智能计算平台项目	西安沣东仪享科技服务有限公司	0.92
2019 年 11 月 26 日	横琴先进智能计算平台（二期）采购项目	珠海市横琴新区管理委员会商务局	4.44
2020 年 12 月 18 日	南京智能计算中心项目（一期）智能计算设备采购项目	南京市科技创新投资有限责任公司	3.00

2021 年 11 月 30 日	江苏昆山高新技术产业投资发展有限公司关于智能计算中心基础设施建设项目	江苏昆山高新技术产业投资发展有限公司	5.09
2022 年 11 月 18 日	南京智能计算中心项目（二、三期）设备采购项目	南京市科技创新投资有限责任公司	5.01
2023 年 6 月 30 日	沈阳市汽车城新型基础设施建设项目-智能计算中心	沈阳市大东区城市建设局	1.55
2023 年 8 月 17 日	浙东南数字经济产业园数字基础设施提升工程（一期）项目	台州市黄岩置成物产管理有限公司	5.28

资料来源：公司中标公告，公司招股说明书，全国公共资源交易平台，长江证券研究所

风险提示

- 1、需求不及预期的风险。当前全球范围内 AI 技术正处于快速变革迭代中，且商业模式仍在探索，是否能最终实现技术上的进一步突破和商业闭环存在不确定性，若生成式大模型发展速度放缓，则存在 AI 算力需求不及预期的风险。
- 2、产能供应不足的风险。由于寒武纪被美国列入实体清单，而高端 AI 芯片的生成又依赖于国内外先进芯片制程，美国的限制或导致公司高端 AI 芯片生产和供应上存在较大的不确定性，或存在无法批量供货的风险。
- 3、行业竞争加剧的风险。除海外英伟达等算力巨头，国内科技大厂和互联网平台以及各类智算芯片创业公司均加大对 AI 芯片研发的投入，行业竞争格局变化较快，激励竞争中或存在竞争加剧，格局恶化的风险。
- 4、盈利预测假设不成立或不及预期的风险。在对公司进行盈利预测及投资价值分析时，我们基于算力芯片行业需求及国产化政策支持作为成长支撑，对未来公司营收业绩进行预测。若上述假设不成立或者不及预期则我们的盈利预测及估值结果可能出现偏差，具体影响包括但不限于公司业绩不及我们的预期、估值结果偏高等，极端悲观假设下，若公司业务拓展不及预期、主业增速放缓、产品推广不及预期等，则公司未来收入/业绩增速或受影响。

表 13：公司利润敏感性分析

单位：亿元	基准情形				悲观情形			
	2023A	2024E	2025E	2026E	2023A	2024E	2025E	2026E
营业收入	7.09	14.00	25.03	36.29	7.09	13.30	21.27	25.40
YoY	-2.7%	97.4%	78.8%	45.0%	-2.7%	87.51%	59.93%	19.41%
毛利率	69.16%	70.68%	70.78%	70.78%	69.16%	67.59%	65.62%	66.61%
归母净利润	-8.48	-4.50	1.97	3.97	-8.48	-4.69	0.72	1.88
YoY	-	46.90%	-	101.69%	-	44.76%	-	161.76%

资料来源：Wind，长江证券研究所

财务报表及预测指标

利润表 (百万元)					资产负债表 (百万元)				
	2023A	2024E	2025E	2026E		2023A	2024E	2025E	2026E
营业总收入	709	1400	2503	3629	货币资金	3954	3953	3670	3505
营业成本	219	411	731	1060	交易性金融资产	700	700	700	700
毛利	491	990	1771	2569	应收账款	644	1138	2006	2896
%营业收入	69%	71%	71%	71%	存货	99	186	331	480
营业税金及附加	4	17	30	43	预付账款	148	53	95	138
%营业收入	1%	1%	1%	1%	其他流动资产	102	83	119	156
销售费用	82	98	125	181	流动资产合计	5648	6113	6922	7876
%营业收入	12%	7%	5%	5%	长期股权投资	230	230	230	230
管理费用	154	294	425	581	投资性房地产	0	0	0	0
%营业收入	22%	21%	17%	16%	固定资产合计	142	62	1	1
研发费用	1118	1120	1126	1452	无形资产	150	150	150	150
%营业收入	158%	80%	45%	40%	商誉	0	0	0	0
财务费用	-45	-79	-79	-73	递延所得税资产	0	0	0	0
%营业收入	-6%	-6%	-3%	-2%	其他非流动资产	249	311	311	311
加: 资产减值损失	-138	0	0	0	资产总计	6418	6865	7613	8567
信用减值损失	-134	-130	-130	-130	短期贷款	0	0	0	0
公允价值变动收益	0	0	0	0	应付款项	237	376	670	972
投资收益	74	0	0	0	预收账款	0	0	0	0
营业利润	-876	-464	214	437	应付职工薪酬	160	296	366	424
%营业收入	-123%	-33%	9%	12%	应交税费	30	84	150	218
营业外收支	1	0	0	0	其他流动负债	37	170	284	400
利润总额	-875	-464	214	437	流动负债合计	463	926	1470	2014
%营业收入	-123%	-33%	9%	12%	长期借款	0	0	0	0
所得税费用	3	2	11	26	应付债券	0	0	0	0
净利润	-878	-466	204	411	递延所得税负债	0	0	0	0
归属于母公司所有者的净利润	-848	-450	197	397	其他非流动负债	225	563	563	563
少数股东损益	-30	-16	7	14	负债合计	689	1490	2034	2577
EPS (元)	-2.07	-1.08	0.47	0.95	归属于母公司所有者权益	5650	5311	5508	5905
现金流量表 (百万元)					少数股东权益	80	64	71	85
	2023A	2024E	2025E	2026E	股东权益	5730	5375	5579	5990
经营活动现金流净额	-596	-381	-283	-165	负债及股东权益	6418	6865	7613	8567
取得投资收益收回现金	80	0	0	0	基本指标				
长期股权投资	-7	0	0	0		2023A	2024E	2025E	2026E
资本性支出	-100	0	0	0	每股收益	-2.07	-1.08	0.47	0.95
其他	452	-62	0	0	每股经营现金流	-1.43	-0.91	-0.68	-0.39
投资活动现金流净额	425	-62	0	0	市盈率	—	—	544.73	270.08
债券融资	0	0	0	0	市净率	9.95	20.18	19.45	18.15
股权融资	1804	112	0	0	EV/EBITDA	-	-	526.97	285.20
银行贷款增加 (减少)	0	0	0	0	总资产收益率	-13.2%	-6.6%	2.6%	4.6%
筹资成本	0	0	0	0	净资产收益率	-15.0%	-8.5%	3.6%	6.7%
其他	-147	330	0	0	净利率	-119.6%	-32.2%	7.9%	10.9%
筹资活动现金流净额	1657	442	0	0	资产负债率	10.7%	21.7%	26.7%	30.1%
现金净流量 (不含汇率变动影响)	1486	-1	-283	-165	总资产周转率	0.12	0.21	0.35	0.45

资料来源: 公司公告, 长江证券研究所

投资评级说明

行业评级	报告发布日后的 12 个月内行业股票指数的涨跌幅相对同期相关证券市场代表性指数的涨跌幅为基准，投资建议的评级标准为：
看好	相对表现优于同期相关证券市场代表性指数
中性	相对表现与同期相关证券市场代表性指数持平
看淡	相对表现弱于同期相关证券市场代表性指数
公司评级	报告发布日后的 12 个月内公司的涨跌幅相对同期相关证券市场代表性指数的涨跌幅为基准，投资建议的评级标准为：
买入	相对同期相关证券市场代表性指数涨幅大于 10%
增持	相对同期相关证券市场代表性指数涨幅在 5%~10%之间
中性	相对同期相关证券市场代表性指数涨幅在-5%~5%之间
减持	相对同期相关证券市场代表性指数涨幅小于-5%
无投资评级	由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。

相关证券市场代表性指数说明：A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准。

办公地址

上海

Add /虹口区新建路 200 号国华金融中心 B 栋 22、23 层
P.C / (200080)

武汉

Add /武汉市江汉区淮海路 88 号长江证券大厦 37 楼
P.C / (430015)

北京

Add /西城区金融街 33 号通泰大厦 15 层
P.C / (100032)

深圳

Add /深圳市福田区中心四路 1 号嘉里建设广场 3 期 36 楼
P.C / (518048)

分析师声明

本报告署名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。

法律主体声明

本报告由长江证券股份有限公司及/或其附属机构（以下简称「长江证券」或「本公司」）制作，由长江证券股份有限公司在中华人民共和国大陆地区发行。长江证券股份有限公司具有中国证监会许可的投资咨询业务资格，经营证券业务许可证编号为：10060000。本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格证书编号已披露在报告首页的作者姓名旁。

在遵守适用的法律法规情况下，本报告亦可能由长江证券经纪（香港）有限公司在香港地区发行。长江证券经纪（香港）有限公司具有香港证券及期货事务监察委员会核准的“就证券提供意见”业务资格（第四类牌照的受监管活动），中央编号为：AXY608。本报告作者所持香港证监会牌照的中央编号已披露在报告首页的作者姓名旁。

其他声明

本报告并非针对或意图发送、发布给在当地法律或监管规则下不允许该报告发送、发布的人员。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。本报告内容的全部或部分均不构成投资建议。本报告所包含的观点、建议并未考虑报告接收人在财务状况、投资目的、风险偏好等方面的具体情况，报告接收者应当独立评估本报告所含信息，基于自身投资目标、需求、市场机会、风险及其他因素自主做出决策并自行承担投资风险。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。报告中的信息或意见并不构成所述证券的买卖出价或征价，投资者据此做出的任何投资决策与本公司和作者无关。本研究报告并不构成本公司对购入、购买或认购证券的邀请或要约。本公司有可能会与本报告涉及的公司进行投资银行业务或投资服务等其他业务(例如:配售代理、牵头经办人、保荐人、承销商或自营投资)。

本报告所包含的观点及建议不适用于所有投资者，且并未考虑个别客户的特殊情况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。投资者不应以本报告取代其独立判断或仅依据本报告做出决策，并在需要时咨询专业意见。

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、价值及投资收入可升可跌，过往表现不应作为日后的表现依据；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；本公司不保证本报告所含信息保持在最新状态。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。

本报告版权仅为本公司所有，本报告仅供意向收件人使用。未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布给其他机构及/或人士（无论整份和部分）。如引用须注明出处为本公司研究所，且不得对本报告进行有悖原意的引用、删节和修改。刊载或者转发本证券研究报告或者摘要的，应当注明本报告的发布人和发布日期，提示使用证券研究报告的风险。本公司不为转发人及/或其客户因使用本报告或报告载明的内容产生的直接或间接损失承担任何责任。未经授权刊载或者转发本报告的，本公司将保留向其追究法律责任的权利。

本公司保留一切权利。