2024
中国智能算力行业白皮书

White Paper of China's Artificial Intelligence Computing Power in 2024

天眾智算

FROST & SULLIVAN 沙利文

前言

沙利文联合天罡智算谨此发布《2024年中国智能算力行业白皮书》。本报告旨在分析在中国智能算力市场的现状、应用前景、技术动向及发展趋势，并识别智能算力市场竞争态势，反映该细分市场领袖梯队企业的差异化竞争优势。

名词解释

- **AI**：人工智能
- **生成式AI**：人工智能领域中的一种技术，让计算机系统能够生成新的、原创的内容，这些内容在没有直接编程或提供具体指令的情况下，模仿或类似于真实世界的数据
- **智算**：智能算力，指专门设计用于支持人工智能应用的计算能力，通常由高性能的处理器和加速器组成
- **大模型**：大语言模型，是人工智能领域中一类具有大量参数的自然语言处理（NLP）模型。这些模型通过深度学习技术训练，能够理解和生成人类语言
- **FLOPS**：每秒浮点运算次数，亦称每秒峰值速度，即每秒所执行的浮点运算次数。
- **TFLOPS**：每秒执行浮点运算10^12次
- **PFLOPS**：每秒执行浮点运算10^15次
- **EFLOPS**：每秒执行浮点运算10^18次
- **Token**：指代文本中的最小单位

摘要

- 智能算力，是数字经济发展的重要基础性资源。短期来看，受制于美国的科技禁运政策，长期来看，国产人工智能芯片在设计和制造方面还存在技术差距，以及我国人工智能企业大模型原创力的暂时性缺失，我国距离实现智算资源的完全国产化还有一段距离。为了谋求可用算力资源在物理空间的释放和高效利用，国家层面持续推进“东数西算”工程的实际落地和实践，建设全国一体化算力网络枢纽节点。因此，智算产业的相关布局会提升到国家未来科技发展的战略性高度。

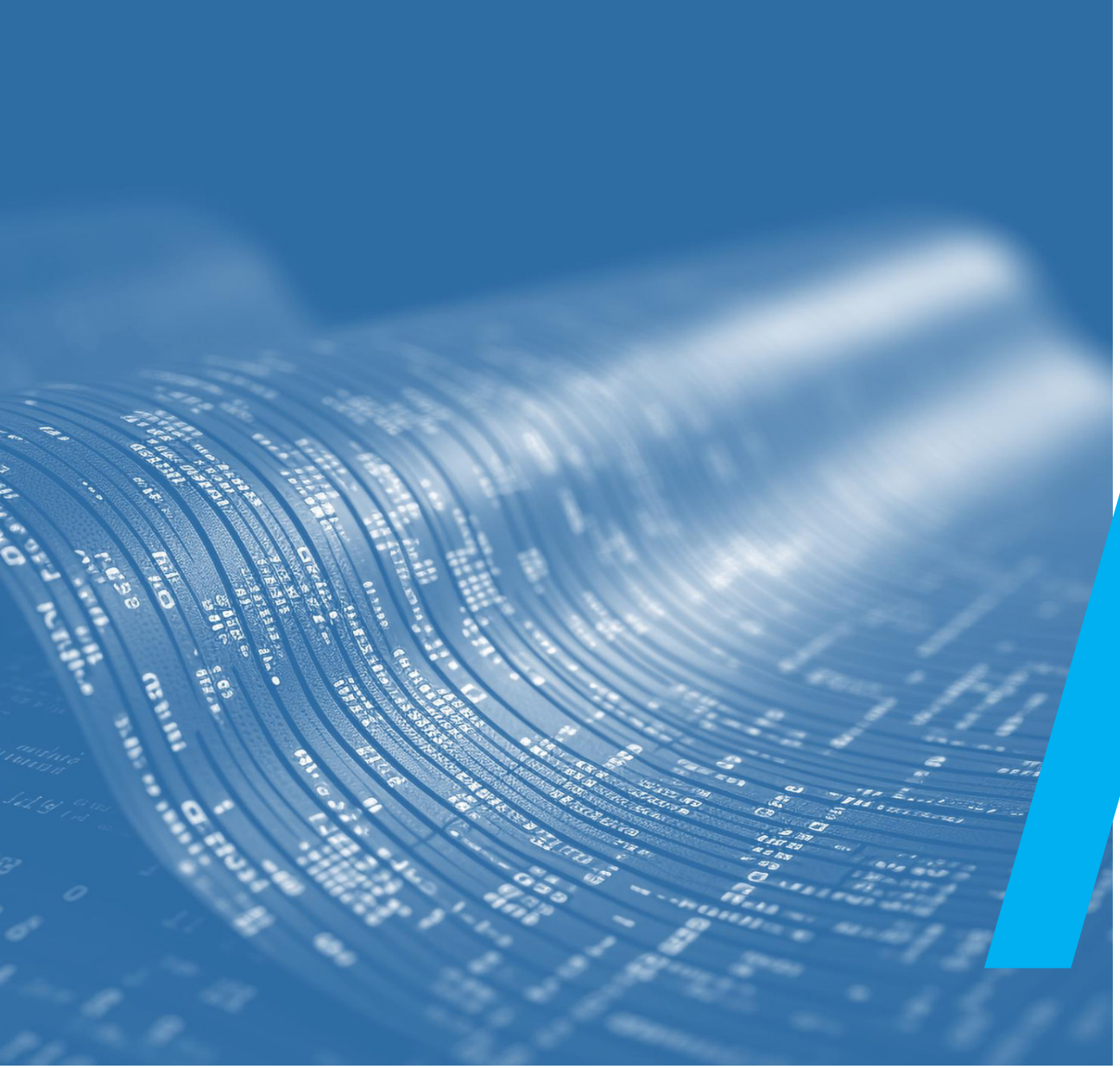
- 梵数智算（以下简称“公司”或者“梵数”），敏锐捕捉到该政策性产业中蕴含的商业机遇，以轻资产模式入局智算产业，构建集GPU计算、存储、网络为一体的一站式智算资源交易平台—天罡智算，由智算资源的供需双方自由报价，平台方进行智能化匹配，实现企业或者个人用户对高性能智算资源的弹性化采买，即联、即取、即用。

- 由于我国智算和智算租赁市场尚处于起步阶段，参与者类型众多，商业模式尚不明确，市场需求仍在探索，但都面临者在算力资源管理的复杂性，算力资源的异构性，算力资源的安全性，算力资源的可靠性以及算力资源商业模式的可行性和稳健性方面的问题和挑战。纵观目前市场上公开的研究资料不难发现，此前市场上大多研究机构的视角主要聚焦于智算需求侧的分析，而对于供给侧的探讨却相对匮乏，略有涉及也大多一笔带过，这无疑限制了对市场全貌的深入理解，以及在商业策略上的精准制定。

- 因此，为了更加准确地把握市场发展脉络，及时掌握市场最新动态，沙利文联合梵数旗下智算交易平台—天罡智算，对智算产业所处的宏观环境，产业链的结构，以及智算租赁产业的市场容量，收入模式，成本结构，竞争格局和潜在下游需求，逐一进行梳理，拆解和分析，尤其是将研究视野拓展至智能算力的供给侧，通过深入剖析不同类型市场参与者切入智算赛道的方式，商业模式，差异化优势和劣势，形成全面的市场洞察。

CONTENTS

目录



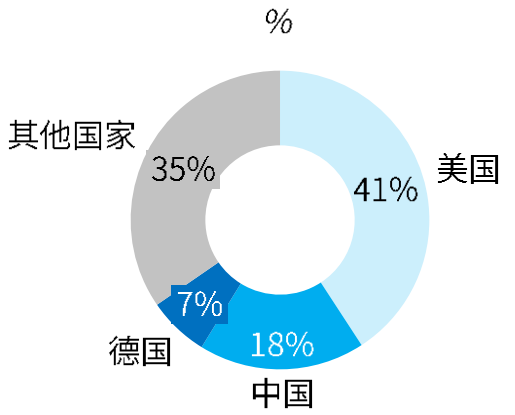
1

- AI时代
人工智能发展概览

全球数字经济规模

2023年，全球数字经济的规模达45万亿美元，占GDP的比重高达44%。相较于2019年大幅提升八个百分点，数字经济已经成为全球经济发展的强大引擎。在全球版图上，美国、中国和德国以其卓越的数字经济实力，形成三足鼎立的领导格局

全球数字经济规模占比，2023

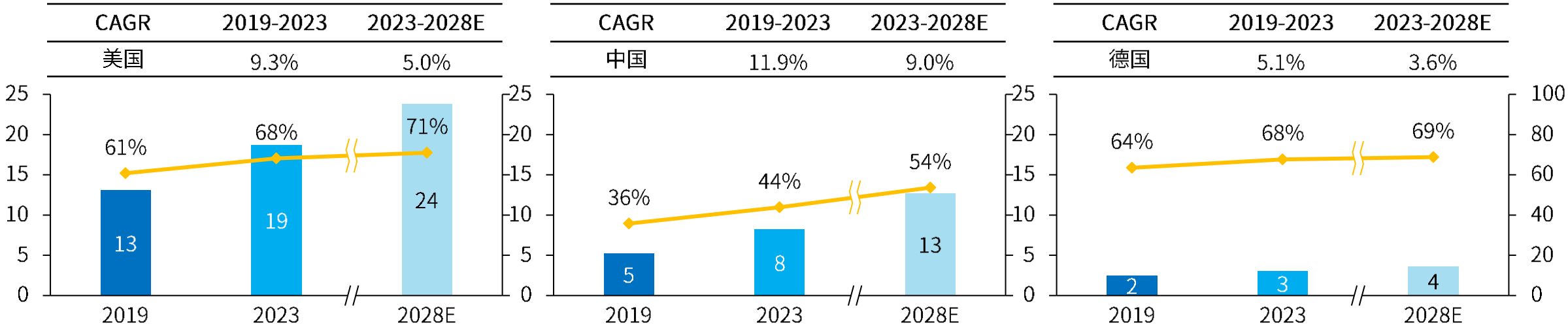


全球数字经济三极格局

- **美国**：美国在产业规模、产业链完整度、数字技术研发实力和数字企业全球竞争力等方面稳居世界前列，因此，数字经济规模以大幅优势领先其他国家，2023年规模达到约19万亿美元，占本国GDP68%。
- **中国**：凭借海量数据资源优势，中国数字经济顶层设计和体制机制建设逐步完善，数字中国建设取得显著成就。2019年至2023年，中国数字经济年复合增长率达到约11.9%。
- **欧盟**：凭借其卓越的科技和创新资源，以及在数字治理上的领先地位，形成了与中美两强优势互补的第三极。

全球主要国家数字经济规模和其在GDP中的占比

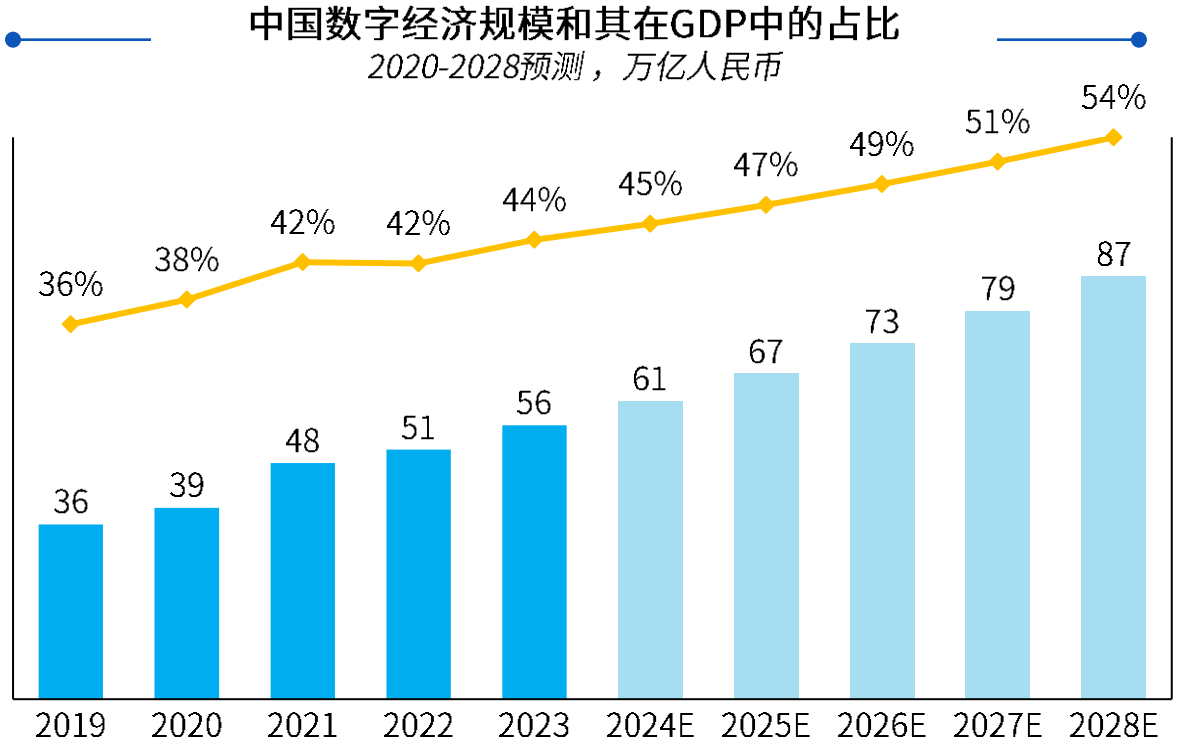
2019-2028预测，万亿美元，%



来源：弗若斯特沙利文

中国数字经济规模

受益于大力推动产业数字化转型的国家战略，中国数字经济规模增速显著超越美国和德国，年复合增长率达到约11.9%，整体规模在2023年实现约56万亿人民币



中国数字经济稳步增长，核心产业贡献显著

- 中国数字经济规模稳步增长，2023年数字经济规模达56万亿人民币，同比增长9.8%，占GDP比重43%左右。
- 2023年中国数字经济核心产业的增加值占GDP比重10%左右，其中以云计算、大数据、物联网等为代表的数字技术为主要增长点。

数字经济发展趋势

基础设施与产业生态融合	数字基础设施建设将与产业数字生态实现更深层次的融合，加速云计算等算力基础设施，区块链等数据基础设施，人工智能等应用基础设施的建设。
经济基础设施布局全面加速	宏观层面，政府加快建立全国统一的数据要素市场规则；微观层面，行业领军企业将积极探索新型数字化生产关系，推动经济基础设施的优化升级。
释放数字要素的市场需求	在数字赋能和消费升级的背景下，数字需求的加速释放将成为推动数字经济发展的新引擎，促进产品和服务的创新，推动产业的数据化、在线化、智能化。
绿色化数字化协同发展	企业以数字技术为核心手段推动绿色化转型；绿色发展对数字采集提出更高要求，牵引数字技术不断升级。数字化绿色化协同发展促进经济发展和生态保护双赢。

制约因素

- 数字基础设施有待进一步优化升级
- 数字化发展水平存在较大的区域差距
- 数字关键核心技术自主可控水平偏低
- 数据安全体系尚未健全，存在数据安全风险

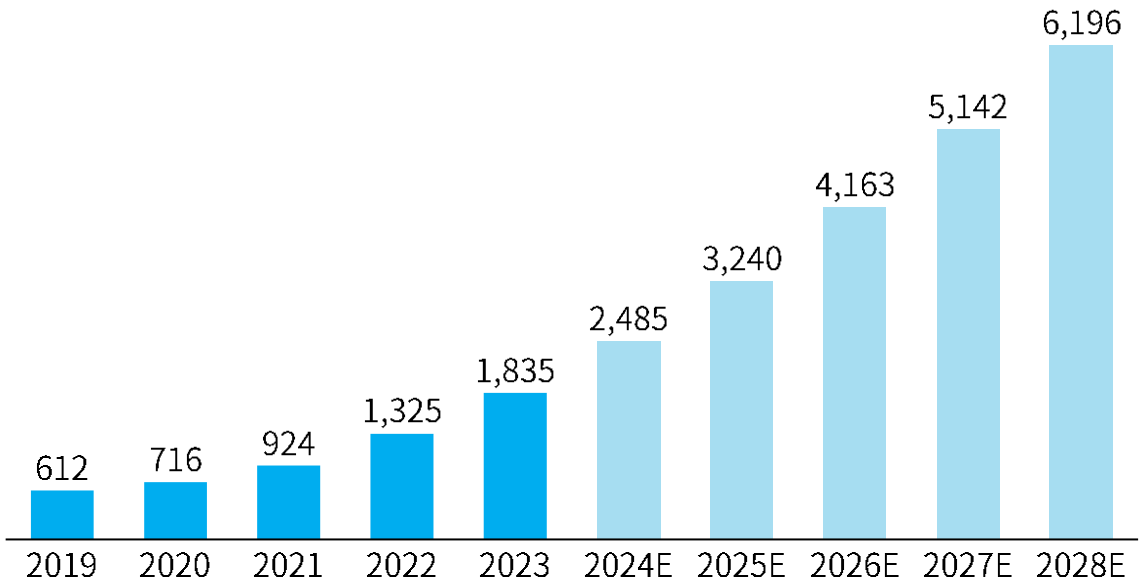
来源：弗若斯特沙利文

全球人工智能市场投资总规模

人工智能产业，作为数字经济生态中的技术推进器和创新加速器，受到全球企业的高度重视和持续性的技术资金加码，预计到2028年，全球人工智能市场的技术投资总额将超过6,000亿美元

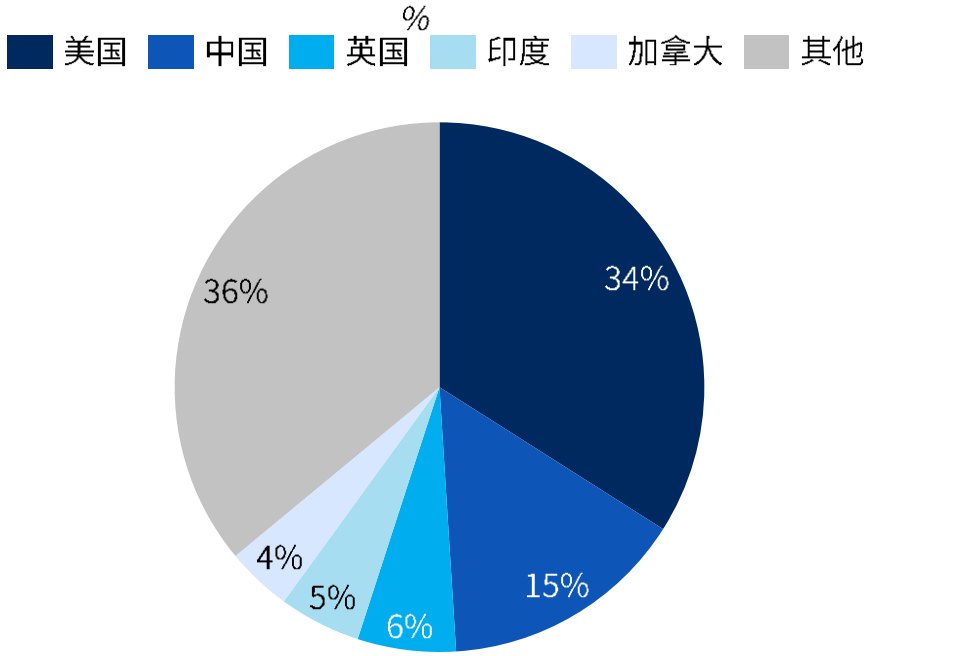
全球人工智能市场总投资额

2019-2028预测，亿美元



备注：总投资规模指企业在包括硬件、软件和服务在内的人工智能市场的技术投资总额

全球人工智能企业数量国家分布，2023



全球人工智能市场蓬勃发展，四大关键领域引领未来增长

- 2023 年全球人工智能市场总投资额达1835亿美元，同比增长38.5%，预计2023年至2028年全球全球人工智能市场规模将保持28%的年复合增长。
- 全球人工智能领域的投资预计将重点聚焦在四个关键业务板块：AI模型开发企业、AI基础设施提供商、AI应用软件开发商，以及企业终端用户。

全球人工智能企业国家分布呈中美主导格局

- 截至2023年第三季度，全球人工智能企业达29,542 家，其中美国和中国的人工智能企业数量占全球总数的49%，成为人工智能领域的引领者。
- 英国、印度、加拿大、德国、以色列、法国、韩国、新加坡在人工智能领域展现创新活力，位列第二梯队，但仍与中美存在较大差距。

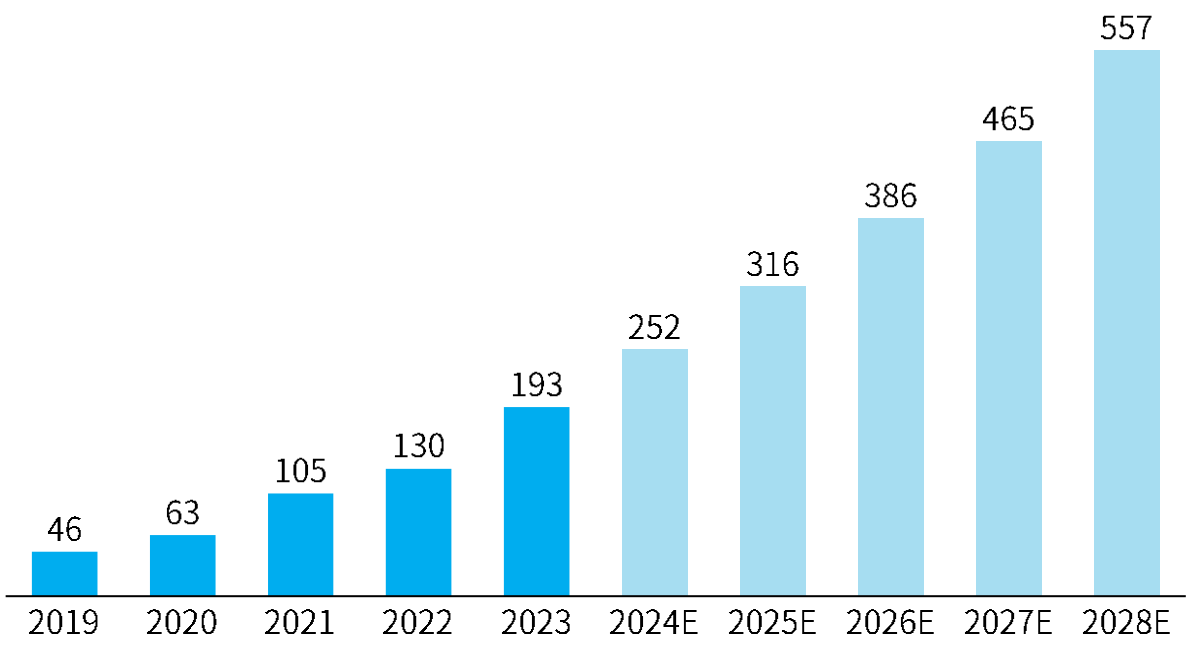
来源：弗若斯特沙利文

中国人工智能市场投资总规模

中国人工智能领域技术投资活跃，投资额在2023年突破193亿美元，同比增长48.2%，创下历史新高。同年，投融资领域有所回温，主要集中在生活服务，智慧医疗，智能制造和汽车物流领域

中国人工智能市场总投资额

2019-2028预测，亿美元



中国人工智能领域投资活跃，软件市场增速领跑人工智能市场

- 2023年中国人工智能市场总投资额突破190亿美元，占全球总量的10.5%，2019-2023年的年复合增长率为43.4%。
- 从技术层面进行分析，中国市场的投资主要集中在硬件领域，其占市场总规模的60%以上；其次投资主要流向软件领域，投资增速位于技术市场首位。

中国人工智能应用领域投资事件和投资额情况（2012-2023）

生活服务

共7,223起投资事件

总投资额为25,301亿人民币

智慧医疗

共2,264起投资事件

总投资额为6,071亿人民币

智能制造

共1,776起投资事件

总投资额为4,904亿人民币

智能汽车

共1,009起投资事件

总投资额为7,540亿人民币

物流仓储

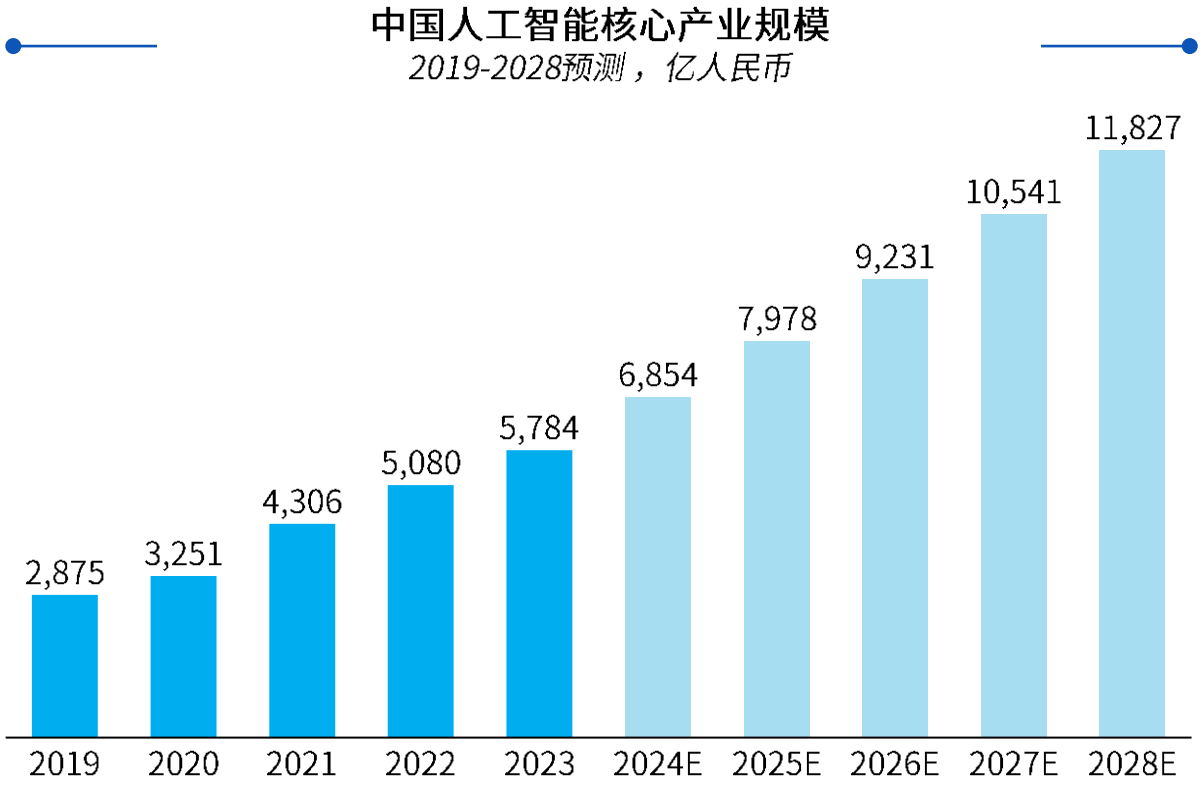
共786起投资事件

总投资额为4,252亿人民币

来源：弗若斯特沙利文

中国人工智能核心产业规模

中国人工智能核心产业规模2023年达5,784亿元人民币，同比增长13.9%，企业数量超过4300家，技术创新活跃，产业体系完备，为各关联层行业提供了良好的基础



中国人工智能核心产业跃升发展

- 2023年中国人工智能核心产业规模达5784亿元，同比增长13.9%，2019-2023年的年复合增长率为19.1%。
- 中国人工智能核心产业下游运用涵盖无人机、语音识别、图像识别、智能机器人、智能汽车、虚拟现实等领域，取得了一系列突破性进展和标志性成果。

人工智能应用场景

生活服务

- 人工智能广泛应用于文化、旅游、体育、健康、养老、教育等生活服务领域，实现服务过程的可视化和可追踪性，显著提升服务业的效率和效益，扩大服务供给，创造出新的产品、服务和商业模式，满足了不同消费群体的需求。

智慧医疗

- 人工智能涵盖了风险评估、筛查、诊断、治疗选择等环节，通过机器学习、深度学习和神经网络技术提高医疗服务效率和诊断准确性。此外，人工智能通过降低医疗成本有效缓解了优质医疗资源分布不均的问题。

智慧制造

- 计算机视觉、机器学习和云部署赋能工业企业的预测、生产、管理、决策，提升各环节智能化水平，加速人工智能在制造业全流程的融合应用，实现从刚性生产到柔性制造的提质增效。

智慧汽车

- 借助鸟瞰视角、自然语言处理等AI大模型技术，人工智能优化了从产品设计、生产流程到供应链管理的各个环节，推动了汽车智能化的进程，其主要体现形式为自动驾驶和智能座舱两个关键领域。

物流仓储

- 通过智能搜索、推理规划和智能机器人等技术，物流行业在运输、仓储、分拨、配送等环节实现了物流资源共享、过程协同以及高度自动化，提升了物流综合服务能力和规模化运行效率。

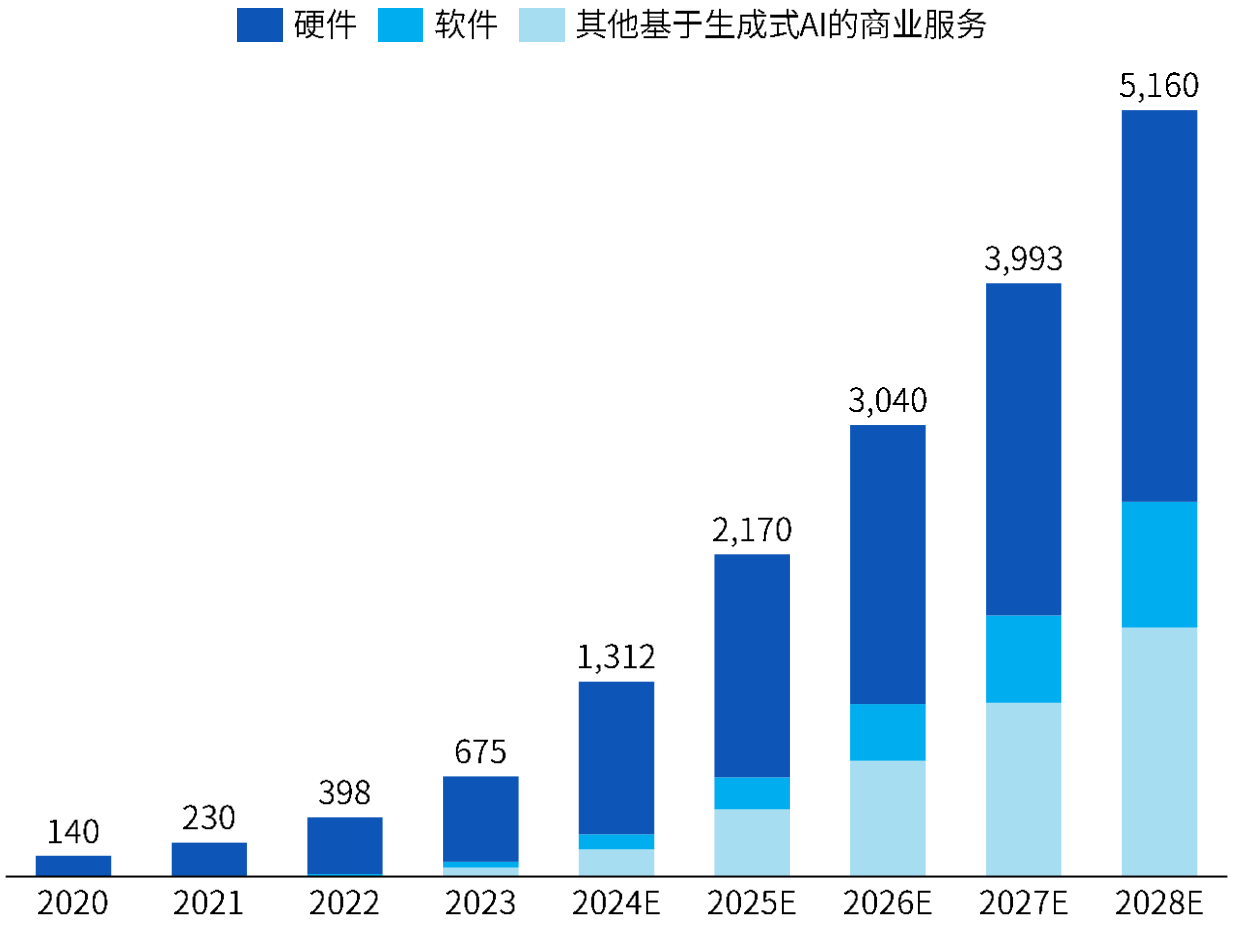
来源：弗若斯特沙利文

全球生成式人工智能应用市场规模

2022年是生成式人工智能技术发展元年，ChatGPT的面世标志着强人工智能拐点的出现，将引发新一轮人工智能革命，2023年全球生成式人工智能的产业规模达到约675亿美元，同比增长高达70%

全球生成式人工智能应用市场规模及拆分

2020-2028预测，亿美元



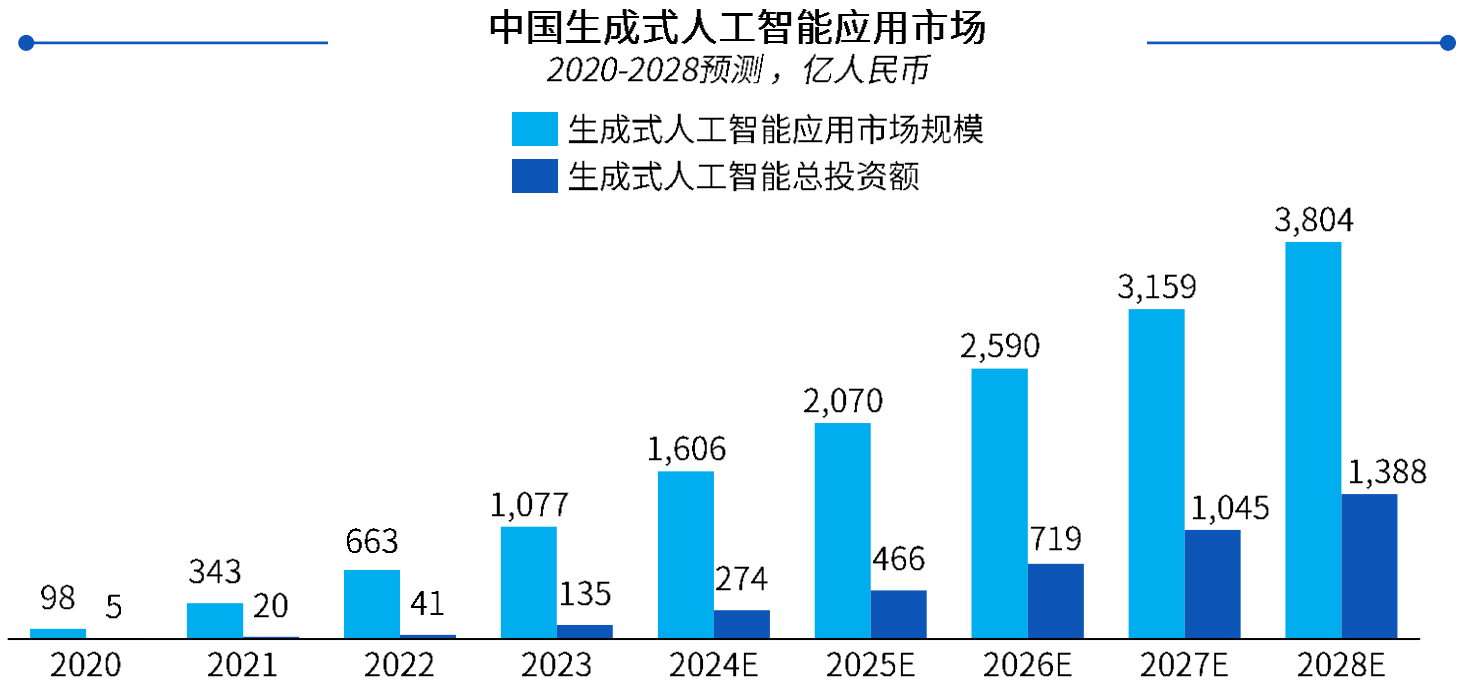
- 人工智能按照其模型类别可被分为**生成式人工智能**与**判别式人工智能**，他们在发展程度、技术角度以及应用方向上都有着较为明显的区别。
- 当前，全球生成式人工智能应用市场规模自2020年开始迎来了快速发展，整体市场规模在2023年大约为675亿美元，预计2028年将增至5,160亿美元，期间年复合增速约为50.2%。其中来自**硬件的占比最高**，而基于生成式人工智能应用的**商业服务增速最快**，这样的高速发展极大的拉动了对于智能算力的需求。

生成式人工智能	VS	判别式人工智能
• 发展**成熟度低**，是近年来发展最快的形式，目前多应用于文本与图像生成场景	发展程度	• **更为成熟**，已经在互联网、零售、汽车及制造业等行业内展开应用
• 利用**大模型**和深度学习等技术可以基于已有的**海量数据**，再生成新的内容。通常生成式人工智能需要**更高性能的算力**和更多训练时间	技术角度	• 主要利用机器学习、深度学习以及计算机视觉等技术针对不同类别的数据进行判断与辨别，对于**数据需求较低，模型较小**
• 主要用于创造或者生成**新的信息或者图像** • 应用行业多为游戏开发、虚拟现实、环境模拟、内容创作、产品设计等	应用方向	• 主要用于**识别或者标签化已有的资料** • 应用行业多为自动驾驶、安防监控、人脸识别、喜好推荐、风控系统等

来源：弗若斯特沙利文

中国生成式人工智能应用市场规模

中国生成式人工智能市场受到上游技术投资持续加码和下游商业化运用持续落地的双轮驱动，展现出强劲的增长潜力和市场活力，预计到2028年应用端规模将突破3800亿人民币



投资端与应用端共同驱动

- 中国政府持续推动人工智能的发展并鼓励生成式人工智能在各行业、各领域的应用，构建出良好的生态体系。而中国生成式人工智能总投资额也在逐年升高，自2020年的约5亿元人民币增长至2023年的约135亿元人民币，期间年复合增速约为191.9%，投资额的高速增长为生成式人工智能应用市场的发展奠定了坚实的基础。
- 同期，中国生成式人工智能应用市场规模也从2020年的约98亿元人民币增长至2023年的约1,077亿元人民币，期间年复合增速约为122.3%，预计2028年，该市场将持续增长至3,804亿元人民币。

金融领域

- 增强风控评估效率，快速生成对于信贷审批候选者的研究报告；
- 提供个性化投资建议，减少人工支出成本；
- 内部运营数据分析，针对性提供优化报告。

医疗领域

- 提高患者应答率，可根据患者描述进行诉求分类或者简单咨询处理；
- 解析部分缺失数据，助力提供精准治疗服务；
- 可以模拟疾病机制，便于医疗人员进行疾病研究，加速新型医药产品研发进程。

生成式人工智能应用方向

工业领域

- 优化生产计划，可根据市场动态或历史订单数据进行弹性生产策略，提升运营效率；
- 构建定制化培训计划，缩短员工学习曲线，提升安全水平与执行效率。

社会治理领域

- 辅助政府持续优化城市规划，高效准确处理复杂且大量的资源信息，形成结构化报告，利于规划研判；
- 支持针对交通流量预测模拟，利于交通规划与优化。

来源：IDC；弗若斯特沙利文

全球大模型发展脉络梳理

全球头部大模型企业引领和推动技术的革新和落地，商业化应用逐步从文本向图像、音频和视频等领域推进，从而催生了对高性能智能算力的需求

萌芽期

- 在2018年前，美国等国家的高科技公司已开始开展人工智能以及计算机学习和算法的相关研究
  - 2013年第一个自然语言处理模型Word2Vec诞生
  - 2014年21世纪最强大的算法之一GAN诞生
  - 2017年Google提出了颠覆性的自注意力机制神经网络结构：Transformer
- 经过了将近10年的沉淀、研发、及尝试，2018年开始国外人工智能大模型开始初步实现突破
  - 2018年Google发布了其BERT大模型
  - 同年OpenAI公司发布了第一代ChatGPT大模型，并在次年推出ChatGPT-2
- 与此同时中小型初创企业开始陆续研究和发布其独立的大模型
  - 艾伦人工智能研究所于2018年发布了ELMo大模型

起步期

- 2020年下半年开始，人工智能大模型的发展逐渐增速，头部玩家的大模型逐渐成熟化
  - 2020年底美国OpenAI公司推出其ChatGPT-3大模型
  - 同年法国公司Hugging Face发布了大模型BLOOM
- 2021年日本、韩国等国家也录取推出初代的人工智能大模型
  - 2021年韩国最大的搜索公司Naver推出拥有2040亿参数的大模型HyperCLOVA
  - 同年韩国互联网巨头Kakao发布基于ChatGPT-3的KoGPT
  - 日本公司Rinna于2018年发布GPT2-medium的模型，参数达到13亿

快速发展期

- 随着ChatGPT-3的优异表现以及成功，国外人工智能大模型开始了快速发展
- 2022年
  - 前谷歌员工创立的Cohere公司推出Cohere For AI
  - 东京大学松尾研究所的AI初创公司ELYZA 推出大语言模型产品ELYZA Pencil
  - 德国初创公司Aleph Alpha发布了拥有700亿参数的预训练模型Luminous
- 2023年
  - OpenAI公司发布了ChatGPT-4并迅速迭代到了ChatGPT-4 Turbo
  - Google推出了Gemini和PaLM 2大模型
  - Meta推出Llama大模型并在同年迭代到Llama2
  - Mistral AI推出Mistral 7B
  - Anthropic推出初代Claude
  - LMSYS推出Vicuna 33B
  - Mosaic ML推出MPT-30B
  - Kakao Brain推出KoGPT 2.0
  - SKT推出A Dot

应用探索期

- 国外大模型持续发展，配套法规逐渐完善，头部玩家引领商业化落地尝试
- OpenAI与微软签订独家合作并将大模型迭代至ChatGPT-4o版本
- Meta对其Llama系列大模型进行商业化开源尝试
- Google将Gemini迭代至1.5版本，希望以其出色的代码编程能力面向软件开发群体开拓业务

2018.01 2018.07 2019.01 2019.07 2020.01 2020.07 2021.01 2021.07 2022.01 2022.07 2023.01 2023.07 2024.01 2024.07

来源：弗若斯特沙利文

全球智能算力政策解析

高性能智算资源是未来大模型产业发展的重要基石，全球主要国家相继出台扶持政策，从资金、基础建设、数据供给、人才、下游应用等多方面谋篇布局

加拿大

- 2024年投资20亿美元用于提升人工智能领域的数据计算和处理能力；
- 2024年开展了针对大众的人工智能算力设施的咨询服务，推广人工智能算力资源的知名度，使其对研究机构、企业、以及个人来说更容易获取。

法国

- 2024年由法国总统创立的人工智能协会强调了要在中短期内让法国成为世界算力资源的中心；
- 截至2022年对人工智能研究、发展、及应用的投资达到15亿欧元。

韩国

- 2024年创立针对人工智能算力芯片企业及项目的扶持基金，基金总额预计可达10亿美元；
- 预计在2025年建成其第7个超级计算机；
- 计划在2030年成为全球范围内对人工智能应用最好的国家。

日本

- 2024年落地和欧盟国家合作的高性能算力发展合作项目；
- 2024年对5家公司投资4.7亿美元用于人工智能算力设施的建设；
- 2020年允许有条件自动驾驶合法上路，并在2023年4月允许高度自动化驾驶汽车合法上路。

美国

- 2023年，行政命令第 5 节要求扩大行业间以及国际盟友的合作关系，利用高性能计算能力和人工智能建立新技术的基础模型与新应用；
- 2022年，政府鼓励以加速节点技术为支撑，克服挑战并最大限度地发挥超大规模计算和其他高性能计算的优势；
- 同年用于人工智能算法及系统研发的总统预算较2020年增长100%，在人工智能及相关机构上的政府投资金额超8亿3千万美元。

德国

- 截至2024年对人工智能相关的超级计算基础设施、技能发展、以及专业人士的培养提供5亿欧元的资金支持；
- 计划到2025年投资2亿欧元用于超级量子计算技术和设备的研发和制造；
- 2018年起新颁发不少于100个人工智能教授职位，加强领域人才的培养；
- 2017年起同其它6个欧洲国家签署成立了欧洲超级计算组织。

澳大利亚

- 2023年起4年内投资5380万美金用于国家人工智能中心以及人工智能计算处理能力中心的建设；
- 2019年起通过为全体居民提供高质量的人工智能医疗保障服务来降低医疗方面的总体支出。

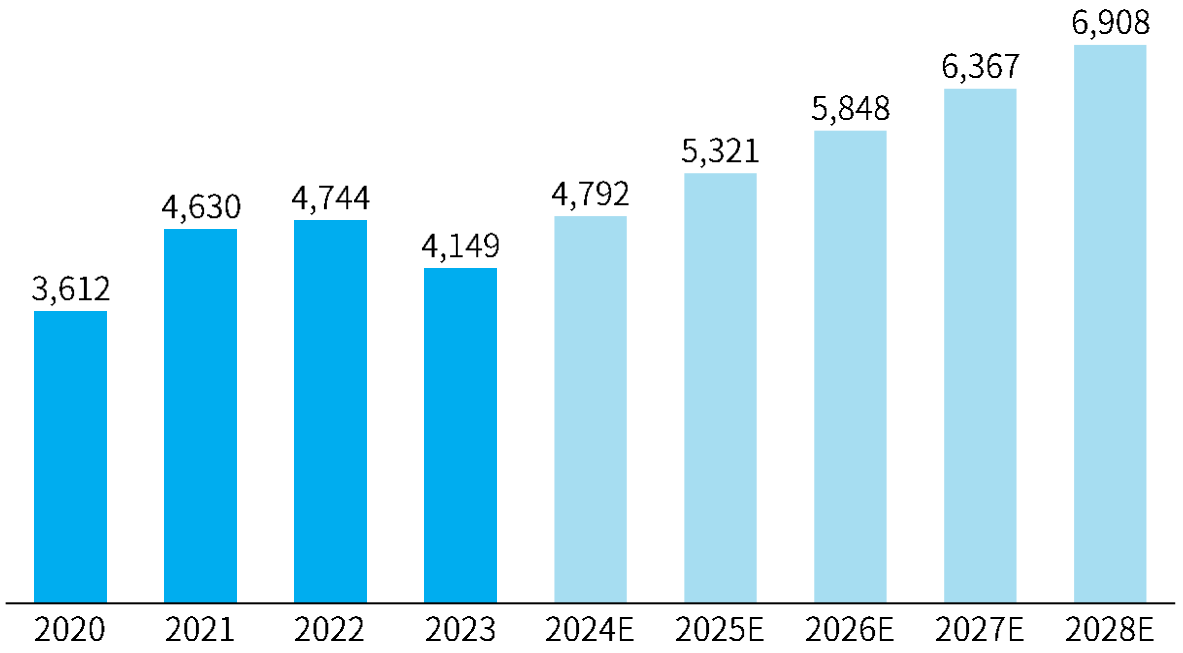
来源：弗若斯特沙利文

全球芯片市场规模

芯片是算力的基石，全球芯片市场预计2024年将逐步复苏，规模同比增长16%，并且突破4700亿美元。同时，生成式AI的浪潮席卷全球，对智算资源的需求跃升至新阶段，带动了人工智能芯片的规模化扩张

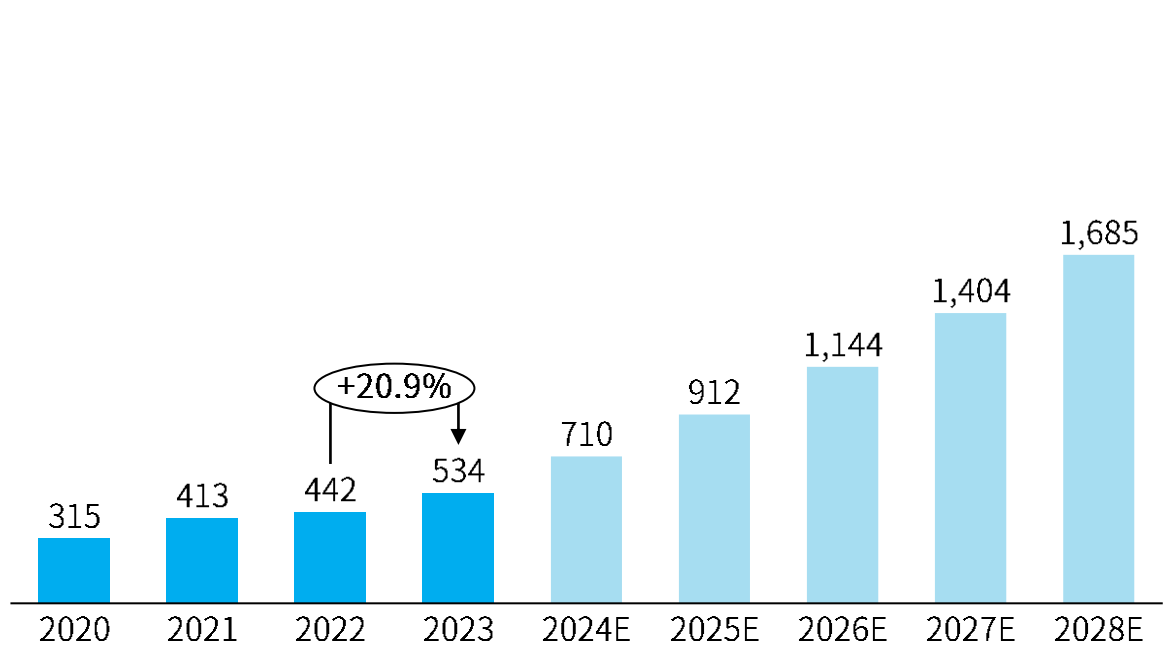
全球芯片市场规模

2020-2028预测，亿美元



全球人工智能芯片市场规模

2020-2028预测，亿美元



全球芯片市场稳步增长，AI芯片引领市场变革

- 受全球经济和需求放缓的影响，2023年全球芯片市场规模小幅下降至4,149亿美元。未来随着芯片厂商的产能提升，AI研发和应用的逐步扩展，全球芯片市场预计将稳步提升，在2028年将达6,908亿美元，2023年至2028年期间年复合增长率达10.7%。
- 随着大规模部署定制AI芯片的趋势愈发明显，传统的芯片架构正逐渐被替代，以更好地满足各类AI工作负载的需求。2023年全球人工智能芯片市场规模达534亿美元，预计到2028年将突破1,600亿美元，其增速显著高于全球芯片市场。

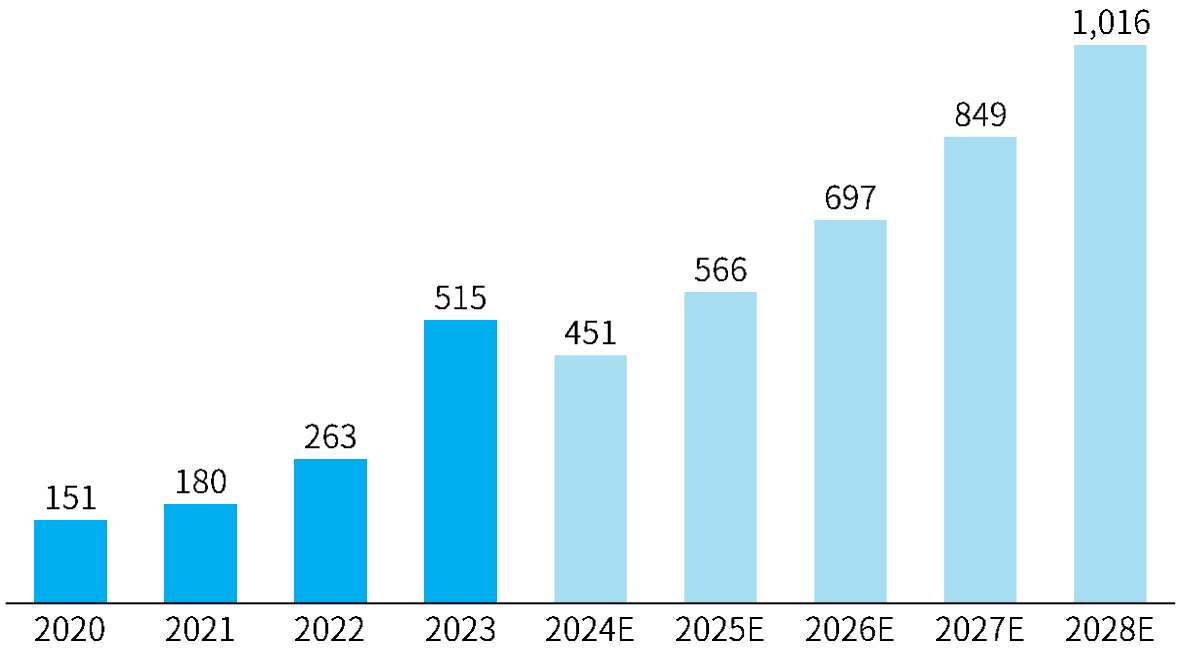
来源：弗若斯特沙利文

全球人工智能服务器市场规模

人工智能服务器是智能算力的载体，2023年全球人工智能服务器市场规模首次突破500亿美元，增幅高达96%，侧面反应了全球人工智能企业加速了对后续AI及智算业务的资源储备

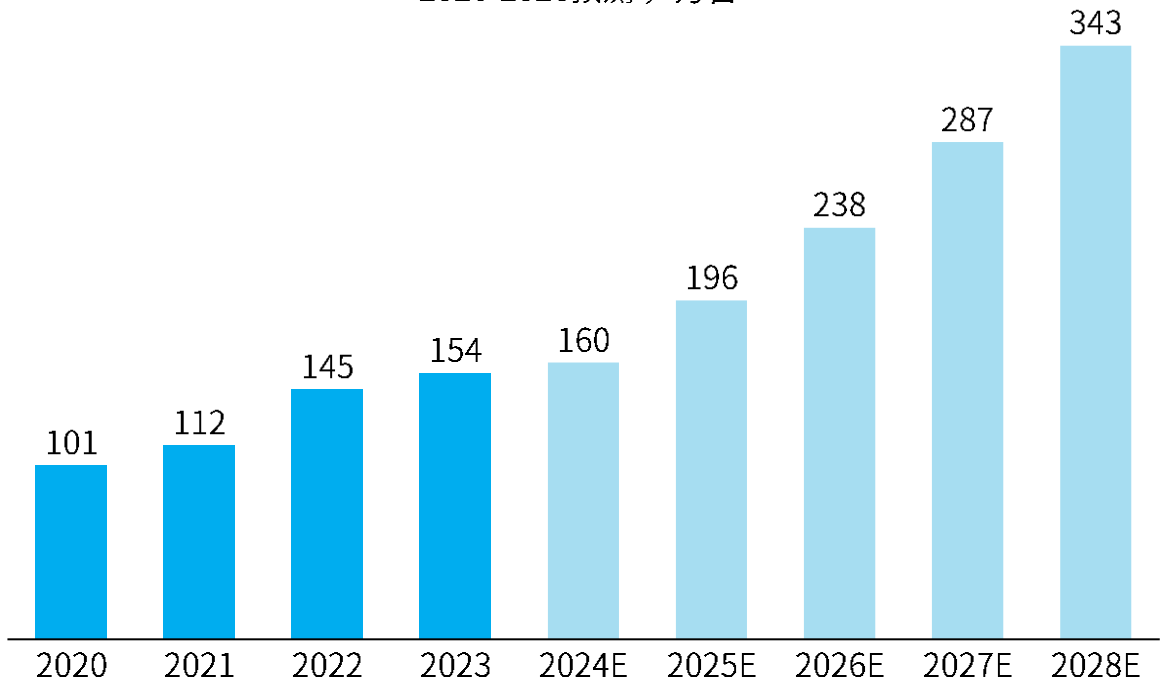
全球人工智能服务器市场规模

2020-2028预测，亿美元



全球人工智能服务器出货量

2020-2028预测，万台



AIGC浪潮推动AI服务器需求大幅增长，全球市场规模与出货量齐涨

- 在全球AIGC的驱动下，基于GPU、ASIC等加速芯片的AI服务器需求上升，2023年全球人工智能服务器市场规模达515亿美元，同比增长96.3%，预计在2028年将突破1000亿美元，五年年复合增长率达14.5%。
- 作为数据中心的核心硬件设备和算力的重要载体，全球AI服务器的出货量保持高速增长。2023年，全球AI服务器的出货量达154万台，占整体服务器出货量的12.6%，预计到2028年全球AI服务器出货量将达到343万台，五年年复合增长率达17.4%。

来源：弗若斯特沙利文

中国大模型发展历程拆解

中国大模型产业虽然起步晚于西方企业，但在从2023年开始进入发展爆发期，国内各大高科技厂商、高校、科研机构及创业团队相继推出自研大模型，大模型市场进入“百团大战”的火热场面

萌芽期

- 中国大模型起步较全球相比稍晚，在萌芽期的企业多在摸索阶段，仅有个别企业/机构推出大模型产品进行试水
  - 华为2021年推出其盘古系列大模型
  - 智源研究院2021年推出其悟道系列大模型产品

盘古大模型　悟道 WUDAO

追赶期

- 随着全球大模型时长的快速发展，国内外人工智能大模型已出现一定的差距，国内多方领域的头部玩家开始积极布局大模型赛道

互联网企业

- 百度于2023年推出文心一言大模型并且在同年迭代至3.0版本
- 阿里巴巴在2022年推出通义千问（Qwen）模型，次年迭代至2.0版本并推出通义千问Audio版大模型
- 360公司2023年推出其智脑大模型并在同年迭代至4.0版本
- 京东2023年推出言犀大模型
- 网易2023年推出教育领域大模型子曰
- 腾讯2023年推出混元大模型

文心一言　通义千问　言犀　子曰大模型

其他企业

- 科大讯飞2023年推出星火大模型系列，并在同年将其迭代至3.0版本
- 百川智能2023年推出Baichuan系列产品，并在一年内迭代至Baichuan2-Turbo版本并进一步扩大训练参数数量至530亿
- 月之暗面科技公司2023年推出KimiChat大模型
- 小米2023年推出MiLM大模型
- 理想2023年推出MindGPT大模型
- 商汤科技2023年推出SenseChat系列大模型并迭代至3.0版本

百川智能 BAICHUAN AI　Mind GPT　商量 SenseChat

高校及研究机构

- 清华大学发布开源大模型ChatGLM
- 复旦大学发布MOSS大模型
- 北京大学发布ChatExcel大模型
- 中科院自动化研究所推出紫东太初大模型并迭代至2.0版本
- 上海AI实验室推出书生系列大模型并在同年推出3个衍生系列大模型产品

ChatGLM　MOSS　ChatExcel　书生

发展期

- 中国的大模型市场逐渐进入相对稳定迭代发展阶段，主流大模型框架已初步获得市场的认可，初创企业积极入局，大模型生态构建逐步形成
  - 阿里巴巴发布70亿参数通义千问Qwen2-72B
  - 月之暗面科技公司突破200万字长文本迭代限制并推出API工具接口
  - 百度文心一言迭代至4.0 Turbo版本
  - 商汤科技SenseChat迭代至5.0版本
  - 上海AI实验室“书生”系列产品逐渐根据不同的下游应用开发出相应的衍生产品并普遍迭代至2.0版本
  - 科大讯飞的星火大模型迭代至3.5版本

文心一言　通义千问　“商量商量，都能解决”商汤自研中文语言大模型应用平台　书生　讯飞星火

2021.01　2022.01　2022.07　2023.01　2023.07　2024.01　2024.07

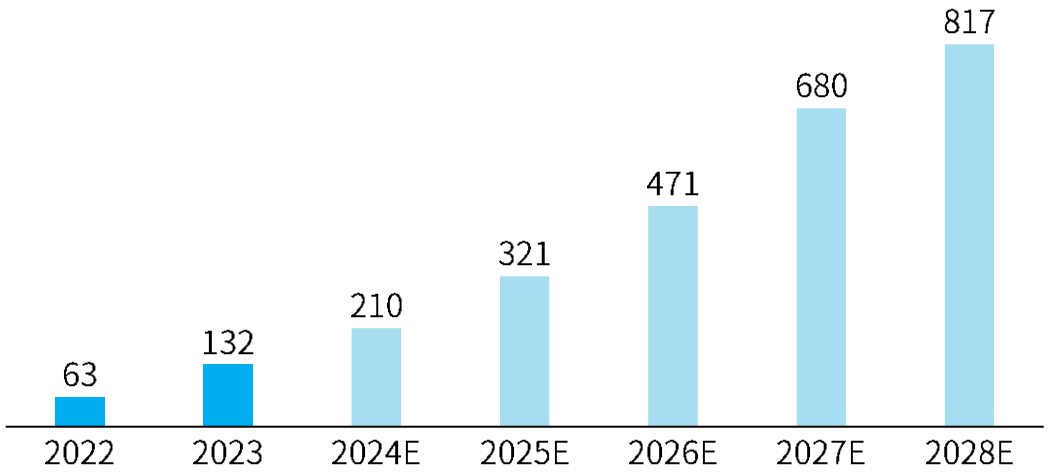
来源：弗若斯特沙利文

中国大模型市场规模

中国大模型市场正在经历快速的技术迭代和商业化落地，目前头部企业的大模型水平已经追平国际大模型均线，预计在2028年市场规模将突破800亿人民币，未来可期

中国大模型市场规模

2020-2028预测，亿人民币



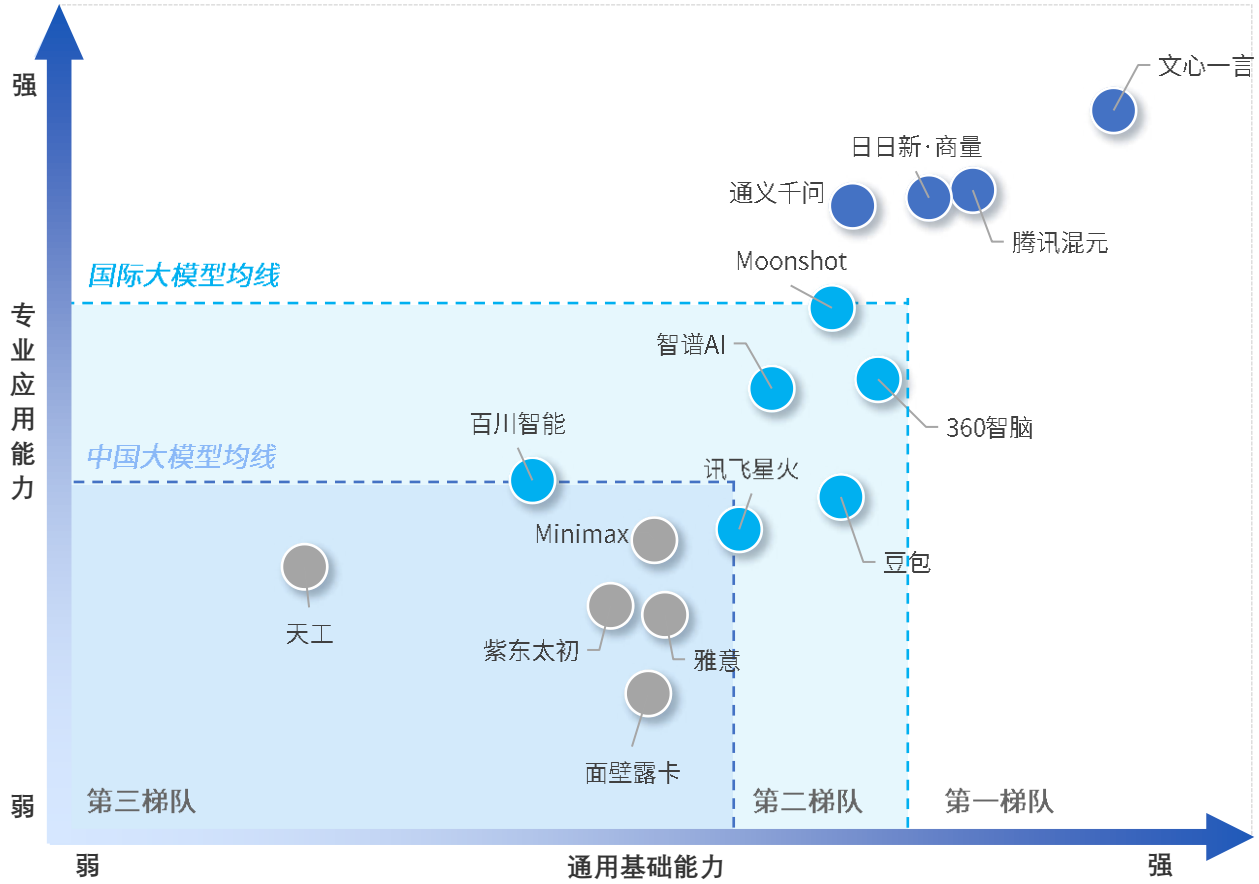
中国大模型发展时长较短、但增长潜力巨大

- 中国大模型市场规模在历经了短期的加速发展后，在2023年达到了约132亿元人民币，预计在未来随着大模型技术不断提升，也会逐步向轻量化小模型，垂直化，以及多功能化方向发展，在2028年将达到817亿元人民币。
- 《2024年中国大模型能力评测》中根据综合考量数理科学、语言能力、道德责任、行业能力及综合能力等5大核心维度的表现来看，中国大模型均线在通用基础能力与专业应用能力上大多弱于国际领先大模型。而根据国际及中国大模型均线可将目前国内主流大模型分为三个梯队。第一梯队为已经达到国际均线的百度文心一言、腾讯混元、商汤日日新·商量和阿里巴巴通义千问；第二梯队包含Moonshot·Kimi、360智脑、智谱AI、豆包、讯飞星火、百川智能。

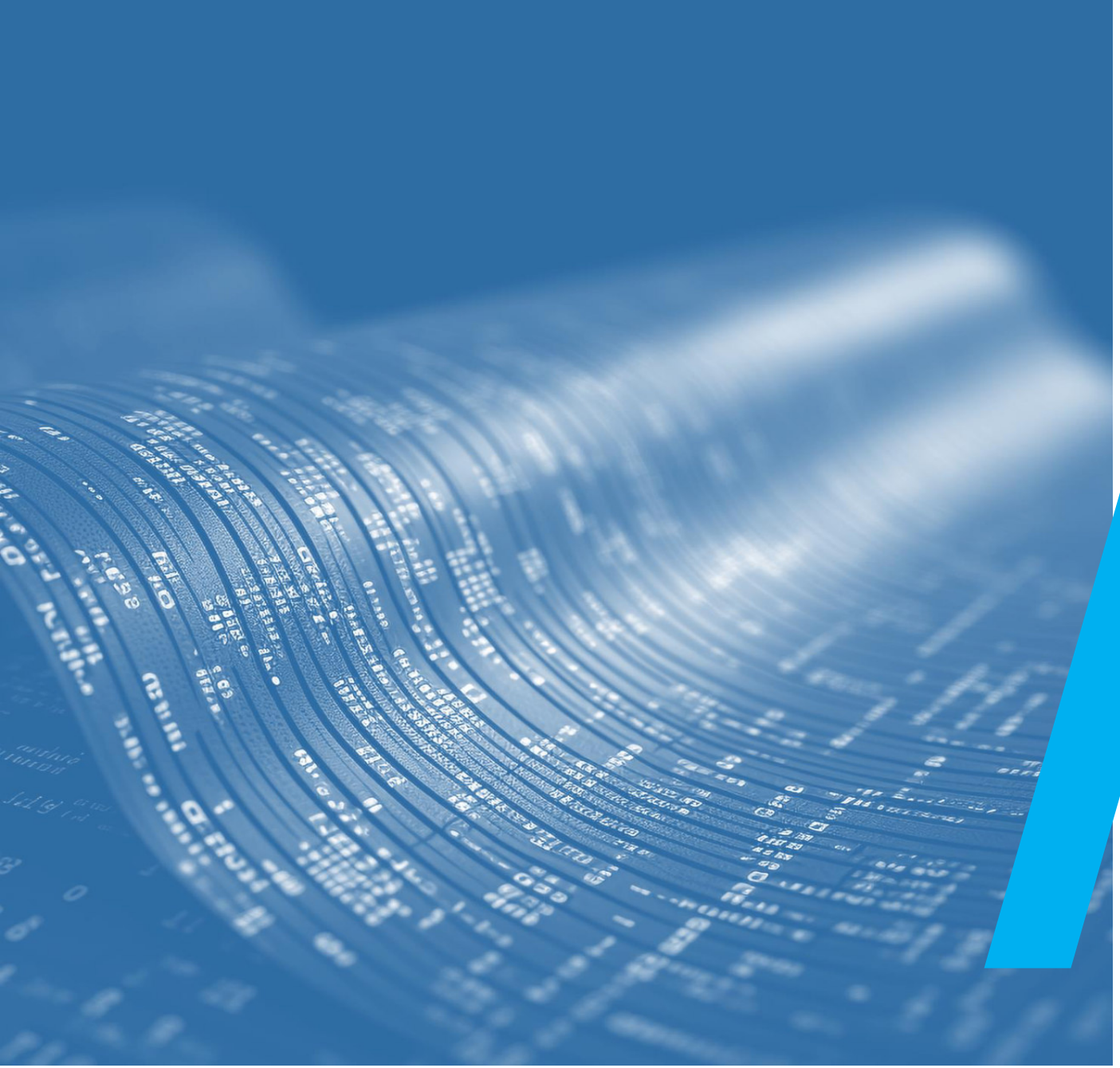
中国大模型综合竞争力气泡图

中国大模型均线= *15个参评大模型均分*

国际大模型均线 *Gemini 1.0* + *GPT4* + *GPT3.5* + *Claude 2*



来源：头豹研究院《2024年中国大模型能力评测》；弗若斯特沙利文



2

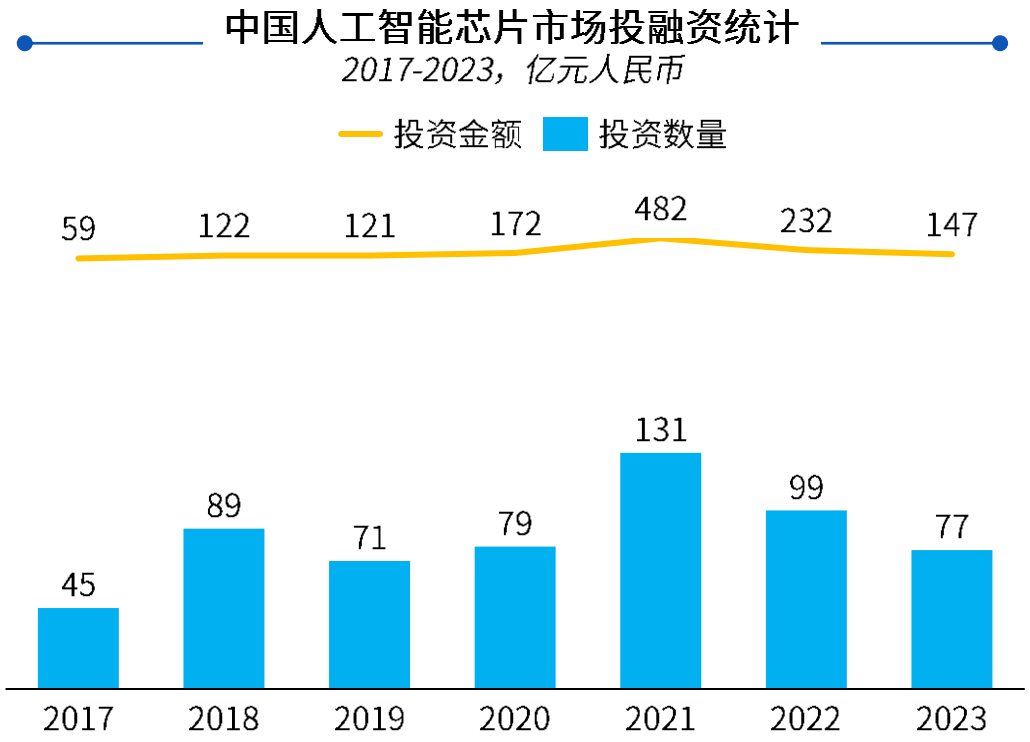
- 产业升级
中国智能算力崛起

智能算力定义与分类

智能算力以人工智能芯片为载体，为人工智能应用提供算法模型训练和推理所需的高性能计算能力

- **人工智能算力**（简称“智能算力”或“智算”）是指用于处理人工智能训练与推理中的大量数据、模型以及其他计算任务的能力。目前，智能算力的提升已成为推动人工智能应用创新和产业升级的关键因素，也是实现大模型快速迭代和优化的关键。
- 智能算力源于专门用于处理人工智能应用中大量数据计算任务的模块，又常被称为人工智能芯片。人工智能芯片目前按照发展路径可以分为两种，一种支持传统计算架构，此种架构中CPU（Central Processing Unit，中央处理器）作为核心逻辑处理器，统一进行任务调度，人工智能芯片将会作为CPU的运算协处理器。另一种则是非冯诺依曼计算架构，此种架构是利用电子技术构建出神经拟态芯片。
- 2021年我国人工智能芯片行业融资达到顶峰，发生融资事件共131起，融资金额共482亿元。2021年后受宏观环境影响，我国人工智能芯片行业融资事件数开始下降，但中长期来看，随着扶持政策落地和人工智能产业逐渐成熟，未来人工智能芯片行业将持续高速发展态势，并拉动智能算力的增长。

人工智能芯片	架构类型		应用类型
	传统计算机架构		训练算力
	GPU	*一种专门用于图像处理的微处理器，其采用数据并行计算模式完成计算任务.常用于数据密集的科学与工程计算中。*	*对芯片需求算力要求高，多为FP32与FP16精度，功耗小、可编程性、高内存与高带宽*
	FPGA	*具有无限次编程的特点，具有模块化，规则化的架构，支持现场重新编辑的功能。整体开发时间短、延迟低、能耗低。*	
	ASIC	*应特定用户要求和特定电子系统的需要而设计、制造的集成电路，在特定算法下能效更加高。可以分为全定制及半定制两种，整体开发时间短、能耗低。*	推理算力
	类脑计算架构		*对芯片需求算力要求低，多为FP32与FP64精度，低延时、分布式、可拓展性*
	NPU	*与服务传统计算机架构深度神经网络计算不同，类脑芯片是模仿大脑神经结构的脉冲神经网络式计算的芯片。目前多为研究型芯片，商业化程度低。*	



来源：弗若斯特沙利文

中国智能算力政策梳理和解析（1/2）

智算产业的发展得到了国家层面的高度重视和政策支持，近年来，国家出台一系列政策文件，旨在推动智算产业的快速发展，智算产业被提升至国家战略高度

■ 国家和地方相继出台智算相关政策，积极发展算力产业生态，强化算力供需对接，推动算力产业区域协同发展。

政策名称	颁布主体	政策要点
深入实施“东数西算”工程加快构建全国一体化算力网的实施意见	国家发改委等五部门	• 到2025年底，通用算力、智能算力、超级算力等多元算力加速集聚，国家枢纽节点地区各类新增算力占全国新增算力的60%以上，国家枢纽节点算力资源使用率显著超过全国平均水平
算力基础设施高质量发展行动计划	工信部等六部门	• 到2025年算力规模超过300EFLOPS，智能算力占比达到35% • 推动算力结构多元配置，逐步提升智能算力占比，推动智能算力与通用算力协同，满足不同类型算力业务需求
数字中国建设整体布局规划	国务院	• 系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局
国家发展改革委等部门关于同意京津冀地区启动建设全国一体化算力网络国家枢纽节点的复函	国家发改委等四部门	• 同意在京津冀地区启动建设全国一体化算力网络国家枢纽节点，发展高密度、高能效、低碳数据中心集群，通过云网协同、云边协同等优化数据中心供给结构，扩展算力增长空间，实现大规模算力部署
“十四五”国家信息化规划	中央网络安全和信息化委员会	• 统筹建设面向区块链和人工智能等的算力和算法中心，构建具备周边环境感应能力和反馈回应能力的边缘计算节点，提供低时延、高可靠、强安全边缘计算服务
“十四五”大数据产业发展规划	工信部	• 加快构建全国一体化大数据中心体系，推进国家工业互联网大数据中心建设，强化算力统筹智能调度，建设若干国家枢纽节点和大数据中心集群。建设高性能计算集群，合理部署超级计算中心

智能算力资源科学布局

✓统筹通用算力、智能算力、超级算力的一体化布局

✓建立东西部联动机制，降低东西部数据传输成本

智算设施建设自主化

✓加强智算基础设施软硬件产品的技术研发，提升核心软硬件的自主可控水平

智能算力的创新及应用

✓结合人工智能发展和业务需求，探索创新融合泛在的智算应用场景

智算中心坚持绿色低碳发展

✓提升智算中心的设施能源利用效率和算力碳效水平

来源：弗若斯特沙利文

中国智能算力政策梳理和解析（2/2）

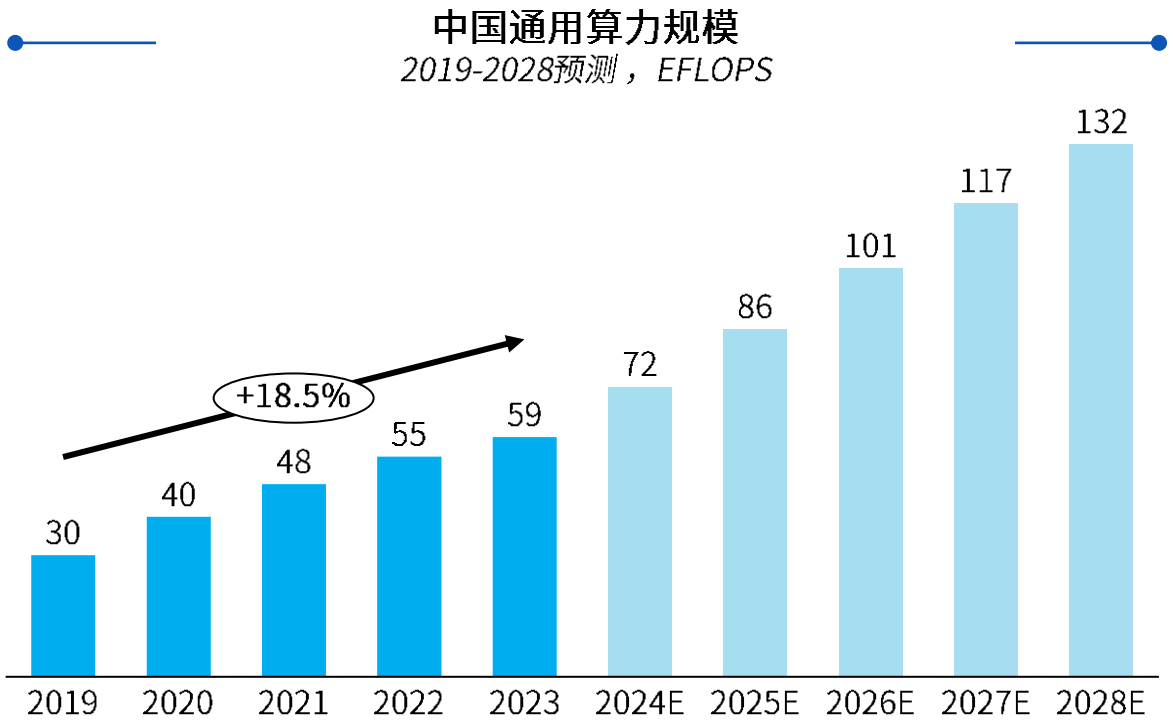
各级政府响应国家号召，政策引导和鼓励覆盖包括基础设施建设、区域协同发展、人才培养与技术创新等多个维度

省份/城市	政策名称	政策要点
陕西	陕西省加快推动人工智能产业发展实施方案（2024-2026 年）	• 整合省内算力资源，建设省级算力统筹调度平台，实现“算力一网化、统筹一体化、调度一站式”，2026年建设运营智能算力达到3000P以上，可统筹的公共智能算力达到西部领先水平
北京	北京市算力基础设施建设实施方案（2024—2027年）	• 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 • 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
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	• 到2025年，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%，PUE值将降至1.25以下 • 到2025年，上海市智能算力规模将超过30EFLOPS，占比达到总算力的50%以上，智算中心内先进存储容量占比达到50%以上
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）	• 到2025年，在计算力方面，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
深圳	深圳市算力基础设施高质量发展行动计划（2024-2025）	• 到2025年，通用算力达到14EFLOPS（FP32），智能算力达到25EFLOPS（FP16），超算算力达到2EFLOPS（FP64），存储总量达到90EB。先进存储容量占比达到30%以上，重点行业核心数据、重要数据灾备覆盖率达到100%
山东	山东省数字基础设施建设行动方案（2024-2025 年）	• 到2025年，全省数据中心在用标准机架总数达到45万个，总算力达到 12.5EFLOPS，智能算力占比达到35%，存力规模达到65EB，先进存储占比达到35%以上
河南	河南省重大新型基础设施建设提速行动方案（2023—2025年）	• 到2025年智算和超算算力规模超过2000P FLOPS，高性能算力占比超过30% • 持续提升国家超算郑州中心超算能力，建设智算中心和郑州城市算力网调度中心，资源利用率达到70%
上海	临港新片区加快构建算力产业生态行动方案（2023-2025年）	• 到 2025 年，新片区算力供给形成以智算算力为主、基础算力和超算算力协同的多元算力供给体系，总算力超过 5EFLOPS（FP32），AI 算力占比达到 80%，新建数据中心 PUE 控制在 1.25 以内
贵州	面向全国的算力保障基地建设规划	• 围绕高可靠、高可用目标，从备份中心提升为计算中心、效益中心，重点布局智算基础设施，形成低时延人工智能算力基地、全国低成本中心、高安全中心，到2024年通用算力、智算算力、超算算力的总规模达到5EFLOPS

来源：弗若斯特沙利文

中国通用算力规模

2023年中国通用算力规模达到59EFLOPS，2019至2023年的年复合增长率达到约19%，通用算力因其复杂任务处理能力以及高适应性使其有广泛的应用场景，但难以适应大模型训练的需求



- 中国通用算力规模从2019年的30EFLOPS增长至2023年的59EFLOPS，年复增长率18.5%。随着需求量的上升以及政策的持续推动预计未来将保持17.3%的增长率至2028年的132EFLOPS；
- 5G及物联网的大规模部署使中国的数据总量进一步提高，在用数据中心机架总数近5年年均增速超过30%，在此推动下中国市场对通用算力需求依旧稳步增加；
- 《算力基础设施高质量发展行动计划》指出，要推动不同计算架构的智能算力与通用算力协同发展，满足均衡型、计算和存储密集型等各类业务算力需求。

优势

- 基于CPU芯片的服务器所提供的计算能力，兼顾计算能力以及处理复杂计算任务能力；
- 其主要优势在于适合用于处理从串行计算到数据库运行等类型的工作，并可以处理复杂的数学算术、逻辑计算等;
- 通用算力能适应不同的应用场景下的复杂的任务需求。

劣势

- 基于CPU的通用计算适应各类复杂运算，但难以支撑人工智能应用中所需要的超大规模且简单的并行数据计算；
- 随着语音识别、视觉识别等人工智能技术的兴起以及机器学习、深度学习等模型的广泛应用，通用计算计算性能一般，难以满足激增的算力需求。

主要头部玩家

- 通用算力的主要应用场景为互联网（39%）、电信（13.16%）、政府（12.26%）；
- 通用算力市场资金壁垒、研发壁垒、及项目资源壁垒较高，因此市场主要头部玩家市占率较高且初创企业进入难度较高；主要的供应厂商可根据具体业务功能分为三大类：

云服务厂商

阿里巴巴 Alibaba.com

Tencent 腾讯

电信运营商

中国电信 CHINA TELECOM

中国移动 China Mobile

数据中心

GDS万国数据

数据港

光环新网 SINNET

CHINDATA 秦淮数据

世纪互联 VNET

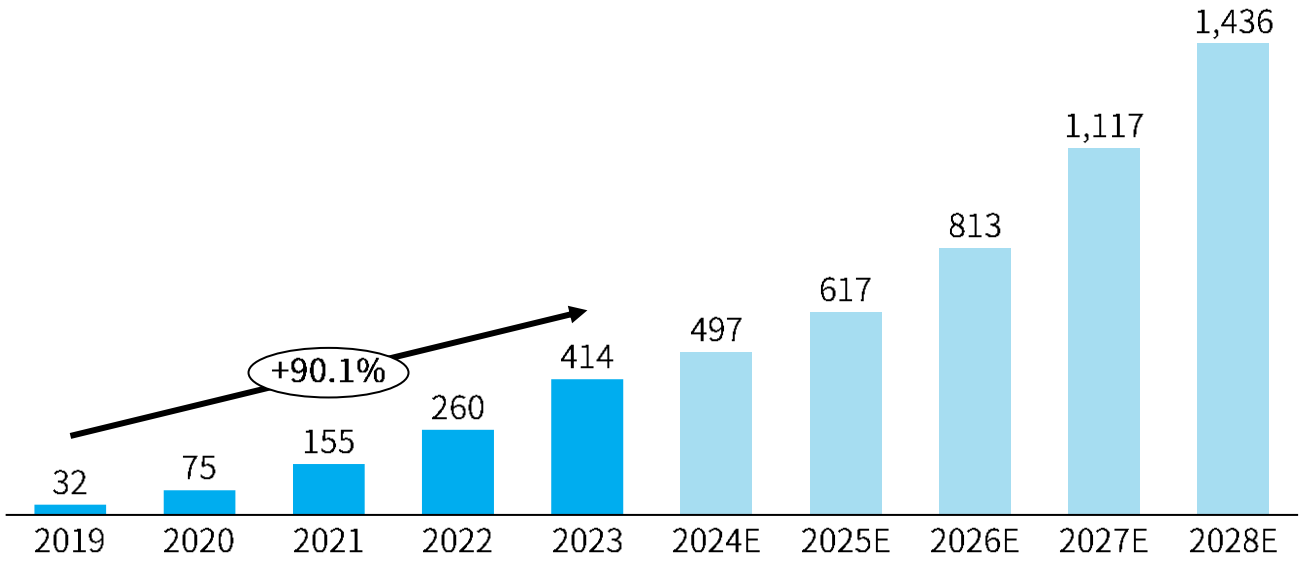
来源：弗若斯特沙利文

中国智能算力规模

生成式AI技术的爆发式增长催生对高性能算力的需求，中国智能算力规模在2023年达到约414EFLOPS，主要用于互联网领域大模型的训练和推理。未来随着大模型的行业分化，通用算力、智能算力和超算算力越来越呈现出融合趋势

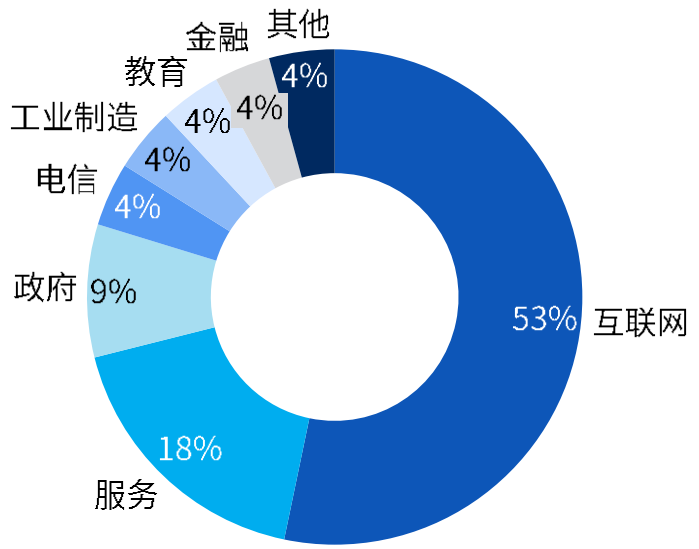
中国智能算力规模

2019-2028预测，EFLOPS



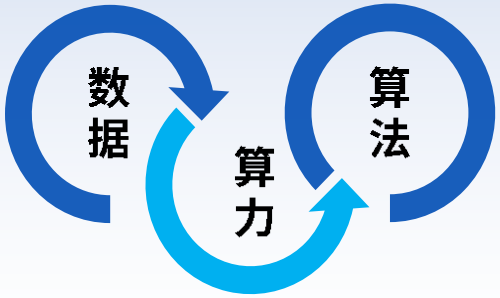
中国智能算力应用分布图

2023，%



人工智能发展三要素

- 大量的场景**数据**为AI训练构筑**基础**
- 高性能计算机提供**数据处理计算能力**
- 算法是基于数据分析而构成的**基础规律**并进行**预测**



互联网领跑智能算力应用，下游行业渗透率加速提升

- 自2019年至2023年，中国智能算力规模从32EFLOPS增长至414EFLOPS，期间年复合增长率约为90.1%，已然超越了同期中国通用算力规模的增速，整体增长态势十分强劲。
- 针对于智能算力下游应用行业的分类来看，目前智能算力主要应用的行业是互联网，其占比超过五成，这是由于互联网企业相对其他行业而言对于大模型的投入较多，故而对于计算需求较大。其他行业目前还在探索人工智能的应用阶段，整体对于智能算力的需求还处于培养期。未来随着人工智能应用加速落地，其他行业对于智能算力的需求将会大幅提升。展望未来，预计到2028年，中国的智能算力规模将会达到1,436EFLOPS。

来源：中国信息通信研究院；弗若斯特沙利文

中国智能算力产业图谱总览（1/2）

在智能算力产业的上游环节，我国目前虽然在基础设施建设和部分核心设备方面实现了国产化替代，但是在人工智能芯片环节，市场份额依旧掌握在海外厂商手里，对我国智能算力供给的稳健性造成隐患

上游 – 智能算力基础设施

IT设施

智算服务器

DELL Technologies、安擎 NGINETECH、H3C、HUAWEI、inspur 浪潮、KUNQIAN 坤前、Lenovo、Nettrix 宁畅、宝德 PowerLeader、SITONHOLY、Suma、超云 SUPER CLOUD、xFUSION、ZTE 中兴

AI芯片

AMD、壁仞科技 BIREN TECHNOLOGY、Cambricon 寒武纪、Enflame 燧原科技、HISILICON、天数智芯 Iluvatar CoreX、intel、昆仑芯、NVIDIA

存储设备

H3C、Hewlett Packard Enterprise、HIKVISION 海康威视、HITACHI Inspire the Next、HUAWEI、IBM、inspur 浪潮、宏杉科技 macrosan、SAMSUNG、SEAGATE、中科曙光 Sugon、同有 TOYOU、Western Digital

交换机设备

ARISTA、Asterfusion、cisco、H3C、DELTA 台达、Hewlett Packard Enterprise、HUAWEI、JUNIPER NETWORKS、Ruijie 锐捷 Networks、ZTE 中兴

光模块

Accelink、eoptolink、FINISAR、HGTECH、LUMENTUM、中际旭创 ZHONGJI INNOLIGHT

基础设施

供电系统

DELTA 台达、EAST 易事特、HUAWEI、KSTAR 科士达、Schneider Electric、VERTIV

后备电池

DELTA 台达、EAST 易事特、HUAWEI、Narada 南都电源、SUNWODA 欣旺达、VERTIV

制冷系统

Envicool 英维克、HUAWEI、GRC、IBM、intel、Schneider Electric、STÄUBLI、Tencent 腾讯

弱电系统

ABB、阿里巴巴 Alibaba.com、Baidu 百度、HUAWEI、Schneider Electric、SIEMENS、VERTIV

来源：弗若斯特沙利文

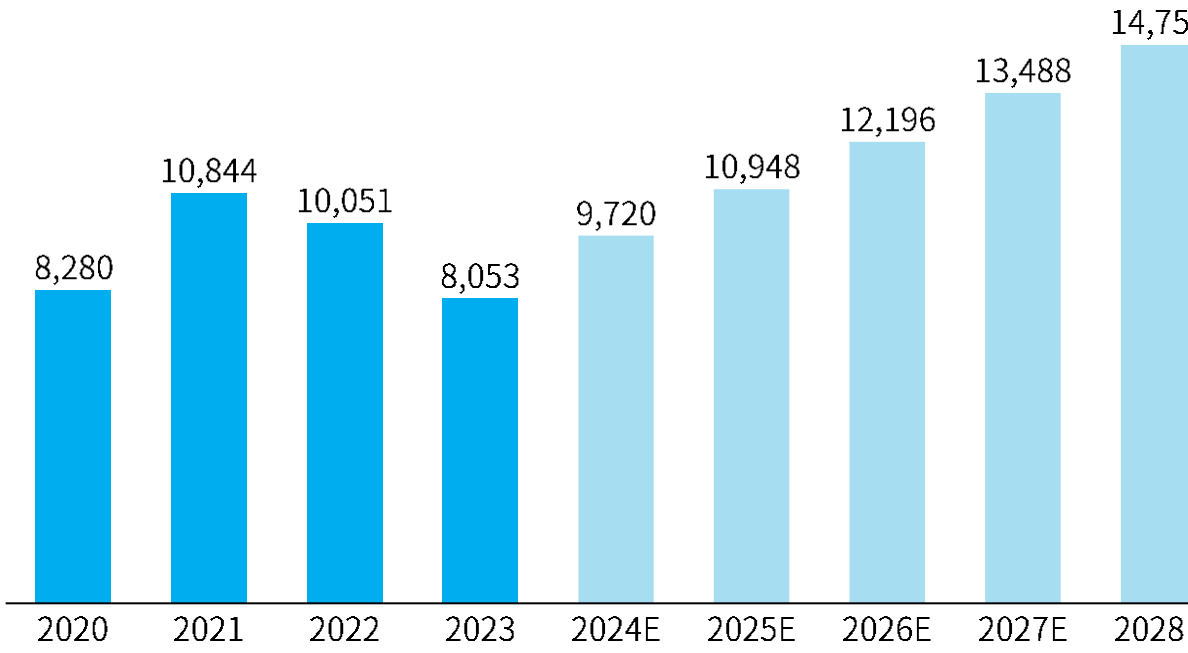
注：此版为第一版产业图谱，未来将根据市场变化持续更新。

中国智能算力产业链透视-上游基础设施层

芯片是算力产业的基础硬件，而人工智能芯片更是为智能算法和应用提供必要的计算能力，其设计和制造水平直接影响中下游智算资源的供给和使用。因此，拥有自主可控的人工智能芯片技术是推动智算产业长期发展的坚实基础

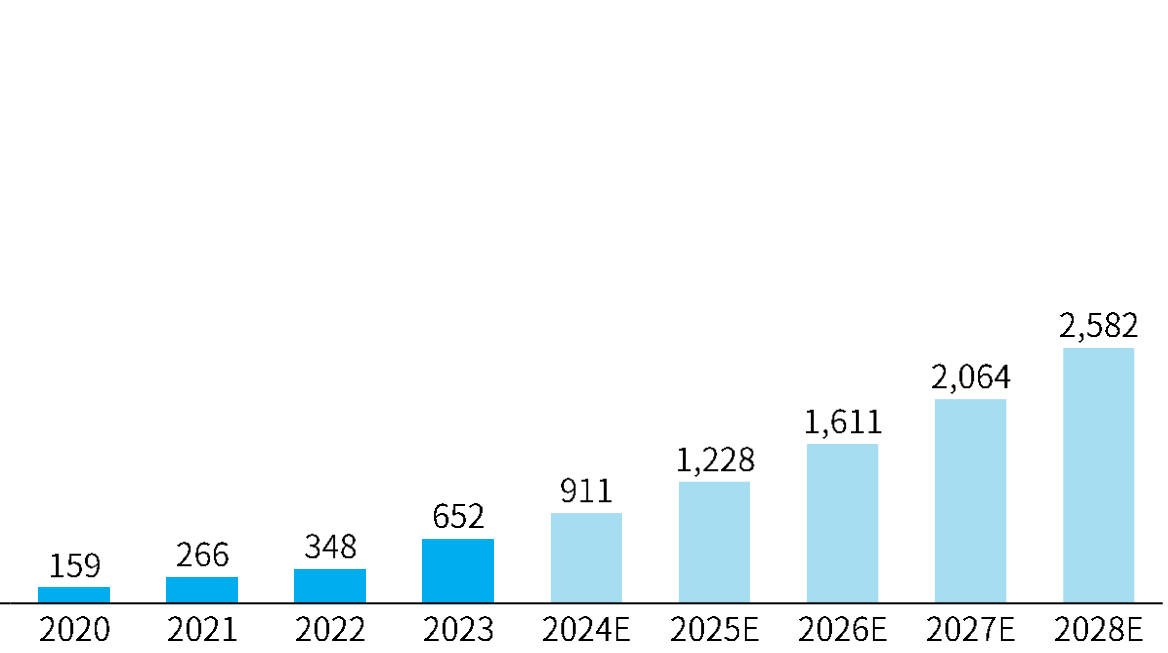
中国芯片市场规模

2020-2028预测，亿人民币



中国人工智能芯片市场规模

2020-2028预测，亿人民币



场景应用落地协同利好政策大力中国驱动人工智能芯片发展

- 由于美国对于中国半导体监管趋严，叠加全球经济放缓，中国芯片市场规模自2021年呈现下滑趋势，但随着芯片产业链结构不断优化，例如中国芯片设计保持高速增长，封装占比逐渐下降，未来中国芯片市场将恢复增长态势预计至2028年增长到1.5万亿人民币。
- 人工智能芯片市场作为中国芯片市场的一个重要组成部分，近年来得益于国家对该产业的高度重视，其展现出了高速增长的趋势。2023年中国AI芯片市场规模约为652亿人民币。随着算力中心的增加以及终端应用的逐步落地，中国AI芯片需求也持续上涨。此外类脑等新型AI芯片也在探索量产，因此未来市场增长潜力巨大，预计市场规模将于2028年达到2,582亿人民币。

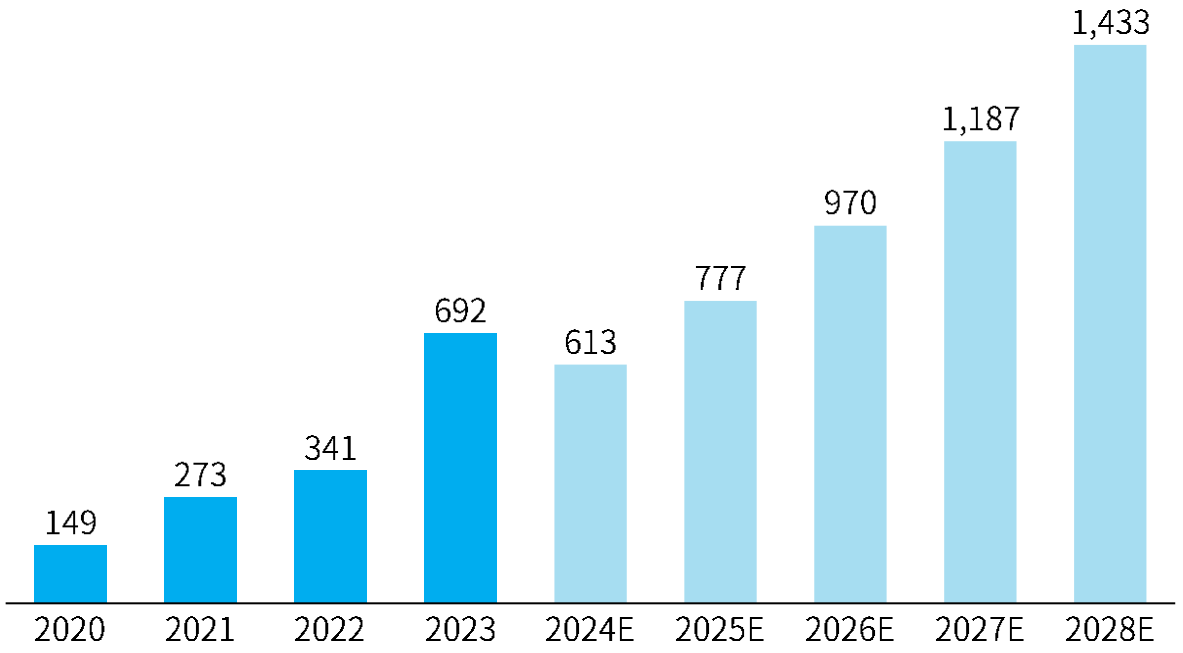
来源：弗若斯特沙利文

中国智能算力产业链透视-上游基础设施层

搭载AI芯片的人工智能服务器为智算产业提供了必要的算力支撑，使得大规模数据处理、模型训练和推理计算等复杂任务得以高效完成，2023年我国人工智能服务器市场规模达到约692亿人民币，其中GPU加速服务器占比超过90%

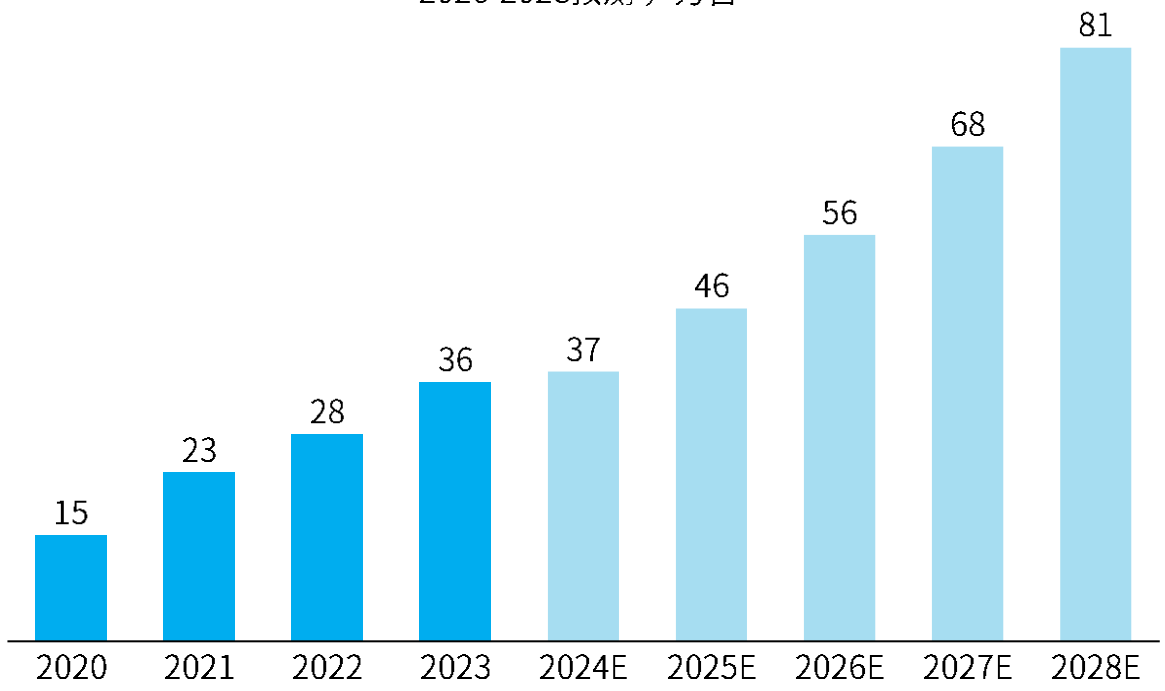
中国人工智能服务器市场规模

2020-2028预测，亿人民币



中国人工智能服务器出货量

2020-2028预测，万台



智能时代的引擎，中国AI服务器稳定增长

- 中国人工智能服务器市场在人工智能行业及智算行业高速发展的带动下持续走高，自2020年的近149亿人民币增长至2023年的692亿人民币。从服务器出货台数来看，整体出货量也是保持了同步的增长，由15万台增长至36万台。 其中，2023年浪潮、坤前与新华三三家提供了超过50%的人工智能服务器。
- 展望未来，随着人工智能应用的成熟度提升，市场对于人工智能服务器的需求将持续走高。但鉴于国内人工智能芯片与服务器的生产技术及产业链尚不稳定与健全，以及美国对中国制裁的不确定性，导致2023年许多企业进行了大量的采购，预计到2024年，市场需求会逐渐回落至正常水平并恢复稳定增长态势。

来源：弗若斯特沙利文

中国智能算力产业图谱总览（2/2）

不同类型的中游企业凭借自身在供应链、渠道、产品、平台和技术等方面的差异化优势，整合上游智能算力的基础设施和核心硬件设备，为下游智能算力需求方交付高性能智算资源

中游 – 智能算力资源供给方

电信运营商

中国移动 China Mobile　中国联通 China unicom　中国电信 CHINA TELECOM

第三方数据中心服务商

BESTER　CDS 首云　中金数据 CENTRIN DATA　CHINDATA GROUP

GDS万国数据　数据港　奥飞数据 www.ofidc.com

润泽科技发展有限公司 Range Technology Development Co., Ltd.　润建股份 RIGF

AI Infra厂商

天罡智算　AutoDL　北京超级云计算中心 BEIJING SUPER CLOUD COMPUTING CENTER

九章云极 DataCanvas　潞晨科技 HPC AI TECH　厚德云

INFINIGENCE 无问芯穹　蓝耘 LANYUN　清昴智能 QINGMAO INTELLIGENCE

ROTH　SILICONFLOW　SILINEST 普云硅巢

天羽蜂智能算网平台　天云融创软件 纳百云·算未来　TRANSWARP 星环科技

趋动云 VirtAI Cloud　猿界算力　云脑科技

跨界企业

安诺其集团　CHT·世纪华通　创业黑马 DARK HORSE　entive 亿田

福能东方 FOET　协鑫能科 GCLET　Greattown 大名城 股票代码600094 900940　航锦科技

杭州钢铁集团公司 HANGZHOU IRON&STEEL GROUP COMPANY　鸿博股份有限公司 HONGBO CO.,LTD.　honflex 弘信电子

IAT　江阴市恒润重工股份有限公司 股票代码:603985 JIANGYIN HENGRUN HEAVY INDUSTRIES CO., LTD　锦鸡股份

利通电子 LETTALL Lettall Electronic　莲花控股　朗源 LONTRUE　南兴股份

宁波建工 NINGBO CONSTRUCTION　TIMEVERSE 天娱数科　旋极科技 WATERTEK　威星智能 VIEWSHINE

人工智能企业

百川智能 BAICHUAN AI　4Paradigm 第四范式　科大讯飞 IFLYTEK　零一万物

MEGVII 旷视　MINIMAX　商汤 sensetime　智谱·AI

云服务厂商

阿里云　百度云　火山引擎　PARATERA 并行　QINGCLOUD 青云　天翼云 State Cloud　腾讯云　UCLOUD 优刻得

ICT硬件集成商

DELL Technologies　安擎 NGINETECH　H3C　HUAWEI　inspur 浪潮　Lenovo　KUNQIAN 坤前　Nettrix 宁畅

宝德 PowerLeader　思腾合力 SITONHOLY　中科曙光 Sugon　Suma　超云 SUPER CLOUD　xFUSION　新紫光集团 TSINGHUA UNIGROUP　ZTE 中兴

来源：弗若斯特沙利文

注：AI Infra厂商是指在人工智能生态系统中，链接算力和应用的中间层，其可覆盖硬件、软件、工具链或优化解决方案提供商。

天罡智算　FROST & SULLIVAN 沙利文

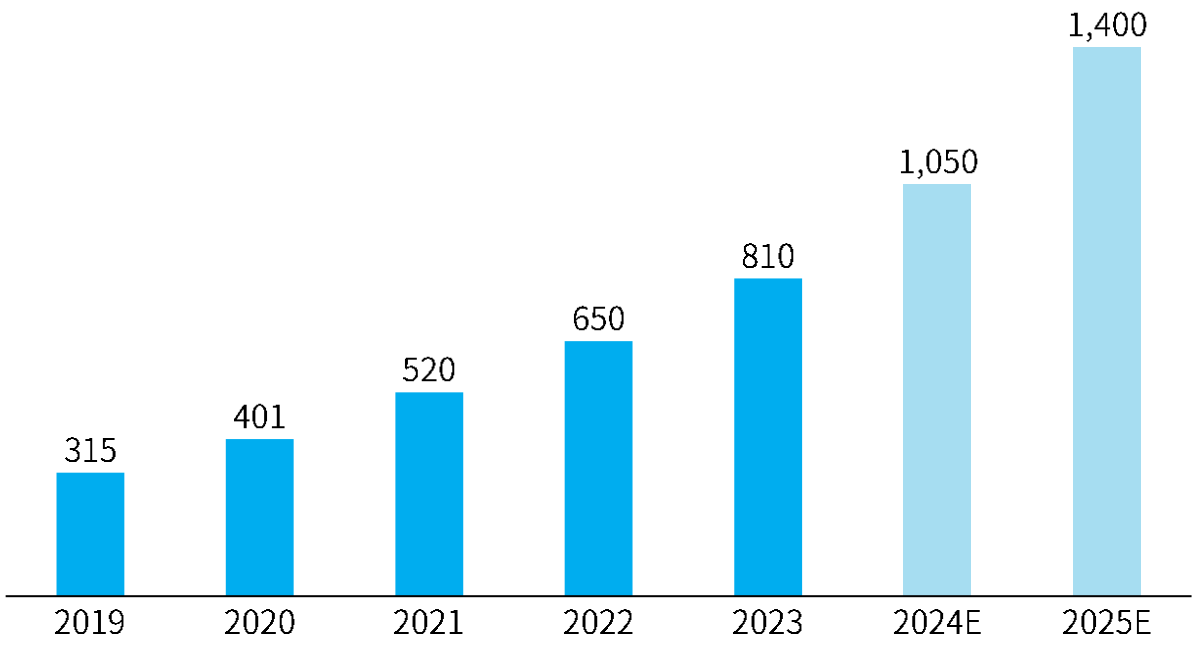
注：此版为第一版产业图谱，未来将根据市场变化持续更新。

中国智能算力产业链透视-中游智算资源供给层

数据中心为智算资源供给方搭建和运维物理平台，并为其提供海量的数据存储，高速的网络传输和实时处理服务，为配合高性能智算资源的供给，传统数据中心在架构、性能、可扩展性和安全性等方面都进行了升级改造

中国数据中心机架累计数量

2019-2025预测，万个



运营商数据中心布局情况

	移动云	天翼云 State Cloud	联通云 Unicom Cloud
机架数量（万台）	47.8	53.4	38.0
资源布局	4+3+X	2+4+31+X	5+4+31+X
中心区域	4：京津冀、长三角、粤港澳、成渝	4：京津冀、长三角、粤港澳、成渝	5：京津冀、长三角、粤港澳、成渝、鲁豫陕
低成本中心	3：呼和浩特、哈尔滨、贵阳	2：内蒙、贵州	4：内蒙、贵州、甘肃、宁夏
省级节点	X：省级数据中心和地级市节点等	X：省级数据中心和云节点	X：省级数据中心

中国大力推动数据中心发展 机架数量持续提升

- 随着我国数字经济的发展，政府和行业需求方大力推动数据中心的建设，2023年我国数据中心机架累计已达810万台。
- 2023年，我国政府印发《数字中国建设整体布局规划》，明确提出要夯实数字基础设施和数据资源体系，预计未来我国数据中心机架数量将持续增长，在2025年达到1400万台。

运营商加速布局数据中心 打造完整数字化网络

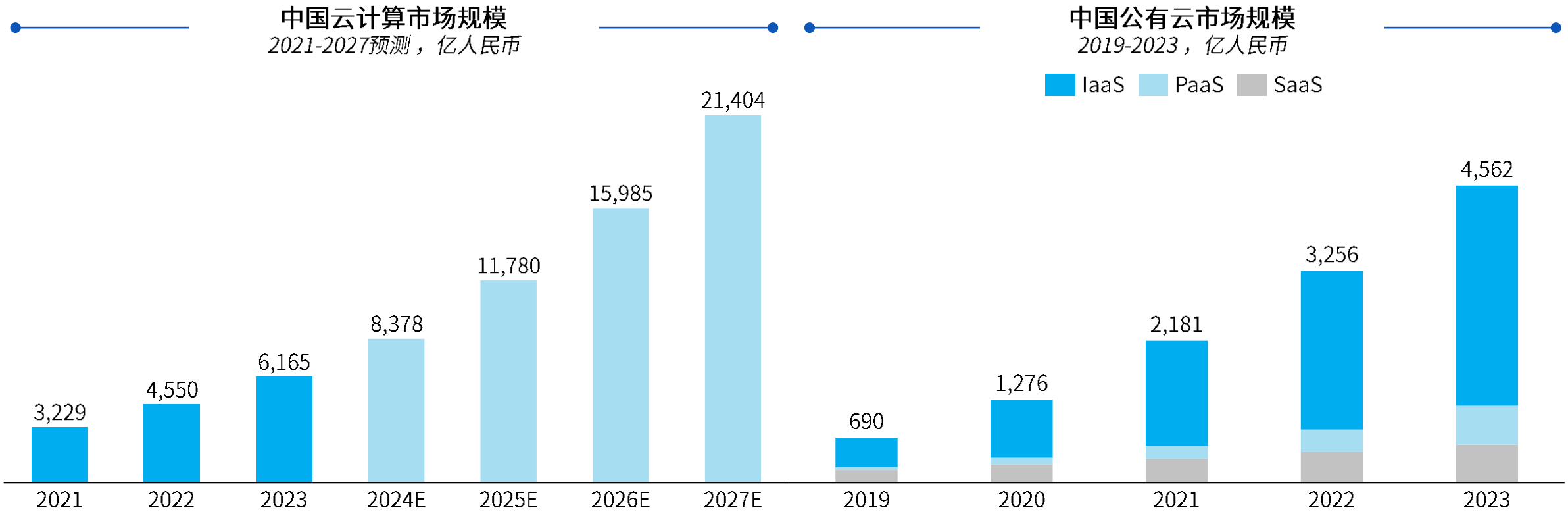
- 在中国数字经济的蓬勃发展背景下，三大电信运营商作为国内数据中心建设的主导力量，正积极整合其在互联网数据中心（IDC）和网络资源方面的领先优势。同时，依托其在政企市场的深厚渠道优势，这些运营商正在加速数字基础设施资源的战略布局，以满足日益增长的数据处理和存储需求。

来源：中国信息通信研究院；工信部；弗若斯特沙利文

注：电信运营商数据中心机架布局情况截取2023年6月数据

中国智能算力产业链透视-中游智算资源供给层

云服务商是目前智算市场上的主要智算资源供给方，凭借其自身在云计算方面全面的技术能力和完善的产品矩阵，为下游客户提供算力资源和算法增值服务。2023年中国云计算市场规模超6000亿人民币，其中公有云服务占比75.4%



技术革新引领云服务市场新发展

- 我国云服务市场多年来一直保持高速发展，2023年中国云计算市场规模达6,165亿人民币。随着AI技术革新和未来大模型的应用落地，我国云服务市场即将开始新一轮增长，预计到2027年市场规模将达到21,404亿人民币。
- 2023年，中国公有云市场规模达4,562亿人民币，占整体云计算市场的74%。IaaS仍是公有云市场的最大组成部分，2023年市场总额达3,383亿人民币，占公有云总市场规模的74%；PaaS和SaaS市场也有所增长，2023年市场总额分别达到598亿人民币和581亿人民币。

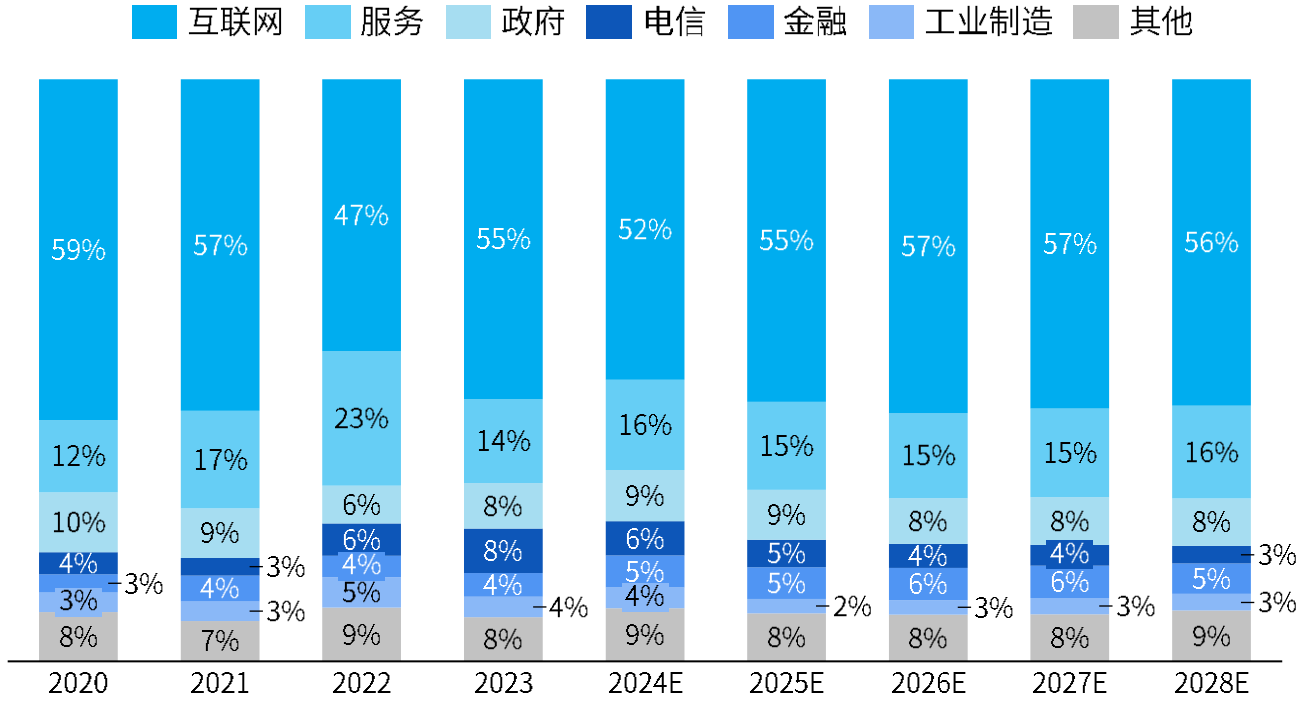
来源：中国信息通信研究院；弗若斯特沙利文

中国智能算力产业链透视-下游智算资源需求层

互联网一直处于大模型训练和推理领域的第一梯队，占据每年人工智能加速服务器市场的半壁江山，随着垂直行业大模型的逐步渗透和商业化落地，金融、能源、交通运输、教育、政府和智能制造等板块对智算资源的需求将逐渐凸显

中国人工智能加速服务器出货量按行业拆分

2020-2028预测，万台



各行业人工智能加速器出货量增速概览

2020-2028，%

行业	CAGR20-23	CAGR23-28E	行业	CAGR20-23	CAGR23-28E
互联网	31.7%	18.5%	教育	2.7%	17.4%
服务	41.4%	20.4%	公用事业	37.3%	21.4%
政府	22.8%	19.1%	资源	46.8%	28.2%
电信	69.5%	-1.8%	医疗健康	123.0%	15.2%
金融	47.3%	23.8%	媒体	58.4%	13.1%
工业制造	37.2%	13.1%	建筑	501.6%	17.6%
交通运输	25.9%	25.7%	流通分销	222.4%	6.0%

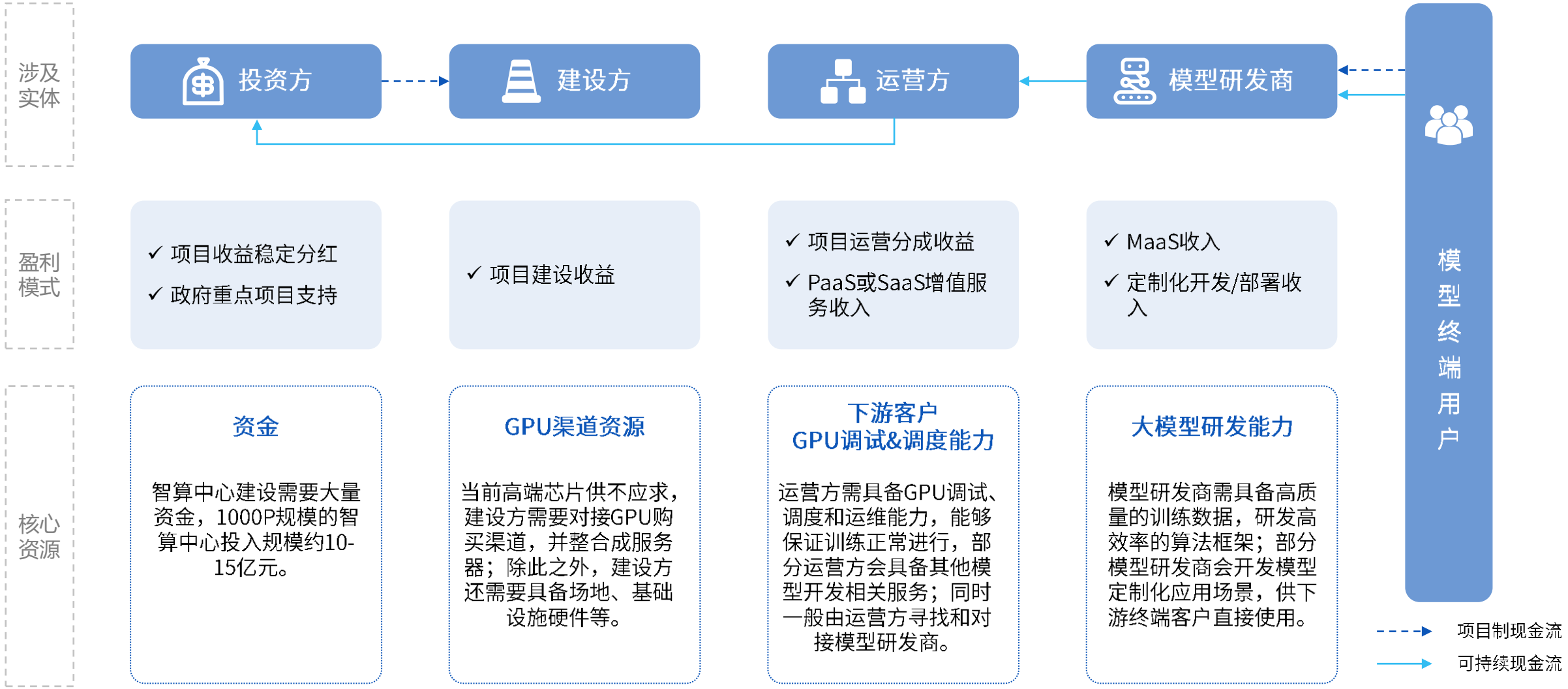
下游市场蓬勃发展 新应用推动增长

- 人工智能的应用已经渗透到社会的各个角落，成为推动各行各业技术革新和效率提升的关键力量。在过去几年中，人工智能技术在互联网、服务业、政府部门、电信业、金融业以及工业制造业等领域得到了广泛的应用。这些行业的AI加速服务器出货量占据了整个人工智能加速服务器市场的重要份额，显示出人工智能在这些领域的核心地位和重要性。
- 2020-2023年间，随着人工智能技术的不断成熟和普及，建筑业、流通分销业、医疗健康业、电信业和媒体业等行业的人工智能加速服务器出货量呈现出显著的增长趋势。展望未来，预计除电信行业之外，其他各行业的AI加速服务器出货量将保持持续增长的态势。AI加速服务器预计将在更多行业中发挥更加关键的作用，推动整个社会的数字化转型和智能化升级。

来源：IDC；弗若斯特沙利文

中国智能算力产业合作共建模式

智能算力产业属于技术密集型科技产业，从算力的感知，到算力的调度，再到算力的运营涉及众多的关键技术，而目前国内智能算力市场正处于先行先试的发展初期，因此，需要全产业链参与者凭借自身优势，实现合作共建



来源：弗若斯特沙利文

中国智算产业发展驱动因素分析

01 政府出台政策加快智算产业布局

- 中国政府高度重视智能算力产业发展，从多方面统筹考虑算力供给建设，出台了《十四五大数据产业发展规划》、《算力基础设施高质量发展行动计划》等政策文件，对算力设施布局、算力结构配置、边缘算力部署、标准体系建设作出了宏观规划，并启动“东数西算”重大工程，积极推动智算中心建设和大模型研发以满足快速迭代的算力市场需求。
- 地方政府围绕智能算力的规模、目标、能效和发展重点积极发布相关政策，推进智算基础设施的有序建设，提供普惠易用的算力服务，致力于形成规模化的先进算力供给。北京、上海、成都等地通过提供AI算力券，降低算力使用门槛，使中小企业和个人能够获得低成本、普惠的算力支持，助力中小企业实现数字化转型。

02 人工智能蓬勃发展带动智算需求

- 从需求侧看，人工智能正加速向金融、电信、工业制造等领域渗透，在产品设计、市场营销、供应链和生产流程管理、风险评估、客户服务、自动驾驶等方面发挥关键作用，推动产业智能化改造和数字化转型。需求侧传统产业的转型升级驱动智算发展。
- 人工智能在新兴领域的应用不断涌现，诸如元宇宙、人形机器人、生物制造、未来网络、新型储能等。人工智能的创新应用带动智能算力需求激增，进一步推动了智算市场的高速发展。

03 算力技术迭代推动智算高速发展

- 随着人工智能不断发展，大模型所使用的数据量和参数规模呈现指数级增长，对算力提出了更高要求。得益于智算处理大规模数据和复杂计算的能力，智算需求快速增长。
- 面对人工智能场景覆盖增加的趋势，算力和网络技术正向泛在化、确定性和协同化方向演进。从供给侧看，算力技术加速迭代创新，推动算力智能化和一体化算力输出，以面对海量数字内容需求和智能化终端，推动智算长效发展。

来源：弗若斯特沙利文

中国智算产业痛点分析

智能算力供给痛点

数据安全保障难度大 + 基础设施配置投入大 + 芯片供应链不稳定 + 专业技术要求极高 + 综合运维成本高 + 算力闲置量预判困难

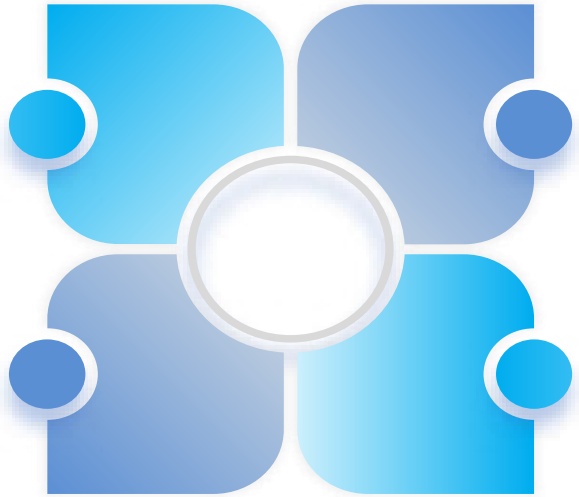
智能算力生态痛点

智算资源分散

◆ 智能算力资源提供主体多样，地域分布广且规模不一，呈现东部不足、西部过剩的不平衡局面。

算力调度能力不足

◆ 大多数智能算力运营主体缺乏成熟调度体系，导致在大规模数据处理和大模型训练中，智能算力资源浪费严重。



AI应用跨架构迁移困难

◆ 计算芯片种类繁多，计算框架各不相同，使得AI应用在进行跨平台迁移时面临兼容性和性能优化等挑战，增加了开发和运维的复杂性。

算力供需不平衡

◆ 智能算力的规模占比不足，难以满足不断增长的AI应用需求，限制了整体算力资源的高效利用和服务能力的提升。

智能算力需求痛点

- 部分AI模型对集群算力的要求较高，需配备大量的计算资源以支持其训练和运行。该类智能算力需求对算力的集成和调度提出了更高的挑战。
- 当前集中化的供需以及技术水平限制限制了智算资源灵活配置，导致智能算力多样化服务较少。目前主流智能算力需求多呈现为稳定长期、大流量的特点，但独立开发人员、小规模智算项目、以及大型项目中短期但大量的算力需求难以得到有效支持。

来源：弗若斯特沙利文

中国智算产业发展趋势分析

智算中心的区域化协同

- 多模态大模型的发展推动智算中心分布式训练，东西部跨区域协同增强，智算中心间通过全光网等实现超低延迟和超高带宽的交互。
- 大模型与业务场景的深度融合将成为主流应用模式，推动集中式训练和边缘分布式推理的两极化部署。
- 终端设备性能提升加速大模型从云端向终端的部署，云-边-端协同应用成为主流。

智算服务的普适普惠

- 智算中心的运营向平台调度算力的模式转变，旨在降低算力的使用门槛和智算中心的运维难度，提高算力的利用效率，实现算力资源的按需分配、动态调整。
- 地方级算力公共服务平台的服务加速算力普惠，为中小企业、高校等提供使用快速便捷的人工智能算力调度和个性化开发服务，算力服务模式从资源式向任务式转变。

智算的绿色低碳转型

- 算力生产、算力运营、算力管理和算力应用的绿色化是智算产业发展的重要方向之一。智算中心积极引入绿色能源，通过源网荷储协同互动等方式提升算力设施绿电使用率，推进绿证交易，呈现绿色转型新趋势。
- 绿色算力与电力、工业等重点碳排放领域的深度融合有助节能提效，实现环境和业务的双重可持续发展。

来源：弗若斯特沙利文

中国智算产业竞争壁垒分析

复合型人才

- 通用大模型向垂直大模型的转变意味着专业人才需要在数据处理、行业知识挖掘和大模型搭建优化方面具备综合能力。垂直大模型不仅要求技术人员具备深厚的人工智能和机器学习基础，还需要他们深入理解特定行业或领域的专业知识，将专业知识融入大模型的构建过程中。例如，在医疗、金融、制造等垂直领域，模型的训练必须结合行业特有的规则、流程和数据。此外，垂直大模型的有效性依赖于海量行业数据的积累与分析，专业人才需要通过迭代模型来实现持续改进和创新，以确保模型能够适应不断变化的应用场景和业务需求。复合型人才是智能算力市场当前发展面临的重要壁垒之一。

资金壁垒

- 智能算力资源提供商（如智算中心，ICT硬件集成商等）具有较高的资金壁垒，主要体现在高额的前期投入和持续运营成本中。
- 前期投入主要聚焦于硬件采购和系统集成。例如，高性能计算服务器、GPU等核心设备的采购成本极为高昂。
- 随着企业规模的扩张，其运营成本也同步上升，包括日益增长的能源消耗、设备的维护和更新、以及高素质技术人才的招聘和留用等。高建设成本与高运营成本的特点使新进入企业面临较大的资金压力，市场资源和机会逐渐向少数资本雄厚的大型企业集中。

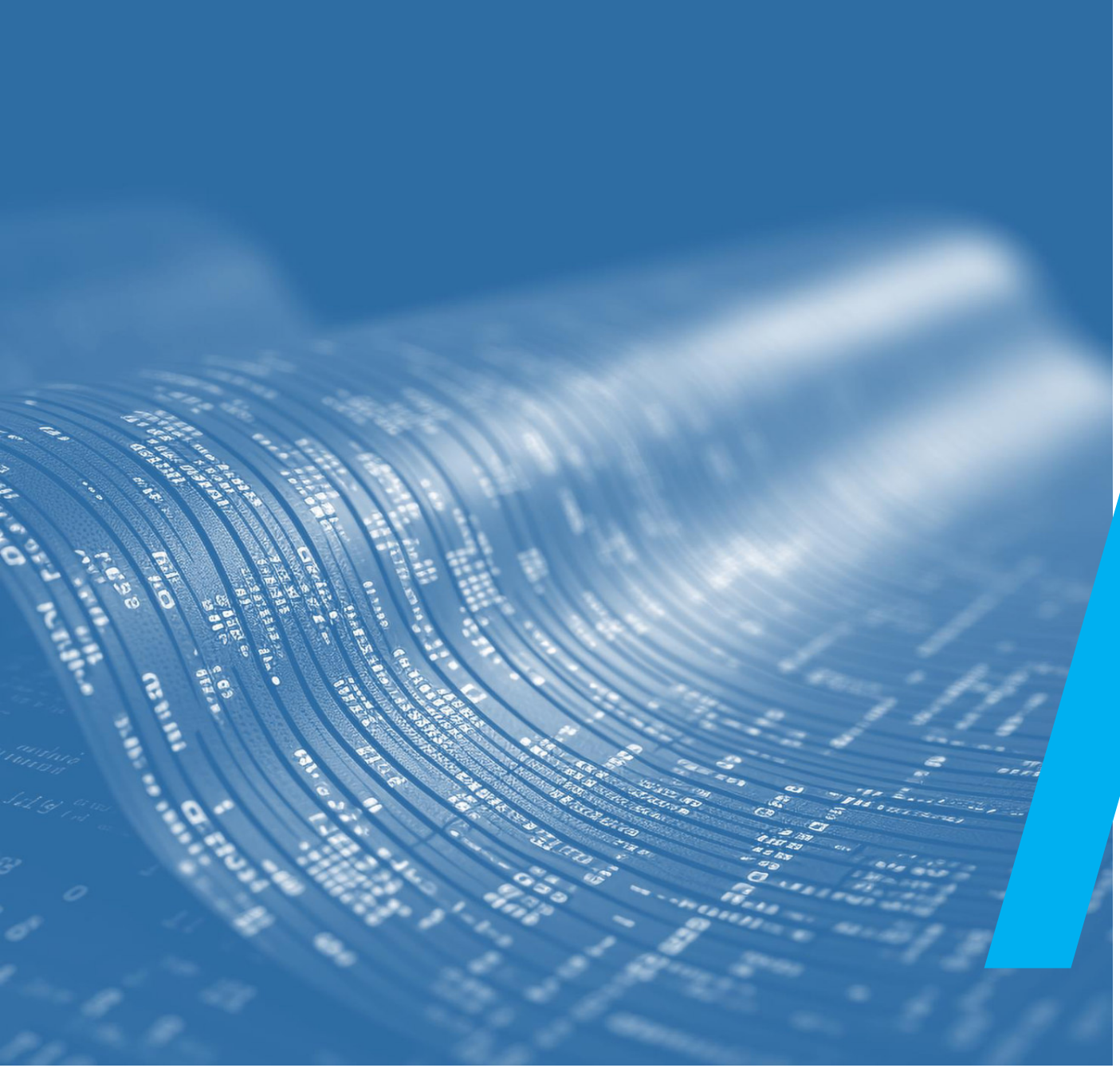
品牌效应

- 云服务厂商的品牌效应构成了强大的竞争壁垒，主要体现在技术领先性、市场渗透度以及客户信任度上。诸如阿里云、腾讯云、火山云等云服务巨头，经过多年的发展，其公有云平台在系统稳定性、功能丰富度和用户体验上已经过市场的充分验证，积累了广泛的用户基础和高度的品牌认可度。
- 客户在转向新的云平台时还面临着较高的转移成本，包括数据迁移和系统重构。因此，新进入企业由于缺乏时间积累和用户沉淀，难以在短时间内建立起同等的市场信任度和品牌影响力，使其在竞争中处于劣势。

增值服务

- 智能算力资源提供商增值服务的技术壁垒尤为突出，成为企业竞争的关键因素。增值服务不仅涵盖了从数据处理、模型开发到部署的全流程，还依赖于多种先进技术的综合应用，如通信加速、内存管理优化、算法加速和并行处理等。通过这些优化措施，企业能够显著提升硬件效能，减少资源浪费，为客户提供更具竞争力的解决方案。
- 企业通过对不同算力资源的适配、纳管，以及跨智算中心的算力调度，显著提升算力资源的整体利用率，进一步强化了行业领先者的市场主导地位。

来源：弗若斯特沙利文



3

- 需求跃迁
智算租赁开启新篇章

智算租赁兴起的底层逻辑拆解

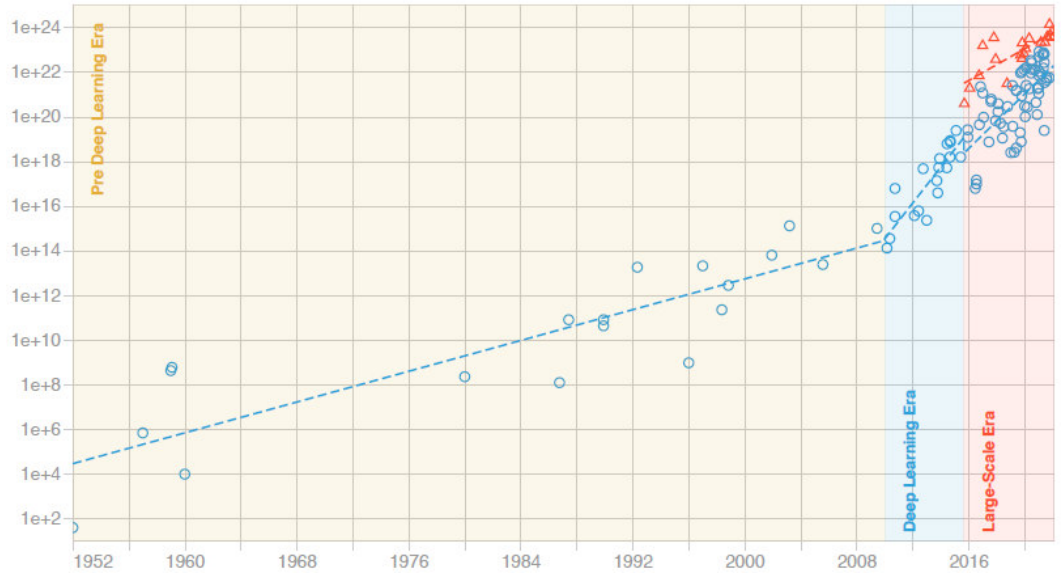
大模型所需的算力规模与算法结构、参数量、数据量以及训练轮数等紧密相关，而当模型规模突破某一阈值时，其性能会显著提升。因此，依托大规模智能算力集群的模型训练，成了模型优化的必然选择

- 自2017年Transformer架构发布以来，模型在处理更长序列和更复杂任务方面取得了显著的进步。模型的训练效果与参数量、数据量及训练轮数紧密相连；并且，研究指出，当模型规模达到某一临界值时，其性能将实现显著提升。为了优化模型的训练效果，必须利用大量计算资源以加速训练过程。
- 到了2022年，基于大模型的人工智能生成内容（AIGC）技术迅速崛起，标志着人工智能技术的又一次革命。全球众多大型企业正在积极投资于大型模型和AIGC领域，这一趋势进一步加剧了对智能算力的大量需求。

	前深度学习时代	深度学习时代	大模型时代
	~2010年	2010年~2015年	2015年~至今
描述	2010年之前，模型训练所需的计算资源增长与摩尔定律的预测相吻合，呈现出大约每20个月翻一番的稳定增长趋势。	自2010年前后深度学习技术的出现与兴起，模型训练所需的计算资源快速增长，大约每6个月便实现翻倍。	自2015年末大模型技术的兴起，模型训练所需的智能算力需求迎来了更为剧烈的增长，短时间内扩大了10到100倍。
算力需求	3e+04 to 2e+14 FLOPs	7e+14 to 2e+18 FLOPs	4e+21 to 8e+23 FLOPs
算力翻倍所需时间	21.3个月	5.7个月	9.9个月

主要机器学习系统训练算力

1952-2022，FLOPs



注：蓝色虚线代表深度学习模型训练所需算力随时间的变化，红色虚线代表大模型训练所需算力随时间的变化

来源：《Compute Trends Across Three Eras of Machine Learning》；弗若斯特沙利文

智算租赁兴起的底层逻辑拆解

训练阶段，大模型需要处理和学习大量的数据，从而识别数据模式和特征，微调阶段，是对预训练模型的调整，以适应特定的应用场景或数据集，推理阶段，则要求模型对新数据做出反应，可见，全周期的算力消耗，是大模型功能的关键。

训练阶段

模型训练是一个使用大量已知数据（通常包括特征和标签）来训练机器学习模型的过程。在训练完成后，通常需要对模型的参数进行微调。训练和微调的目的让模型更精确地学习数据中的特征和模式。

推理算力核心影响因素

- 模型数量
- 参数量
- 训练数据量
- 训练次数

推理阶段

模型推理，也称为模型预测或部署，是在模型训练和微调阶段完成后，投入实际使用并生成结果的阶段。推理阶段的目标是应用经过充分训练的模型来对新出现的、未知的数据进行预测或分类。

推理算力核心影响因素

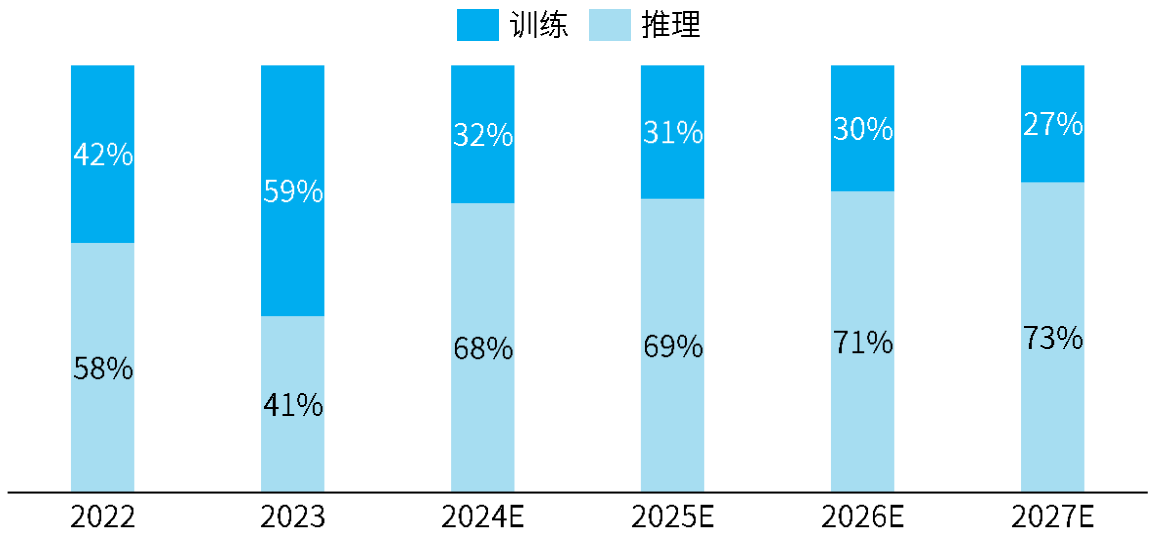
- 模型数量
- 应用场景
- 单用户数据量
- 用户日活
- 应用时间

热门大模型消耗卡的数量及训练时长

模型	消耗卡的数量	训练时长
LLaMA-65B模型	2048张A100 80GB	21天
GPT3-175B模型	1024张A100 40GB	34天
GLM-130B模型	768张A100 40GB	60天
Falcon-40B模型	384张A100 40GB	60天
Inflection-2模型	5000张H100	-
Megatron-Turing模型	4480张A100	-
GPT4-1800B模型	2.5万张A100	90~100天

人工智能服务器推理和训练工作负载

2022-2027预测，%



来源：《LLaMA(Large Language Model Meta AI)》；《GLM-130B：开源的双语预训练模型》；弗若斯特沙利文

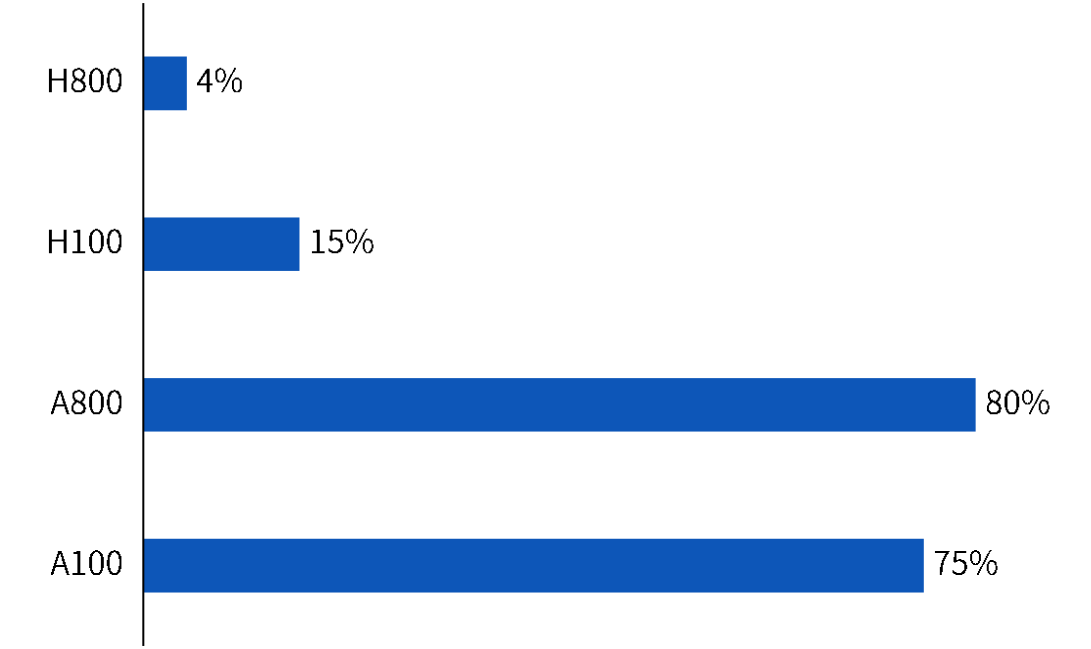
智算租赁兴起的底层逻辑拆解

现阶段，大模型的训练、微调和推理主要依赖以英伟达为代表的高性能GPU来提供强大的并行计算能力、显存容量和成熟的软件生态支持，算力需求的跃迁式增长导致了市场供需失衡，价格上涨

- 目前，大型模型的训练主要依赖于英伟达A100/H100等高性能GPU的算力，这些GPU不仅能够提供高效的数据处理能力，还有助于最大限度地减少智能算力的闲置。
- 由于市场需求的激增，英伟达GPU的价格一直在不断攀升。
- 尽管英伟达不断增强其生产能力，且出货量持续增加，但市场的需求仍未得到完全满足。

英伟达GPU价格涨幅

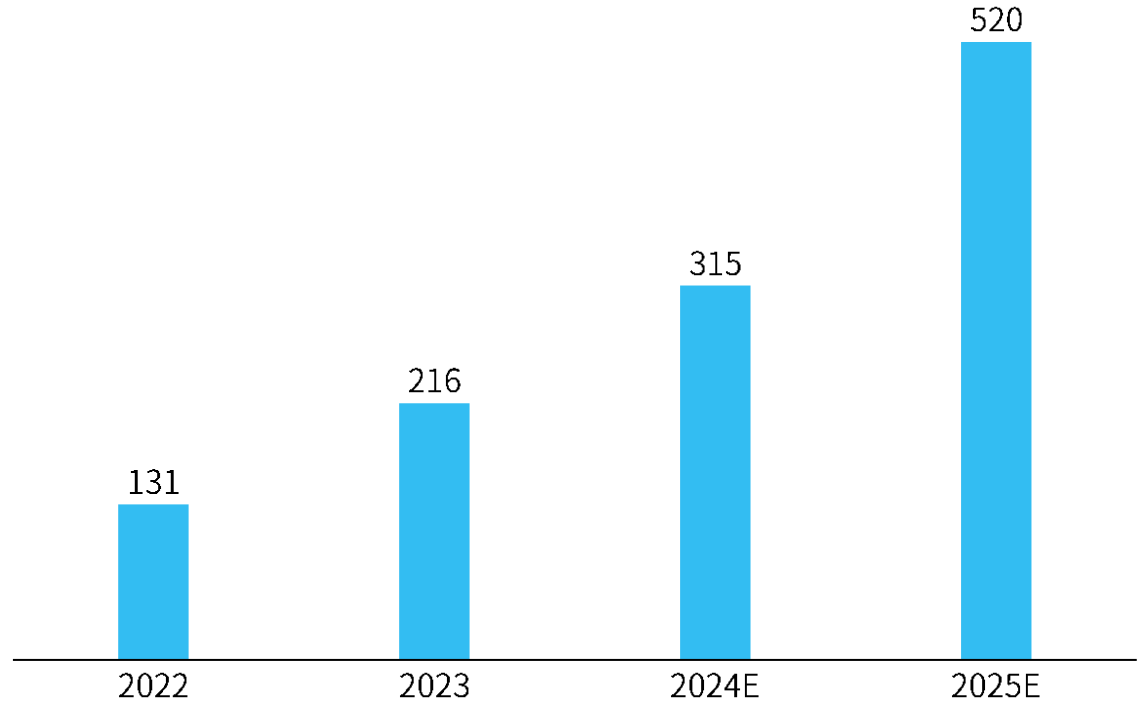
2023年1月-2023年10月，%



注：H800累计涨幅统计自2023年7月起

英伟达训练芯片出货量

2022-2025预测，千张



来源：弗若斯特沙利文

注：英伟达训练芯片包括A100、A800、H100、H800、H200、H20和B100

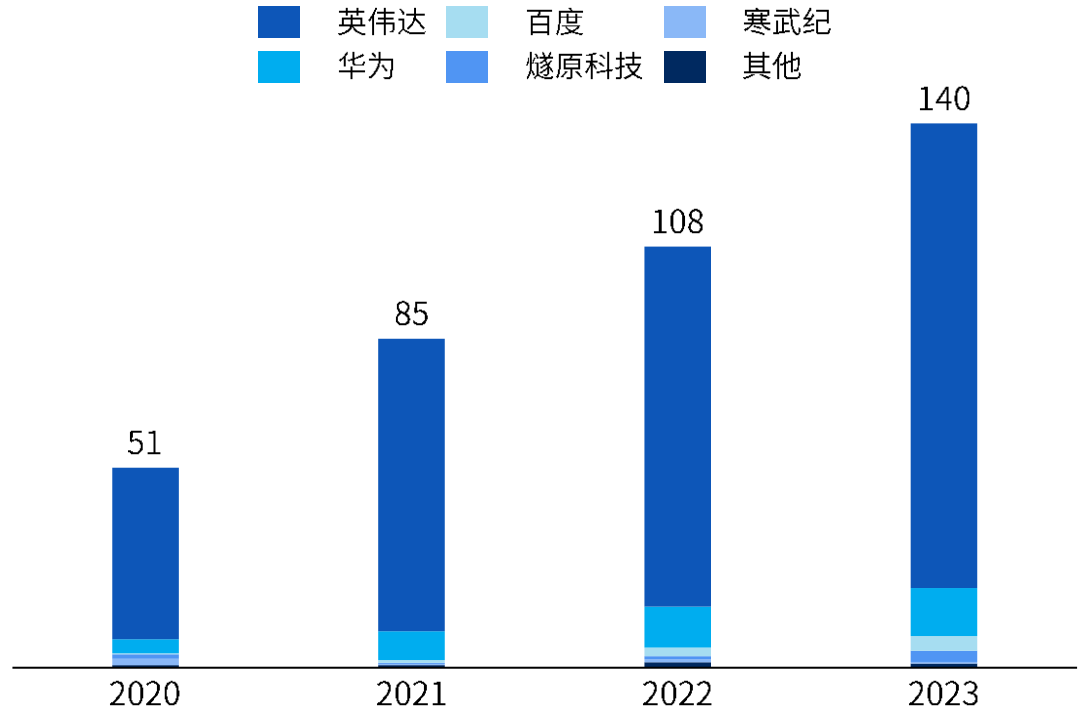
智算租赁兴起的底层逻辑拆解

2023年中国市场加速芯片出货量达140万，其中英伟达以119万张占据85%的市场份额，同时，这些高性能芯片大部分被中国的头部互联网企业所持有，中小公司和创业公司芯片购买获取难度增加

- 在过去数年中，英伟达稳居中国加速芯片市场出货量首位。2023年，英伟达向中国市场供应了119万张加速芯片，占该年度中国加速芯片总出货量的85%。在2023年英伟达对华销售的加速芯片中，A100、A800及H800型号的芯片合计出货量达到56万张，占其对中国市场总销量的47%。
- 尽管英伟达对中国市场的加速芯片销售量显著，但高端芯片的流向仍主要集中在互联网行业的领军企业。以H100芯片为例，2023年英伟达全球出货量达到65万张，这些芯片主要被分配给了全球顶尖企业，在中国，仅有腾讯、百度、阿里巴巴和字节跳动等少数几家公司获得了H100芯片的供应。

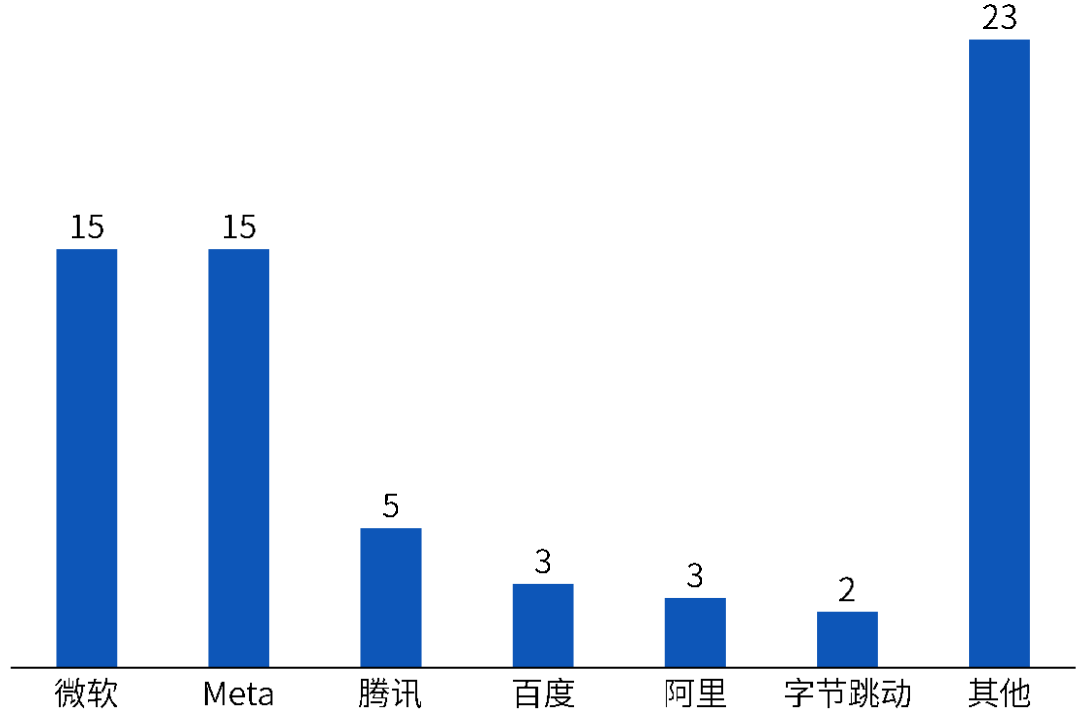
中国加速芯片出货量

2020-2023，万张



英伟达H100芯片流向情况

2023，万张



来源：IDC；弗若斯特沙利文

注：其他厂商为微软、谷歌、亚马逊、Oracle、CoreWeave、Lambda、特斯拉

智算租赁兴起的底层逻辑拆解

叠加美国科技封锁的政策性影响，中国市场上海外高性能GPU供给更为紧缺，而国产GPU短时间内，在技术水平和实际运用方面都还未能实现完全替代，因而，高性能智算资源，成为稀缺产品

- 自2018年起，美国政府相继推出多项针对中国的科技封锁政策。2022年10月，美国政府要求英伟达和AMD等公司停止向中国出口尖端人工智能芯片。2023年，限制措施进一步加剧，英伟达的高端智能计算芯片，包括但不限于A100、H100、A800、H800等型号，均受到出口限制。
- 除了对芯片硬件实施出口管制外，英伟达目前还在监控这些高端芯片的最终用户位置，并禁止其在中国境内运行。受政策影响，中国市场上的高端芯片无论从软件还是硬件角度来看，均已成为稀缺资源。

美国芯片出口限制

2022年芯片限制

芯片出口限制条件：

① 传输速率为前提，当I/O接口传输速率大于或等于600GB/s，且总算力超过300TFLOPS

② 若传输速率低于600GB/s，就算总算力超过300TFLOPS也不会被限制

2023年芯片限制

以下三种情况芯片出口会被限制：

① 总算力超过300TFLOPS

② 总算力超过100TFLOPS，性能密度超过0.2TFLOPS/mm²

③ 总算力超过150TFLOPS，性能密度超过0.1TFLOPS/mm²

2018年
- 未来7年禁止美国公司向中兴通讯开展任何业务往来
- 发布涉及人工智能和机器学习技术、先进计算技术等14项前沿技术的对华出口管制框架

2019年
- 将华为列入出口管制“实体清单”，美国成分超25%的产品或技术，需经美国批准后，方可与华为合作
- 将中科曙光、天津海光等5家与超算有关的企业/机构列入“实体清单”

2020年
- 将中芯国际列入实体清单
- 用于10nm及以下技术节点的产品成技术，采取“推定拒绝”的审批政策

2021年
- 将天津飞腾、申微科技、上海集成电路研发中心等7家中国超算实体及GPU龙头景嘉微、亚成微等34家企业列入“实体清单”
- 限制中国台湾的半导体业务出售给中国大陆

2022年
- 禁止获得美国资助的企业扩大在中国半导体领域的投资，禁止出售14nm以下的半导体设备出售给中国大陆企业
- 英伟达与AMD均被要求停止向中国出口用于人工智能的最先进芯片
- 长江存储、寒武纪、上海微电子等36家中国实体列入美出口管制“实体清单”

2023年
- 限制中国购买和制造中高端芯片的能力，涉及英伟达A800、H800、L40S 及 RTX 4090 等芯片将受到限制
- 壁仞科技、摩尔线程等多家中国GPU芯片企业被列入实体名单

2024年
- 要求美国云计算公司验证其外国用户的身份，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
- 全面限制英伟达、AMD以及更多更先进 AI 芯片和半导体设备向中国销售，搬销高通和英特尔公司向华为出售半导体的许可证

来源：U.S. Bureau of Industry and Security；英伟达官网；弗若斯特沙利文

智算租赁兴起的底层逻辑拆解

除受到上述客观因素的制约外，综合考虑部署成本，运维成本，时间成本，算力资源的灵活性以及数据的安全性，智算租赁都大幅降低了使用高性能智算资源的门槛

考量因素	智算租赁	自建算力
部署成本	以阿里云为例，8卡英伟达A100-NVLink（80GB显存）的GPU服务器的月租金约为13.34万元，对应全年租金约为160万元，无其他成本	8卡英伟达HGX加速显卡A100 SXM模组市场售价约为145万元，除此之外仍需部署交换机（约20万元）、配件（约15万元）、基建等
运维成本	智算租赁提供商负责硬件维护和技术支持，且无需承担额外成本	涉及资源管理审批、调配、监控等工作，需自行设立维护团队，并且需要承担电力、冷却和网费等固定运营成本
时间成本	大部分租赁形式无需等待硬件采购和部署，租用后可以立即开始计算任务，小部分长期租用的情况也仅需对服务器进行微调	当前英伟达出货周期约为3-4个月，服务器送达后需要进行调试部署，从下单到使用计算的时长难以确定
灵活性	可根据需求快速调整使用芯片的型号、租赁时间和带宽等，适应项目需求变化	服务器等硬件可能随时间热迭代升级，带宽一次性购买，无法灵活变更
数据安全	数据存储在智算租赁提供商的存储服务器内，安全性和故障解决能力取决于智算租赁提供商	自行搭建存储设备，数据存储在企业内部存储设备内，故障解决能力取决于企业自身

关键发现

- 智算租赁在资金成本和运营灵活性方面展现出显著的优势。特别是在时间方面，对于模型研发企业而言，能够尽快完成模型研发的企业将更有可能获得市场先发优势。
- 大模型的研发往往需要GPU与CPU的协同工作，这要求对计算资源进行精确的分配和管理。通过租赁方式，模型研发企业能够灵活地部署所需的智能算力资源，从而提高GPU的使用效率，减少资源浪费，实现资源配置的优化。
- 对于互联网大型企业而言，自建算力仍然是一个较为理想的解决方案。这些企业通常拥有庞大的数据中心和专业的运维团队。自建智能算力不仅可以提升运维团队的工作效率，增加人均产出，还可以通过将未使用的GPU资源进行池化并对外出租，进一步提高单位智能算力投资的回报率。

来源：阿里云官网；弗若斯特沙利文

智算租赁市场空间容量测算-训练端

以ChatGPT3.5的训练成本作为估算基础，中国头部AI大模型厂商的参数规模通常在1000亿左右，而初创企业的大模型参数通常在10亿到100以之间，假设平均单个模型参数规模为800亿，则100个模型在训练阶段需要26,380张A100

$$\text{单个大模型训练算力需求(FLOPs)} = \frac{6 \times \text{模型参数数量} \times \text{Token数量}}{\text{单次训练秒数}}$$

$$\text{GPU需求(张)} = \frac{\text{算力需求}}{\text{单张GPU峰值算力} \times \text{GPU利用率}}$$

GPT-3大模型训练端需求

- GPT-3模型训练参数：参数数量=1750亿 Token数量=3000亿。
- 假设1：30天训练完单个GPT-3大模型。

$$\text{单个GPT}-3\text{大模型训练算力需求} = \frac{6 \times 1750 \times 10^8 \times 3000 \times 10^8}{30 \times 24 \times 60 \times 60} = \mathbf{1.2 \times 10^{17}\ FLOPs}$$

- 假设2：采用混合精度（FP16）的A100芯片训练，即每张A100的峰值算力312 TFLOPs。
- 假设3：GPT-3训练期间A100芯片利用率为45%。

$$\text{单个GPT}-3\text{大模型训练端A100需求} = \frac{1.2 \times 10^{17}}{312 \times 10^{12} \times 45\%} = \mathbf{865.6}\ \text{张}$$

GPT-4大模型训练端需求

- OpenAI训练GPT-4的所用算力约为2.15×10^{25} FLOPs，采用A100芯片训练了90-100天，利用率约为32-36%。
- 假设训练100天，训练期间A100芯片利用率为35%，则推算需要22,787.8 张A100芯片。

中国大模型训练端需求测算

- 根据公开资料，截止2024年4月底，中国10亿以上参数规模的大模型数量已超100个。其中，仅百度、阿里等头部互联网企业发布的大模型参数规模达1000亿，其余初创企业大模型参数规模通常在100亿、10亿级别。
- 假设1：目前中国大模型平均每个模型参数数量为800亿，平均单个模型Token数量为2000亿。
- 假设2：每个模型的训练周期为30天。

$$\text{平均单个模型训练算力需求} = \frac{6 \times 800 \times 10^8 \times 2000 \times 10^8}{30 \times 24 \times 60 \times 60} = 3.7 \times 10^{16}\ \text{FLOPs}$$

- 假设3：采用混合精度（FP16）的A100芯片训练，即每张A100的峰值算力312 TFLOPs，训练期间A100芯片利用率为45%。

$$\text{平均单个模型训练端A100需求} = \frac{3.7 \times 10^{16}}{312 \times 10^{12} \times 45\%} = \mathbf{263.8}\ \text{张}$$

- 假设4：仅估算10亿以上参数规模的大模型，**模型数量为100个。**

$$\text{中国大模型训练端算力需求} = 3.7 \times 10^{16} \times 100 = \mathbf{3.7 \times 10^{18}\ FLOPs}$$

$$\text{中国大模型训练端A100需求} = 263.8 \times 100 = \mathbf{26,380}\ \text{张}$$

来源：《Scaling Laws for Neural Language Models》；弗若斯特沙利文

智算租赁市场空间容量测算-推理端

以ChatGPT3.5的推理成本作为估算基础，假设中国头部AI大模型厂商平均单个模型的日均访问量为100万次，访客每次提问需要10,000个Token,每天有50个大模型在进行推理，推理阶段则需要A100数量约为3,300张

推理端每日***Token***数量 = 模型日均访问量 × 平均每次提问数量 × 单次提问所需***Token***数量

$$\text{单个大模型每秒推理算力需求(FLOPs)} = \frac{2 \times \text{模型参数数量} \times \text{推理端每日Token数量}}{\text{单次训练秒数}}$$

$$\text{GPU需求(张)} = \frac{\text{算力需求}}{\text{单张GPU峰值算力} \times \text{GPU利用率}}$$

ChatGPT大模型推理端需求

- 根据Similarweb，2024年7月ChatGPT月访问量达24亿次，平均每天约0.7亿次。
- 假设1：ChatGPT访客平均每次提问10次，每次提问平均所需1000个Token。

$$ChatGPT\text{推理端每日Token数量} = 0.7 \times 10^8 \times 10 \times 1000 = 7.0 \times 10^{11}$$

- 假设2：ChatGPT模型参数数量为2000亿。

$$\text{ChatGPT每秒推理算力需求} = \frac{2 \times 1750 \times 10^8 \times 7.0 \times 10^{11}}{24 \times 60 \times 60} = \mathbf{2.8 \times 10^{18}\ TOPs}$$

- 假设3：采用Int8精度的A100芯片进行推理，即每张A100的峰值算力624 TOPs。
- 假设4：推理期间A100芯片利用率为45%。

$$\text{ChatGPT推理端A100需求} = \frac{2.8 \times 10^{18}}{624 \times 10^{12} \times 45\%} = \mathbf{10,000}\text{张}$$

中国大模型推理端需求测算

- 假设1：中国平均单个模型的日均访问量为100万次，访客平均每次提问10次，每次提问平均所需1000个Token。

$$\text{平均每个模型推理端每日Token数量} = 100 \times 10^4 \times 10 \times 1000 = 1.0 \times 10^{10}$$

- 假设2：目前中国大模型平均每个模型参数数量为800亿。

$$\text{平均每个模型每秒推理算力需求} = \frac{2 \times 800 \times 10^8 \times 1.0 \times 10^{10}}{24 \times 60 \times 60} = 1.9 \times 10^{16}\ \text{TOPs}$$

- 假设3：采用Int8精度的A100芯片进行推理，即每张A100的峰值算力624 TOPs，推理期间A100芯片利用率为45%。

$$\text{平均单个模型推理端A100需求} = \frac{1.9 \times 10^{16}}{624 \times 10^{12} \times 45\%} = 65.9\ \text{张}$$

- 假设4：仅估算10亿以上参数规模的大模型，考虑部分模型仍在训练阶段，假设推理阶段模型数量为50个

$$\text{中国大模型推理算力需求} = \mathbf{1.9 \times 10^{16} \times 50} = \mathbf{9.3 \times 10^{17}\ FLOPs}$$

$$\text{中国大模型推理A100需求} = \mathbf{65.9 \times 50} = \mathbf{3,297.5}\ \text{张}$$

来源：《Scaling Laws for Neural Language Models》；弗若斯特沙利文

智算租赁的商业模式

我国智算租赁市场正处于发展初期，参与者类型众多，各类参与者凭借自身的差异化优势，还在对适配且稳健的商业模式进行探索，主要包括云服务商的一站式解决方案，GPU算力池的租赁，GPU算力池的调度和搭载硬件来进行算力的交付

	商业模式	优势	劣势	代表性厂商
云计算一站式解决方案	先前专注于提供公有云服务的供应商，在部署GPU服务器等计算硬件后，将业务范围拓展至人工智能云服务领域，提供智能算力及模型开发相关服务	• 已有基建设施和运维团队，拓展智算业务可产生额外收入并提升原有基建和人员的利用率 • 已有开发客户基础，获客成本相对较低	• 服务成本较高，因此基本不支持弹性租赁服务，仅适合对智能算力需求较大且资金实力较强的客户	阿里云 百度云 天翼云 State Cloud 腾讯云
GPU算力池租赁	向客户提供自建或外采的智能算力资源，通常以整台服务器、单独GPU或算力规模为基础提供智能算力	• 可根据自有资金情况灵活部署智能算力硬件和软件等，形成规模化的智算租赁服务	• 需具备稳定可调配的智算资源 • 部分自建或共建的智算资源，在空余时仍需承担运维成本	润建股份 RIGF 中科曙光 Sugon CDS 首云 inspur 浪潮
GPU算力池调度	通常不准备智能算力硬件，而是采用轻资产的平台模式，整合和分配智能算力资源，将资源与需求方匹配	• 轻资产的模式对资金要求较低，仅需建设平台即可提供智算租赁服务	• 需具备稳定可调配的智算资源 • 对平台运营能力和营销能力要求较高	潞晨科技 蓝耘 LANYUN INFINIGENCE 无问芯穹 拔界算力
推训一体机模式	以华为昇腾系列产品为代表，与模型开发企业合作开发大模型推理训练一体化设备	• 软件与硬件打包售卖，产品价值较高，并且可从中收取长期稳定的订阅费用 • 本地化部署更易吸引对数据安全敏感的客户群体	• 硬件更新较为困难 • 对产品的售后和运维服务要求较高	科大讯飞 iFLYTEK HUAWEI 智谱·AI 华鲲振宇 HUAKUN

来源：弗若斯特沙利文

智算租赁的商业模式—云服务厂商一站式解决方案

云服务商是目前智算租赁市场的中坚力量，凭借其在云计算方面全面的技术能力和完善的产品矩阵，依托其底层的高性能智算资源，搭载其技术增值服务，为下游客户提供一站式的智算解决方案

- 云计算服务提供商在部署智算服务器等硬件基础设施后，将其业务领域扩展至人工智能云服务。他们在现有的云服务平台基础上增加智能算力服务，允许客户直接在平台上进行智算资源的租赁和交易。
- 凭借在云服务领域的深厚运营经验、丰富的客户资源以及强大的产品研发能力，云计算服务提供商能够提供与人工智能和智能算力紧密相关的软件服务，包括模型优化、模型开发，以及为特定垂直行业定制的应用软件解决方案等。

AI云计算一站式解决方案可提供的服务

应用层	闭源模型应用	**网络平台应用** 基于公共网络平台的应用，单一平台应用	**边缘或端侧应用** 基于本地边缘或端侧交互的应用
中间层	闭源模型应用	**数据中心与模型API** 数据中心工具与模型部署和调用API	**Prompt优化与模型微调** 基于预训练模型的场景/产品适配
模型层	闭源模型应用	**非开源预训练模型** 商业化的预训练模型，通过API调用	**模型Hub** 开源模型共享平台	**开源模型** 开源的预训练模型，一般可免费试用
框架层	**深度学习与中间件** 用于训练或云部署的深度学习框架和中间件等，包括PyTorch、TensorFlow		
算力层	**基础设备** 服务器、交换机、光模块、机架等	**云服务** 智算中心、云计算等	
	基础硬件 芯片、存储、PCB、网卡、电源等	**数据资源** （数据采集、存储、标注等）	

阿里云 AI云计算一站式解决方案案例：阿里云

阿里云创立于2009年，是全球领先的云计算及人工智能科技公司，为200多个国家和地区的企业、开发者和政府机构提供服务。

- 阿里云是全球领先的云计算和人工智能服务商，提供计算、容器、存储、网络与安全、数据库、大数据计算、人工智能与机器学习、企业服务与云通信等产品，基本囊括所有与云计算有关的硬件、软件和运维服务等。
- 在智算领域，阿里云的GPU云服务器提供软件与硬件结合的完整服务体系，助力开发企业在人工智能业务中实现资源的灵活分配、弹性扩展、智能算力的提升以及成本的控制，通常适用于深度学习，视频编解码，视频处理，科学计算，图形可视化，云游戏等场景。

■ 多样化计算能力

拥有大量擅长处理大规模并发计算的算术逻辑单元（ALU），持续采用最新GPU加速芯片，提供FPGA, GPU, ASIC等多种加速卡，为AI，图形，转码，加密等不同业务提供服务。

■ 简单易用

全球部署GPU资源充裕，分布广泛，逻辑控制单元相对简单，可以随时应对客户业务弹性扩容。提供AIACC AI加速引擎，FastGPU套件，cGPU套件等专有的辅助工具。

■ 高网络性能

采用神龙计算架构提升服务器性能降低IO延迟，最大支持2400万pps和64GbpsVPC网络及800G高带宽RDMA网络，能够支持多线程并行的高吞吐量运算。

来源：阿里云官网；弗若斯特沙利文

智算租赁的商业模式—GPU算力池租赁

用户可以根据自己的需求租用不同配置的GPU，并按小时、包月或者包年或者整机租赁的方式支付费用，GPU算力池租赁的盈利模式主要是赚取租金收入和运营成本之间差额，目前微软的GPU算力租赁业务的毛利率高达42%

- 通过自建、与合作伙伴共建或外采的方式获取智能算力硬件的使用权，并根据使用时间将算力硬件或算力资源直接出租给下游智能算力需求方。
- 该类型的商业模式下，租赁方普遍会建设智能算力交易平台，需求方可以自行在平台上选择租赁类型和时间。
- 部分租赁方会额外提供智能计算相关的软件和服务，包括模型优化和模型开发等。

GPU算力池租赁模式

按整台服务器租赁

租金每台服务器（含4-8张GPU）为单位计量，可选择年租、月租等；适用于需要长期稳定使用大量GPU资源的客户。

按单张GPU租赁

租金按照每GPU为单位计量，可选择年租、月租、日租、小时租等；适用于短期或临时性的智算需求，可根据需求随时调整卡的租赁组合和数量。

按算力规模租赁

租金按照每P为单位计量，可选择年租、月租等；可以根据需求选择合适的智能算力规模，实现灵活的资源配置。

GPU算力池租赁案例：润建股份

润建股份（002929.SZ）成立于2003年，是领先的数字化智能运维服务商。

- 2024年，润建股份宣布了其在人工智能领域的发展战略。该战略以智能算力服务和数据服务为核心基础，并且依托公司自主研发的“曲尺”平台，致力于生成和开发人工智能行业模型。
- 在智能算力服务领域，润建股份推出了包括智能算力租赁和智能算力中心运维管理在内的多项业务。公司致力于构建云计算中心和智算中心，旨在为AI大模型的训练、推理以及图形渲染提供专业的智能算力服务，支持AI大模型和特定行业模型的开发与应用。

来源：润建股份官网；弗若斯特沙利文

智算租赁的商业模式—GPU算力池调度

在高性能GPU供需不平衡的情况下，实现高性能智算资源的高效利用成为了破局的关键，因此出现了以提供算力发现、供需撮合、交易购买、调度使用为服务的商业模式

- 智能算力租赁厂商通常不准备算力硬件，而是采用轻资产的平台模式，整合和分配智能算力资源，将资源与需求方匹配。
- 通常会构建自有的平台系统，实现对智算资源的集中管理和调度，客户可以在平台上选择所需的智能算力资源，并根据使用时间或计算需求直接下单。
- 部分企业在提供智能算力调度的基础上，会提供智能算力优化等软件的增值服务。

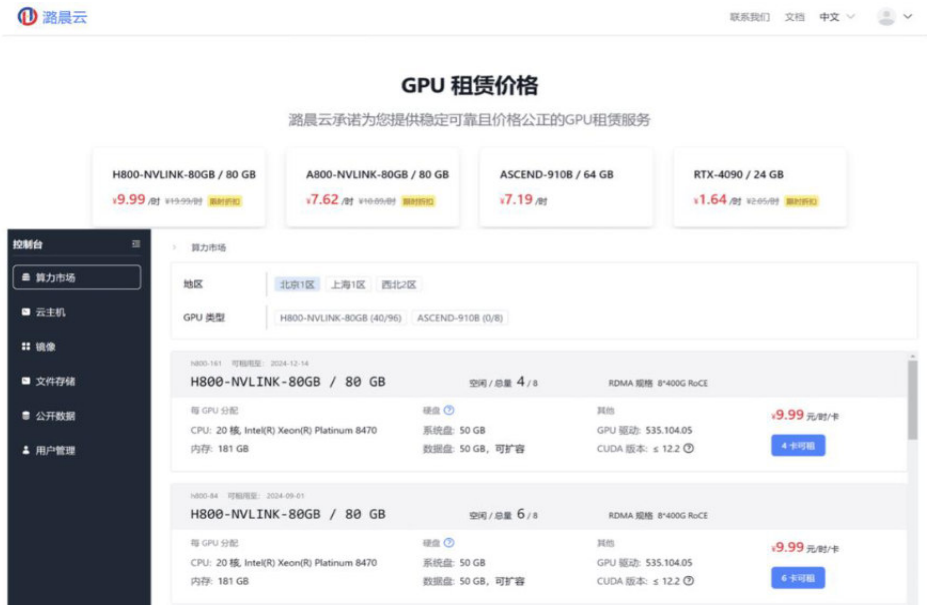
GPU算力池调度涉及的关键环节

智算应用层	金融科技	人工智能	智能制造	生物医药	智能汽车	科创研发
	智慧能源	智慧园区	智慧政务	智慧交通	智慧医疗	智慧教育
服务运营层	算力注册	算力目录	服务申请	意图感知	算力封装	服务计费	服务结算
管理编排层	资源感知	算力解构	智能推理	协同编排	智能匹配	灵活调度	统一管控
基础设施层		计算		存储		网络	
	云计算	异构计算	边缘计算	算力卸载	无服务器计算	存算一体	5G移动网络
	光传送网	SRv6/G-SRv6	新一代软件定义广域网络		确定性网络	物联感知网络	

潞晨科技 GPU算力池调度案例：潞晨科技

潞晨科技成立于2021年，是一家致力于“解放AI生产力”的全球性公司，目标是使AI大模型更低成本、方便易用、高效扩展。

- 在成立之初，潞晨科技以提供软件的形式，助力AI模型开发商优化GPU及大规模基础设施集群，提高模型的运算效率。
- 自2023年起，潞晨科技开始与智能算力资源提供商建立合作关系，并开发了潞晨云平台，依托潞晨的先进技术和运营能力，实现了对智能算力资源的有效整合和调度。通过潞晨云平台，算力用户能够灵活地管理和调度所需的智能算力资源，确保实时满足其智算需求。



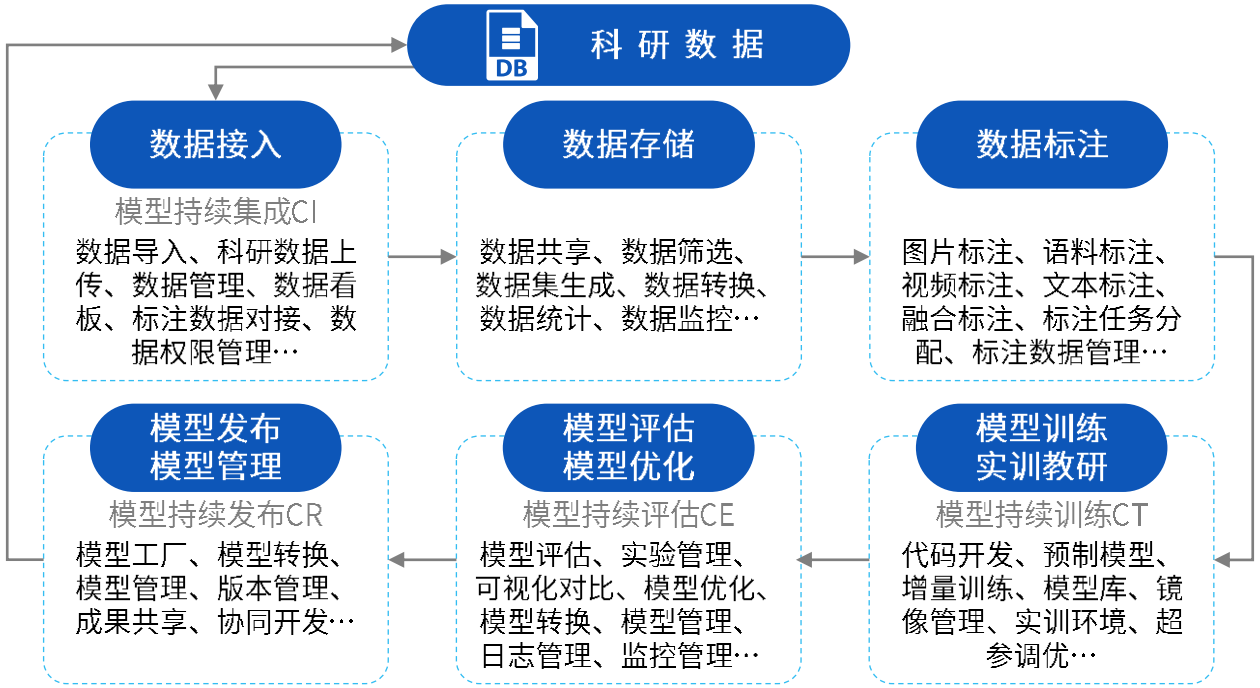
来源：《算力调度关键问题和实施路径研究》；潞晨科技官网；弗若斯特沙利文

智算租赁的商业模式—算力资源搭载硬件交付

大模型厂商是智能算力的主要需求方，为了自身业务能够有稳定的算力支撑，部分企业向上游产业链进行横向拓展，通过自建或者共建智算中心，或者与上游智算产业核心设备厂商合作，以交付硬件的方式获取本地化部署的算力资源

- 2023年7月，华为与合作伙伴联合发布昇腾AI 大模型训推一体化解决方案，主要为解决大模型面对大模型研发周期长、部署难度高、业务安全性要求等问题，加速大模型在各行业应用落地。
- 大模型推训一体机结合了昇腾AI的基础软硬件能力、合作伙伴的大模型资源以及一体化平台的优势，为行业客户提供了“即插即用”的大模型解决方案。
- 主要优势在于支持私有化部署，满足了对企业业务保密性有高度要求的企业需求。

案例：科研推训一体机功能



推训一体机案例：科大讯飞

科大讯飞（002230.SZ）成立于1999年，专注于智能语音、计算机视觉、自然语言处理、认知智能等人工智能核心技术研究。

- 2023年5月，科大讯飞正式对外发布讯飞星火大模型，讯飞星火大模型具备文本生成、语言理解、知识问答、逻辑推理、数学能力、代码能力、多模态能力等七大核心能力。2023年10月，科大讯飞与华为联合发布了国内首个全国产算力平台“飞星一号”。
- 2023年8月科大讯飞宣布联合华为推出讯飞星火一体机。星火一体机提供了从底层算力、AI框架、训练算法、推理能力到应用成效的全栈AI能力，还具备大模型预训练、多模态理解与生成、多任务学习和迁移等能力，为企业打造专属大模型提供了强有力的支撑和持续的创新动力。

来源：华鲲振宇官网；科大讯飞官网；弗若斯特沙利文

智算租赁市场参与者

智算租赁市场处于起步阶段，产业链上中下游不同类型的参与者纷纷入局，依托自身原有业务在产品，资源，供应链，客户以及渠道方面的优势，在智算产业谋篇布局

参与者类型	商业模式	产品矩阵	智算业务设立逻辑	智算业务优势
ICT硬件集成商	根据自身产品提供硬件或整体解决方案服务，结合智算硬件设备，打造大型智能算力平台	CPU&GPU服务器、数据基础设施建设、网络安全、存储、云计算服务、智能算力租赁等	初期以承建政府或国企的智算集群建设为主，业务逐步扩展至帮助已经建设或即将建设的智算集群提升利用率，缩短投资资金回收周期	掌握智算产业核心设备，或者有基础设施的搭建能力
电信运营商	以移动、联通、电信三大运营商为代表，在提供网络数据服务等基建的基础上，逐步部署智算硬件设备和智算中心	智算中心、数据中心、数据和算力调度、网络设施、网络安全等	主要承接全国智算中心项目的建设或改造，在建设后为提升利用率缩短投资资金回收周期，与其他智能算力供应商合作出租闲置算力	作为国家数字基金的主要投资者和建设者，资金实力雄厚，可触达资源丰富
第三方数据中心服务商（IDC）	通过自建或租用标准化机房、带宽等电信资源，为客户提供IDC基础服务的基础上，进行AIDC升级改造	机房和基础设施建设服务、智算中心运维服务、云服务等	主要与ICT硬件集成商、电信运营商或互联网大厂等合作共建智算集群项目，协助采购及建设基础设施；项目建成后由IDC进行后续运维，协助出租智算资源	具备数据中心的搭建能力，大规模硬件集群部署和配套运维服务经验
AI Infra厂商	链接智能算力基础设施和智能算力应用之间的技术服务商，为模型的训练、服务器的部署和运行提供支撑，确保算力的使用体验	AI模型推理引擎、模型训练加速、算力池管理、智能算力租赁等以软件和平台为主的产品	原先主要从事AI基础软件的开发，目前业务拓展至设立平台链接智能算力供给方和下游应用企业，为模型的训练和推理、服务器的部署和运行提供支持	人工智能和智算方面技术能力突出，能够优化客户的智算资源使用体验
云服务厂商	基于原有云服务平台，进一步部署智算业务，自建大模型。云服务厂商可直接出租自有智能算力或在此基础上以MaaS的形式交付算力	计算服务、存储、数据库服务、网络服务、安全服务、AI算力服务、大数据分析等	在原先完整的云服务体系基础上，结合GPU等硬件设施，以增值服务产品或一站式的智算解决方案，满足下游智算需求方的业务需求	全栈式的技术服务能力，完善的产品矩阵以及较高的用户粘性和品牌效应
人工智能应用企业	自身需要大量智能算力，向上拓展智算资源，自建智能算力设施并搭建计算平台。在满足自身模型训练需求的基础上，提供智能算力租赁和其他相关AI服务	算力运维、数据服务、算法补足、服务器租赁、数据存储、网络等	本身需要大量的智能算力，部分企业会自行部署智算中心，在保证自身智能算力需求的同时，将闲置算力出租	智能算力的深度使用者，具备较强AI技术和丰富的行业经验
跨界企业	企业本身从事其他行业业务，投资购买GPU服务器，部署智算相关基础设施和硬件设备	GPU芯片等智能算力资源	资金充足，希望寻找第二成长曲线，因此投资智算产业，并将其智算资源出租给其他智能算力运营商或下游大型智能算力应用企业	在渠道，产品，资金，供应链等方面有差异化的优势

来源：弗若斯特沙利文

主要产品形态：硬件/基础设施　软件

智算租赁盈利模型拆解（1/2）

假设自行购入GPU服务器，托管在数据中心内，然后将此智算资源租出，则成本端需要考虑设备的折旧，托管成本和智算平台的运营成本，收入端参考云服务厂商的定价

成本端，参考数据中心运行成本

- 数据中心的运行成本包括一次性资产购置与建设成本以及日常运营成本。
- 为保证数据中心总体目标的实现，需要进行全周期的优化，日常运营成本的影响更重要。

折旧摊销
(非现金成本)

- 折旧摊销的成本占比一般最高，且占比随配置级别提升。
- 据测算，单台A100服务器价格145万元，交换机约20万元，其他零配件约15万元，整套设备5年摊销对应月折旧成本3万元。

能耗费用
IT设备
制冷系统
供配电系统
照明等

- 能耗费用在智能算力租赁产业中的成本占比可达30%-50%以上。
- 据测算，1台8卡A100服务器每月电费约为6000元。

运营维护
带宽使用
技术服务
维修维保
云服务

- 简单运维的成本占比约为5%，提供云服务的运维成本占比约20%。

其他费用
人工费用
资产租赁
其他

- 其他各类费用的成本占比不足10%。

收入端，参考云服务厂商租赁价格

- 云服务厂商倾向于与客户形成长期稳定的租赁关系，因此以包年与包月形式出租的卡型更多，该租赁模式对于有大量训练需求的用户而言性价比更高。
- 小颗粒度的订单（单卡按小时租赁）可以更好地匹配不同用户的需求，对预算有限且需要少量高性能计算资源的用户吸引力明显，符合智能算力租赁平台出租模式。

部分云服务厂商GPU计算型云服务器售价对表（截至2024.8.15）

按小时计费，折算为元/小时/单卡								
GPU型号	显存	卡传输	阿里云	百度云	火山引擎	腾讯云	天翼云	均值
V100	32GB	NVLink	19.7	15.3	-	-	15.4	16.8
A100	40GB	PCIe	-	-	-	-	21.3	21.3
A100	40GB	NVLink	-	26.1	-	19.3	-	22.7
A100	80GB	NVLink	-	-	40.4	-	-	40.4
按包年包月计费，折算为万元/月/8卡服务器								
GPU型号	显存	卡传输	阿里云	百度云	火山引擎	腾讯云	天翼云	均值
V100	32GB	NVLink	7.6	5.5	5.4	4.8	5.9	5.8
A100	40GB	PCIe	-	-	-	-	8.2	8.2
A100	40GB	NVLink		10.0	-	11.2	-	10.6
A100	80GB	NVLink	-	11.6	17.1	-	-	14.4
按年订阅折扣			8.5折	8.3折	8.3折	-	8.5折	-

来源：弗若斯特沙利文；阿里云官网；百度云官网；火山引擎官网；腾讯云官网；天翼云官网

智算租赁盈利模型拆解（2/2）

自购GPU服务器的重资产模式下，设备的投资回收周期约为16.2个月，而轻资产模式下搭建智算平台，提供算力的调度、匹配和供需撮合服务，期初成本显著低于重资产模式，但技术增值服务成为影响平台利润的核心关键

重资产模式下智算租赁平台投资回收周期测算

- 重资产运行模式下，智能算力租赁平台可以布局数据中心建设，购置GPU，参与到算力池化、算力匹配、算力调度等多个智能算力管理与运营阶段中，最终通过完善的GPU云服务盈利。由于目前市面上H系列服务器暂无公开用例，因此测算主体为1台8卡A100-40G-NVLink服务器。
- 假设1：设备购置成本方面，1台A100服务器价格为145万元，配套交换机约20万元，其他零配件约15万元，整台设备总价格为180万元。
- 假设2：每台设备按照5年进行摊销计算非现金成本，无残值。

每台设备每年非现金成本 = 180 ÷ 5 = 36 万元

- 假设3：日常运营成本方面，每年每台设备的非现金成本与现金成本在总成本中的占比比例为60%：40%。

每台设备每年现金成本 = 36 ÷ 60% × 40% = 24 万元

- 假设4：收入方面，基于云服务厂商平均22.7元/小时/单卡的租赁市价，将1台8卡服务器的租赁价格定为180元/小时。
- 假设5：每台设备每年可全功率运行。

每台设备每年租赁收入 = 180 ÷ 10000 × 24 × 365 = 157.7 万元

$$\text{GPU设备投资回收周期} = \frac{\text{设备购置成本}}{\text{年租赁收入}-\text{年现金成本}}$$

$$= \frac{180}{157.7-24} = 1.35\text{ 年} = 16.2\text{ 月}$$

轻资产模式下智算租赁平台盈利能力分析

- 轻资产运行模式下，智能算力租赁平台的业务可与能够提供GPU服务器与相对完善的算力增值服务的第三方数据中心厂商深度绑定。
- 根据中贝通信2023年9月公告，中贝通信以12万元/P/年的价格，向青海联通提供智能算力服务。中贝通信负责提供H800算力服务器、相应的IB网络交换机和配套光模块、线缆与管理平台，以及提供维保服务。
- 根据中贝通信2023年11月公告，中贝通信以18万元/P/年的价格，向北京中科新远科技有限公司提供智能算力服务。该合同价格包含算力租用价格、技术服务与支持有关费用。
- 假设1：成本方面，参考中贝通信的定价，以及智能算力供给受限后定价上涨的趋势，智能算力租赁平台在自身系统内部署智能算力的价格为20万元/P/年。
- 假设2：收入方面，参考180元/台/小时的定价，根据A100单卡在FP16精度下312TFLOPs的算力，按P数折算1台8卡设备提供的GPU云服务的等值收入。

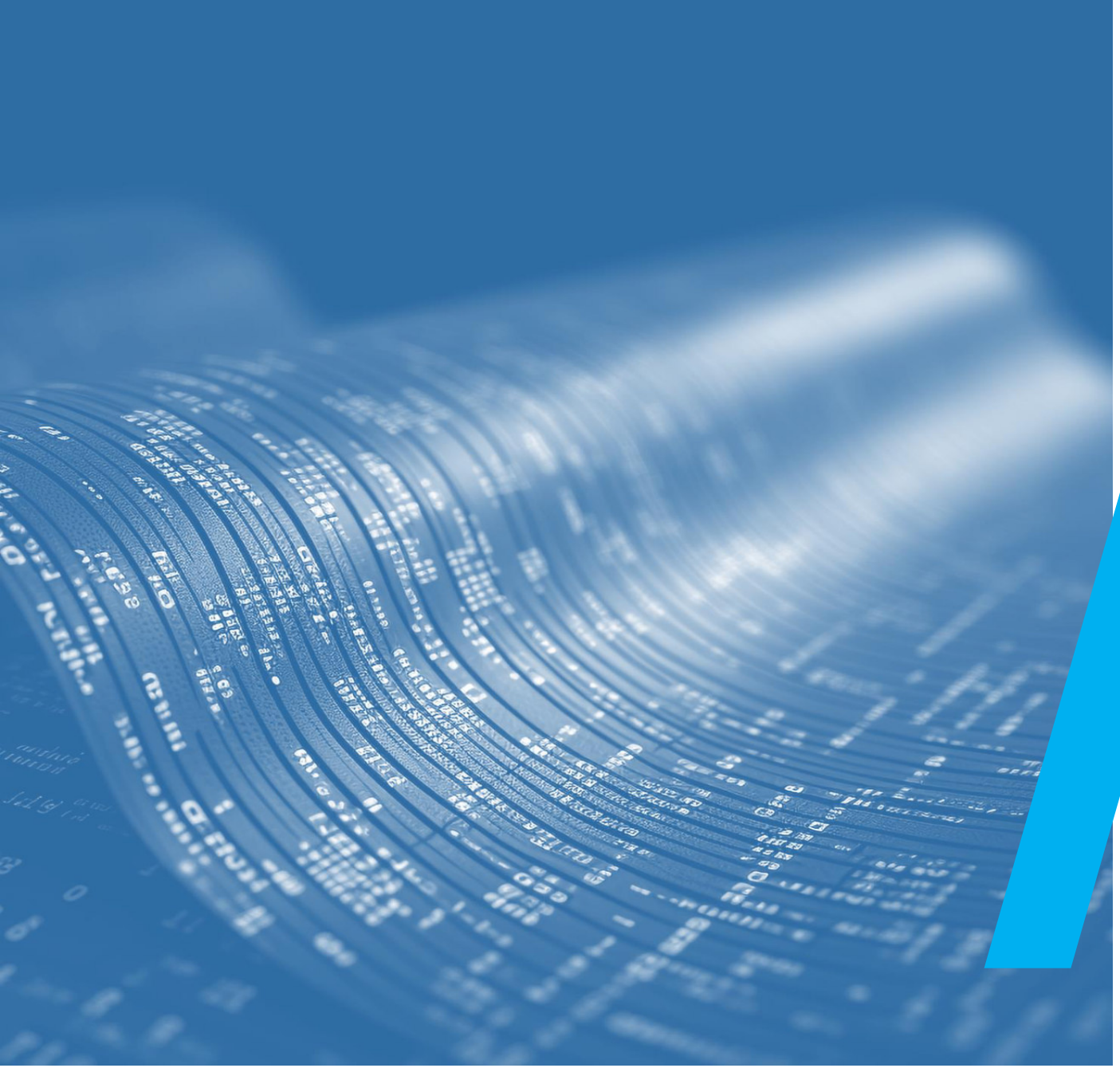
每台设备P数 = 312 × 8 ÷ 1000 = 2.5P

每台设备每P租赁收入 = 157.7 ÷ 2.5 = 63.1 万元/P/年

技术服务创造的额外价值 = 63.1 – 20 = 43.1 万元/P/年

- 轻资产运行模式下，租赁平台无需考虑能耗费用以及运维费用等现金成本，可以将更多资产投入在GPU云服务的附加值创造上，进一步提升平台的盈利能力。

来源：弗若斯特沙利文



4

- 附录

来源：弗若斯特沙利文

注：此版为第一版产业图谱，未来将根据市场变化持续更新。

天翌智算 | FROST & SULLIVAN 沙利文

方法论

- 弗若斯特沙利文布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。
- 弗若斯特沙利文依托中国活跃的经济环境，从中国智能算力领域着手，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，弗若斯特沙利文的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- 弗若斯特沙利文融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在弗若斯特沙利文的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- 弗若斯特沙利文密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- 弗若斯特沙利文秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

- 本白皮书著作权归弗若斯特沙利文所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得弗若斯特沙利文同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“弗若斯特沙利文”，且不得对本白皮书进行任何有悖原意的引用、删节或修改。

- 弗若斯特沙利文作为独立的第三方调研机构，在调研工作过程中，遵守了中国有关法律法规，坚持独立、客观、公正的原则。本白皮书中的信息包含的更改或更新，不应解释为承诺或保证，不应该以此白皮书的任何部分作为任何合同或承诺的基础。

- 弗若斯特沙利文会尽力确保本白皮书的信息的准确性及来源可靠性，但不针对展现信息的准确性、可靠性或完整性做出任何保证或声明。弗若斯特沙利文对本白皮书所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。所有市场价格、数据和其他信息的完整性或准确性，都是基于选定的公共市场数据，反映现行情况，以及当下的看法。在不同时期，弗若斯特沙利文可发出与本白皮书所载资料、意见及推测不一致的报告和文章。白皮书中包含的内容本身不具备投资决策使用，弗若斯特沙利文否认对任何直接或间接依赖白皮书中包含的任何信息、任何错误、遗漏或不准确或由此导致的任何行动而造成的直接或间接损失或损害承担责任。投资者应对本文件中讨论的主题进行独立的尽职调查，并在做出任何投资决策之前对相关市场进行独立的判断。

- 本文件中所载的信息可能包括或通过参考前瞻性假设与预测，其中将包括任何非历史事实的假设。对此类前瞻性假设的准确性不作任何保证。本文件中所载的任何预测和估计都是基于以某些假设为基础的推测性判断。这些前瞻性假设可能是错误的，并可能受到不准确的假设或已知或未知的风险、不确定性和其他因素的影响，其中大多数是无法控制的，可能导致实际结果与预测信息有重大差异。

- 本白皮书仅供客户使用。除非事先获得书面许可，否则弗若斯特沙利文对任何其他用途或拥有本白皮书的任何其他人士不承担任何义务或责任，且不会提供进一步意见、作证或出庭或在任何其他法律程序中出庭，并保留追究未经授权人士由此造成的任何损失的权利。