中国平安 PINGAN
专业·价值

证券研究报告

AI系列专题报告（一）
算力：算力基建景气度高，国产AI芯片发展势头良好

电子行业 强于大市（维持）

平安证券研究所 电子信息团队

分析师： 陈福栋 S1060523070003（证券投资咨询）

分析师： 闫磊 S1060519100002（证券投资咨询）

2025年6月12日

请务必阅读正文后免责条款

核心摘要

- **AIGC蓬勃发展，对底层智能算力产生强劲需求。**行业前期，训练是算力需求的主力，大量大模型训练需要海量算力支撑。2024年末，DeepSeek重磅发布，其轻量化、低成本、高性能特征大幅拉低了AI应用门槛，有望成为各类推理场景爆发的契机，推理算力市场需求潜力巨大。在此背景下，全球科技巨头资本支出维持在高位，国内智算中心建设如火如荼。根据中国信通院&浪潮信息报告数据，截至2023年底，全球智能算力规模为335EFLOPS，同比增长136%，智能算力需求旺盛，增速远超算力整体规模增速。

- **AI算力芯片：ASIC蒸蒸日上，关注AI芯片国产化。**根据中投产业研究院数据，2022年，全球AI芯片市场规模约422亿美元，预计2025年将达到920亿美元，三年CAGR达27.7%。AI算力芯片主要有通用GPU和ASIC两大类，通用GPU是目前的主流，算力强大、生态丰富，英伟达是领导者，AMD紧随其后，加速追赶，国内海光信息、沐熙等也采用该路线；ASIC是专用定制芯片，面对专项任务，计算能力和效率较通用GPU更强，博通和MARVELL是领先者，拥有大量特定任务负载的大型云服务厂商多与其合作开发，典型案例是谷歌TPU系列产品（与博通合作研发），国内华为昇腾、燧原等采用AISC路线。2025年5月19日，英伟达发布NVLink Fusion，允许第三方ASIC芯片接入英伟达计算体系，AI算力基础设施异构融合计算壁垒得到一定缓解，客观上有望推动ASIC芯片的发展。国内来看，AI芯片是美国对华科技制裁的重灾区，其先进的AI芯片产品无法出口至国内，倒逼国内AI算力芯片实现从设计到制造的全产业链国产替代，海光DCU、华为昇腾、寒武纪等已取得一定突破，燧原、沐熙、天数、壁仞等公司也在快速发展，AI算力芯片自主可控已取得长足进步。

- **AI服务器：市场景气度高，国产AI芯片占比提升。**根据IDC&浪潮信息报告数据，2024-2028年，我国AI服务器CAGR将达到约30.6%，景气度持续高企，同时，在美国对华半导体出口管制进一步升级、我国AI服务器厂商积极拥抱国产AI芯片等多重因素影响下，国内AI服务器市场中，国产AI芯片占比将持续提高。此外，随着DeepSeek火爆出圈，仅数月时间，国内市场已有100多家厂商推出AI（DeepSeek）一体机，供应端呈现“百机大战”格局，需求端已在政务、金融、医疗、教育、物流等行业多点开花，DeepSeek大模型一体机未来有望持续向好。

- **投资建议：**DeepSeek火爆出圈，轻量化、低成本、高性能，推理场景逐渐打开，推理端算力需求将逐步超过训练端。当前，AI算力是美国对华科技制裁的重灾区，先进的AI算力芯片无法出口至国内，反向倒逼国内AI算力从设计到制造到整机的全面国产替代，国内通用GPU、ASIC芯片蓬勃发展，同时服务器和一体机厂商也逐渐向国产AI芯片倾斜，全产业链合力，国内AI算力自主可控已取得不菲成果。推荐海光信息、浪潮信息、龙芯中科、芯原股份、紫光股份、深信服、神州数码，建议关注寒武纪、华勤技术、软通动力。

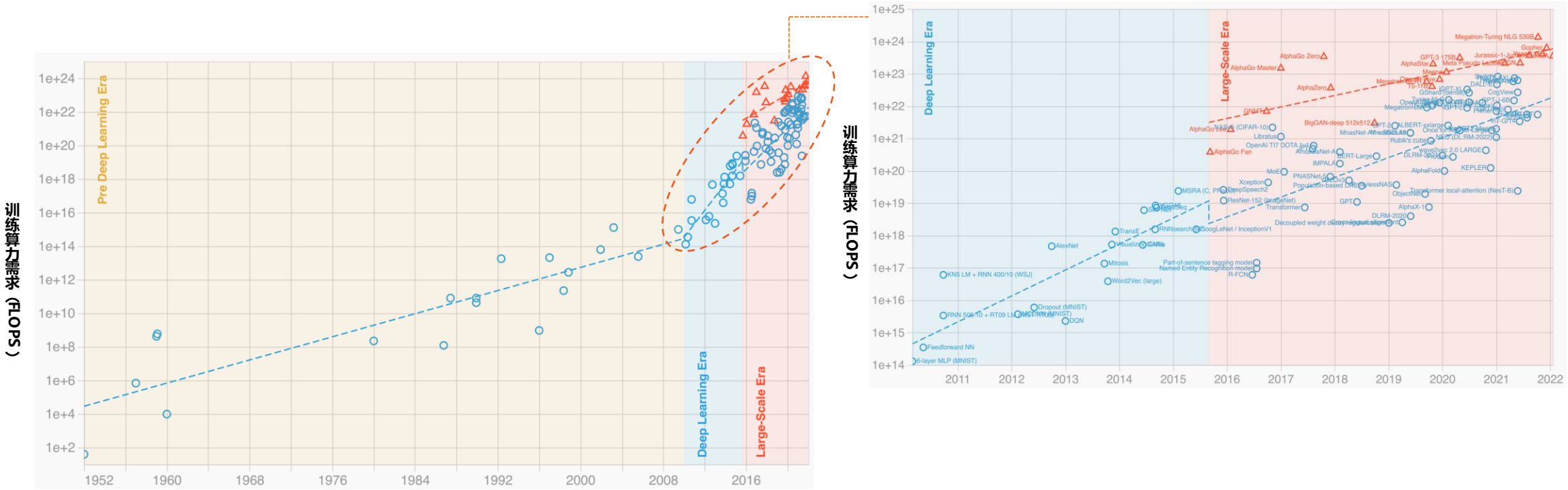
- **风险提示：**（1）大模型应用落地不及预期的风险。（2）国产AI芯片开发不及预期的风险。（3）美国对华科技制裁的风险。

目录CONTENTS

1.1 大模型发展助推智能算力需求加速释放

- **大模型和生成式AI快速发展，拉动智能算力需求加速释放。**2015-2016年，大模型时代开启，整体训练计算量较之前时期大2-3个数量级。2022年底，ChatGPT横空出世，随后，拥有千亿甚至万亿级参数的各类通用大模型相继发布，其训练迭代极大拉动了对智能算力的需求。根据Jaime Sevilla等人的研究，2016-2022年，大模型训练所需的算力从4e+21增长到8e+23 FLOPS，意味着大模型训练算力每10个月即翻一倍。

里程碑级机器学习系统的算力变化趋势

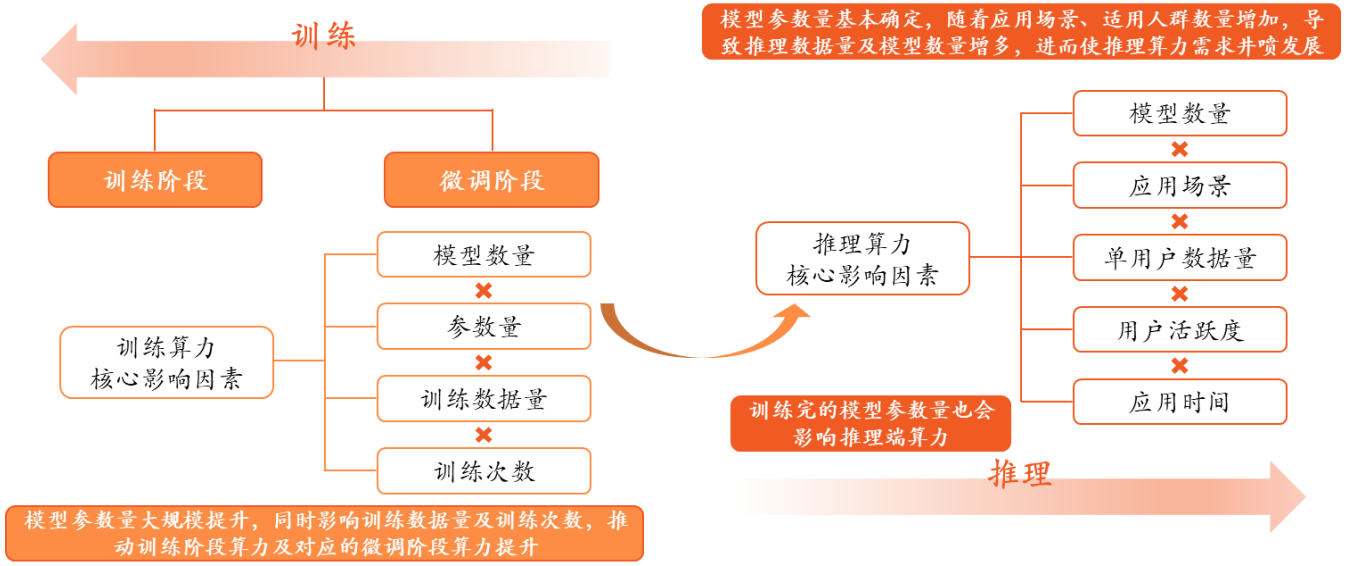


数据来源：Jaime Sevilla等《Compute Trends Across Three Eras of Machine Learning》,平安证券研究所

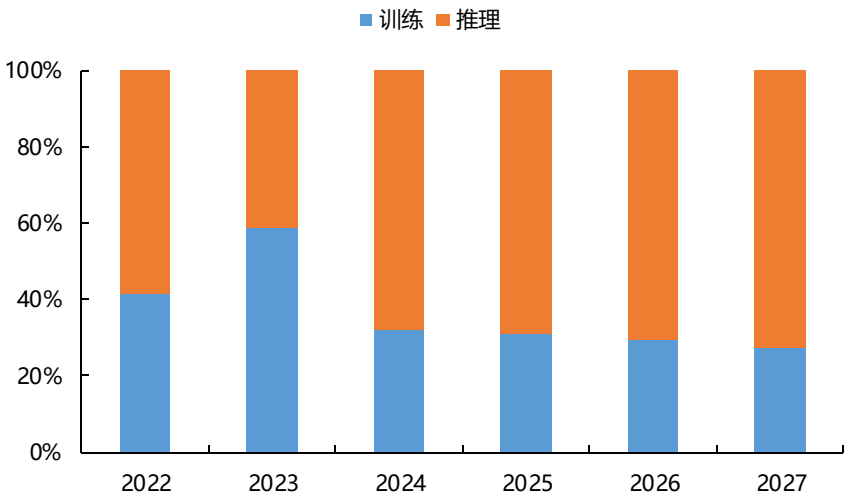
1.2 训练端算力需求阶段性旺盛，未来推理端需求可能远超训练端

- AIGC通常需要训练和推理两个环节，训练是通过数据开发出AI大模型，因此参数量的升级对算力需求影响较大，推理是利用训练好的模型进行计算，推理部署的算力主要在于应用场景的日数据吞吐量。
- **随着大模型逐渐成熟，推理成本逐渐下降，未来推理端算力需求可能远超训练端。**大模型训练很大程度上是阶段性需求，训练数据通常是相对固定的，比如几万亿或几十万亿量级，在发展早期是主要的算力需求端。根据IDC&浪潮信息报告数据，2023年，由于各类大模型层出不穷，当年中国AI服务器工作负载中，训练端算力占比为58.7%。随着大模型逐渐成熟以及推理成本逐渐下降，大模型应用场景将逐渐打开，对推理算力的需求将持续扩大，预计到2027年，推理端算力需求占比将大幅增长到72.6%。长期看，推理算力的需求潜力远超训练端。

训练+推理算力需求



中国人工智能服务器工作负载预测，2022-2027

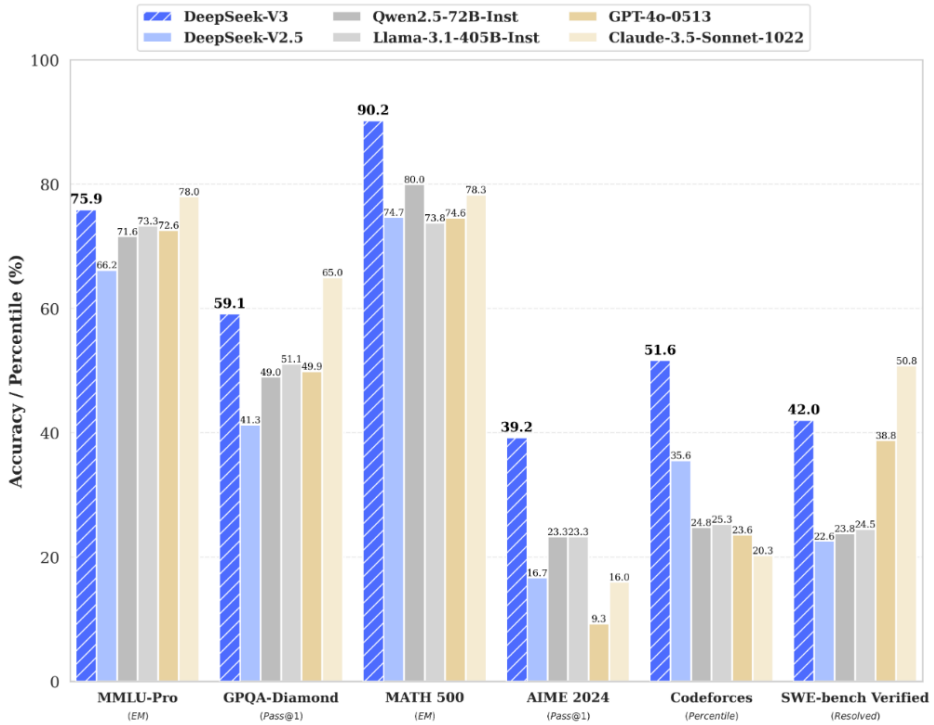


数据来源：甲子光年公众号，IDC、浪潮信息《2023-2024中国人工智能计算力发展评估报告》，平安证券研究所

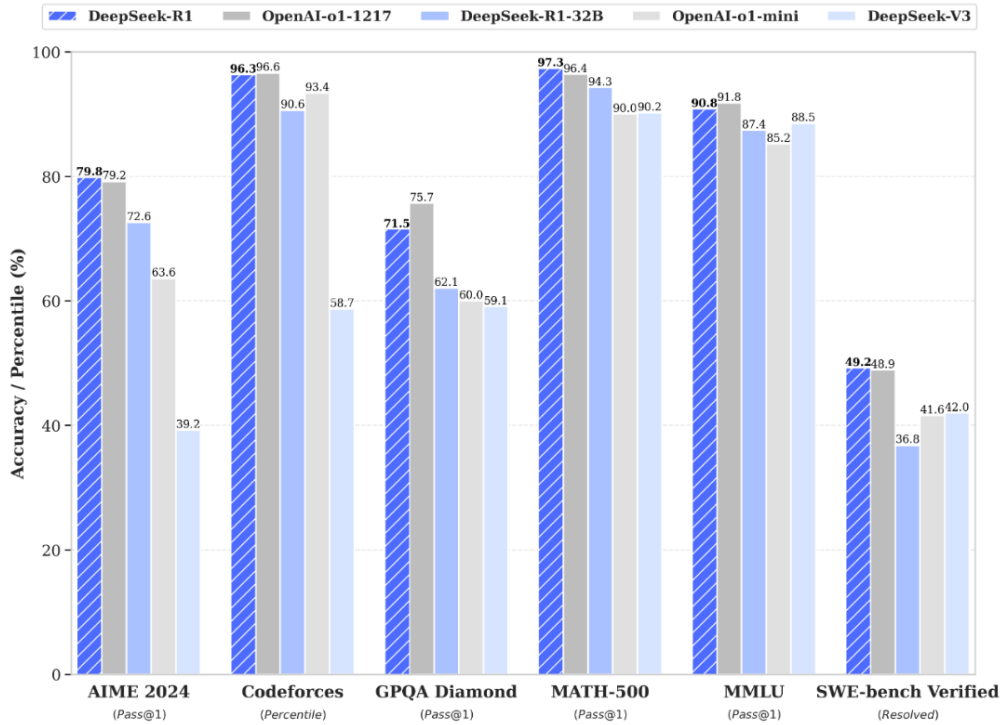
1.3 DeepSeek高性能、低成本大模型有望成为推理场景爆发的契机

- **2024年12月26日，DeepSeek-V3发布，671B参数，性能与GPT-4o不分伯仲。**DeepSeek-V3采用MLA和MoE架构，支持使用FP8混合精度训练，引入了一种无辅助损失的负载平衡策略，并设置了多Token预测训练目标，注重轻量化的同时，性能上与GPT-4o不分伯仲。
- **2025年1月20日，DeepSeek-R1发布，性能对齐OpenAI-o1正式版。**DeepSeek-R1在后训练阶段大规模使用了强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力，在数学、代码、自然语言处理等任务上性能对标OpenAI-O1。

DeepSeek-V3与其他大模型测评成绩对比



DeepSeek-R1与其他大模型测评成绩对比

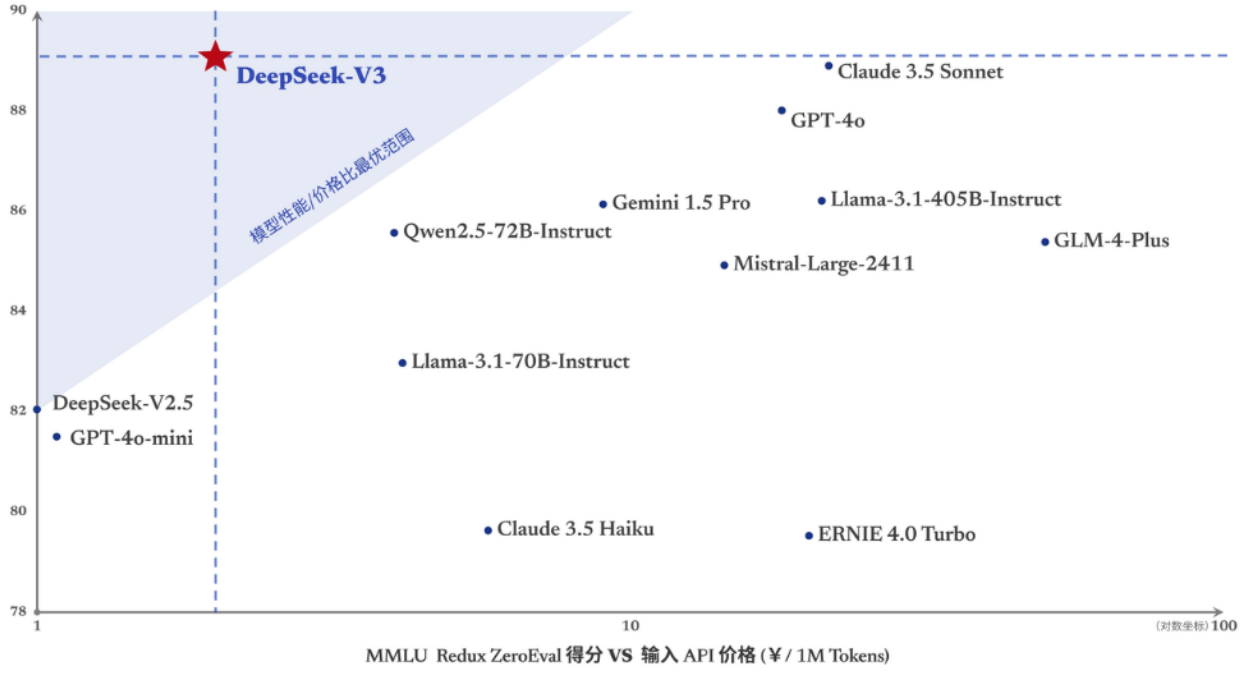


数据来源：DeepSeek公众号，平安证券研究所

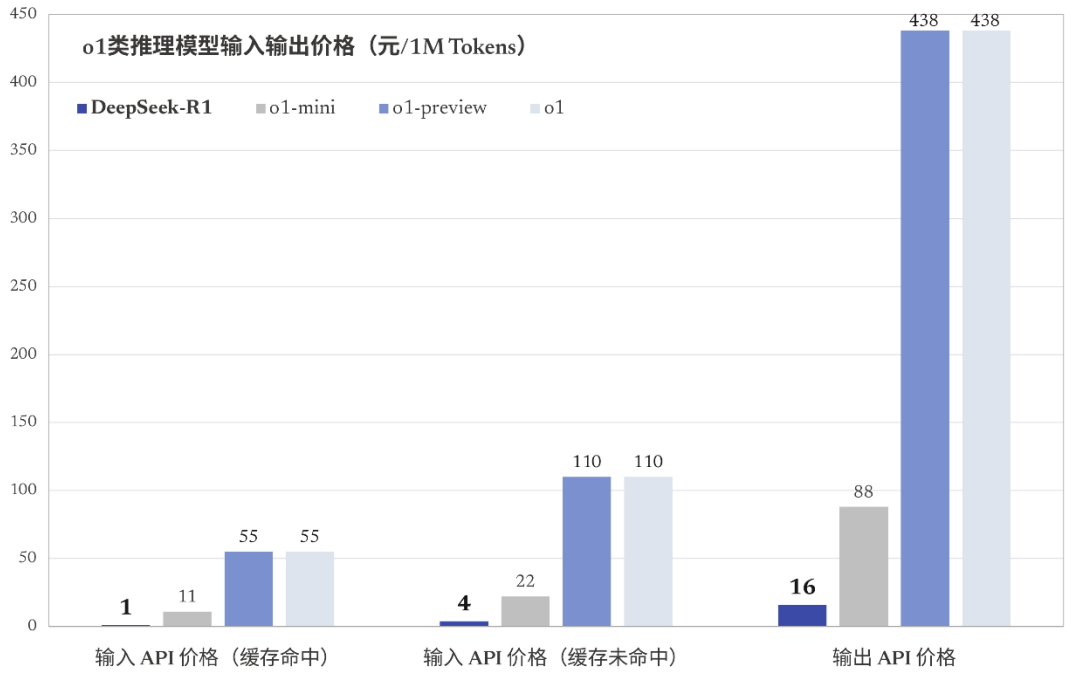
1.3 DeepSeek高性能、低成本大模型有望成为推理场景爆发的契机

- **DeepSeek设计轻量化，训练成本大幅降低。**DeepSeek-V3使用2048块H800 GPU集群训练，训练时长278.8万小时，假设H800 GPU每小时租金2美元，总训练成本仅557.6万美元。
- 受益于低成本训练优势，DeepSeek-V3和R1的API服务定价较GPT-4o和o1大幅下降。DeepSeek-R1 API服务价格为每百万输出Tokens 16元，而o1价格为438元，成本优势明显，大模型推理准入门槛大幅降低。

DeepSeek-V3与其他大模型性能/价格对比



DeepSeek-R1与其他大模型输入/输出价格对比

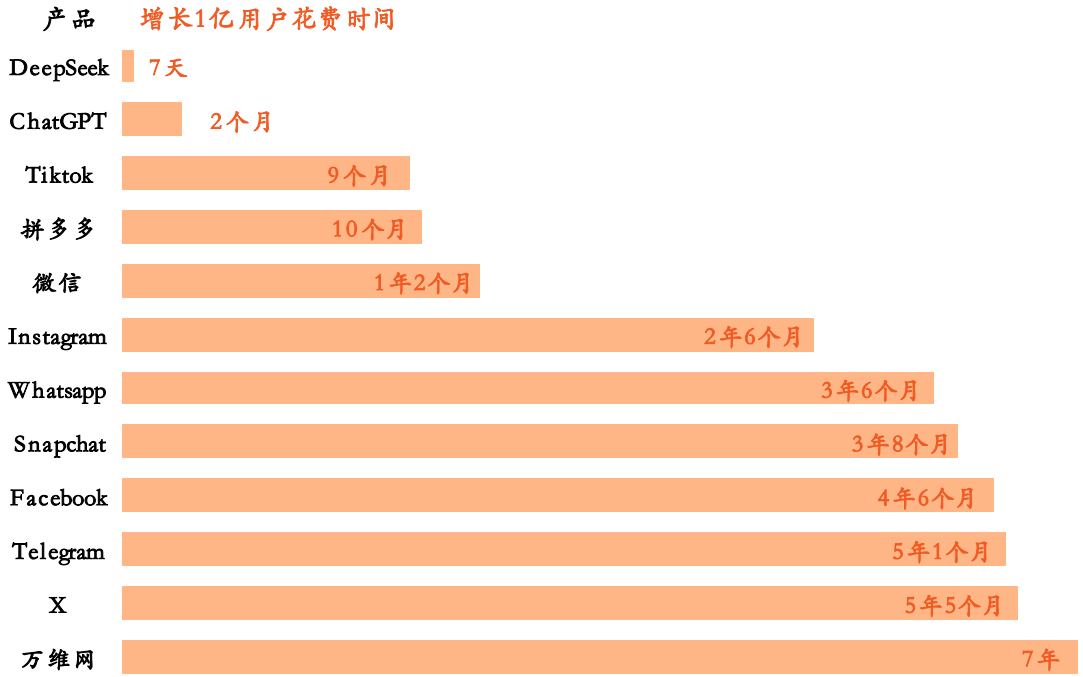


数据来源：DeepSeek公众号，平安证券研究所

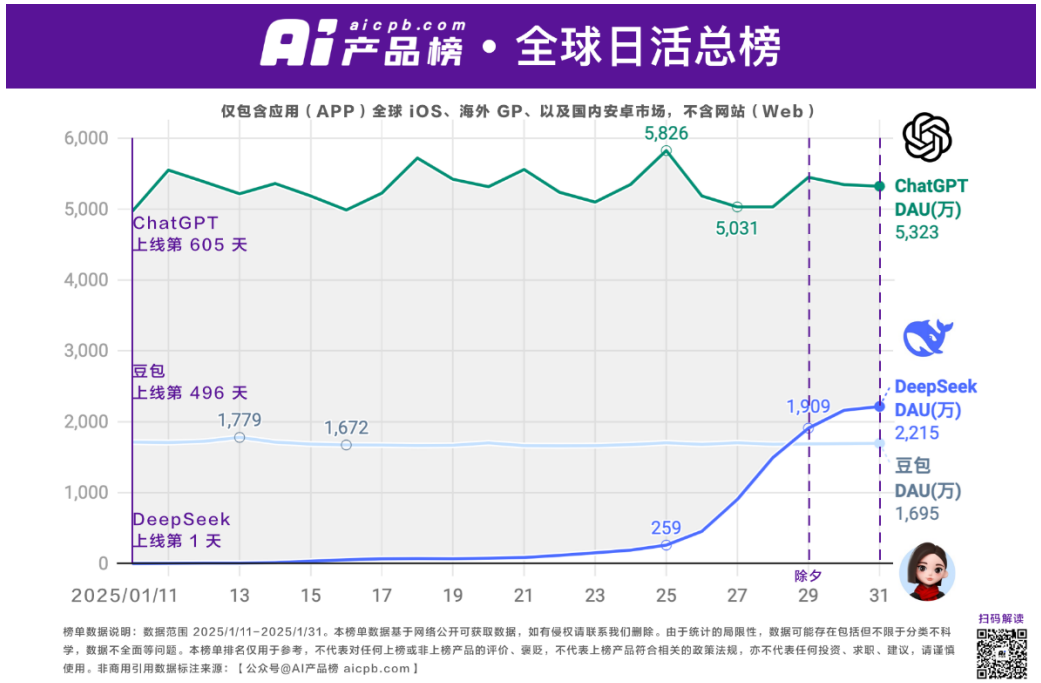
1.3 DeepSeek高性能、低成本大模型有望成为推理场景爆发的契机

- **DeepSeek凭借其低成本、高性能的推理优势，始一推出便赢得市场青睐。**根据AI产品榜数据，DeepSeek仅用7天就完成了1亿用户的增长，且DeepSeek APP发布20天时，全球日活DAU突破2000万，达到ChatGPT的40%。
- **DeepSeek爆火反映出市场对低成本推理大模型的迫切需求，且其开源属性将大力推动各类推理应用场景的爆发，未来推理对算力的需求权重将进一步增长。**

DeepSeek增长1亿用户仅花费7天



DeepSeek APP上线20日日活达到ChatGPT的40%



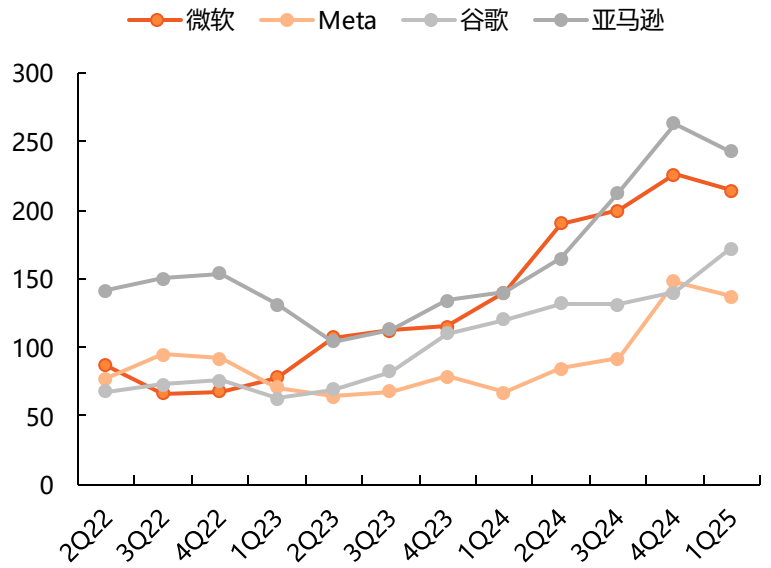
备注：DeepSeek包含网站Web/应用APP累加不去重，Tiktok不包含国内版抖音

数据来源：AI产品榜公众号，平安证券研究所

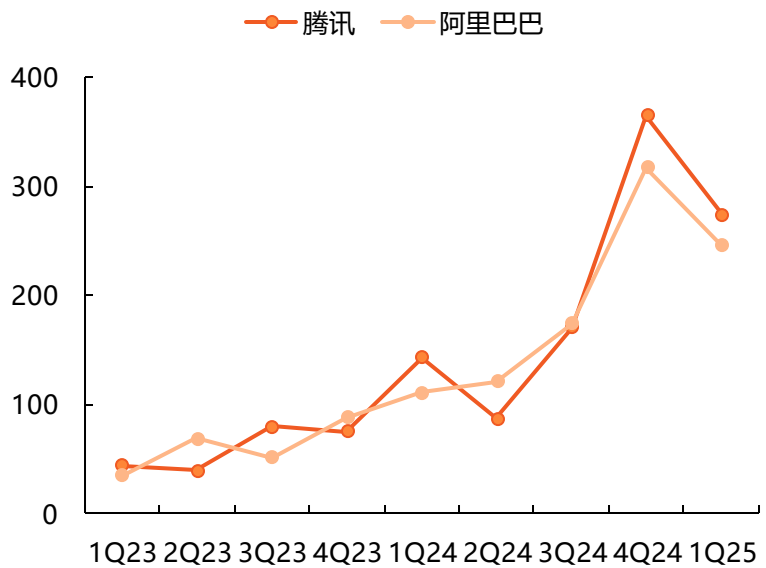
1.4 科技巨头资本支出维持在高位，加速布局万卡集群

➢ **大模型训练和推理带来海量算力需求，科技巨头资本支出维持在较高水平。**根据各公司公告，2025Q1，微软（214亿美元）、Meta（137亿美元）、谷歌（172亿美元）、亚马逊（243亿美元）四巨头资本开支合计高达766亿美元，同比增长64.0%，国内腾讯（275亿元）、阿里巴巴（246亿元）资本开支合计521亿元，同比大幅增长104.2%。此外，科技巨头持续投入算力采购，以H100的采购为例，根据Omdia Research信息，2023年，Meta和微软是H100最大的购买者，谷歌、亚马逊、甲骨文、腾讯其次。

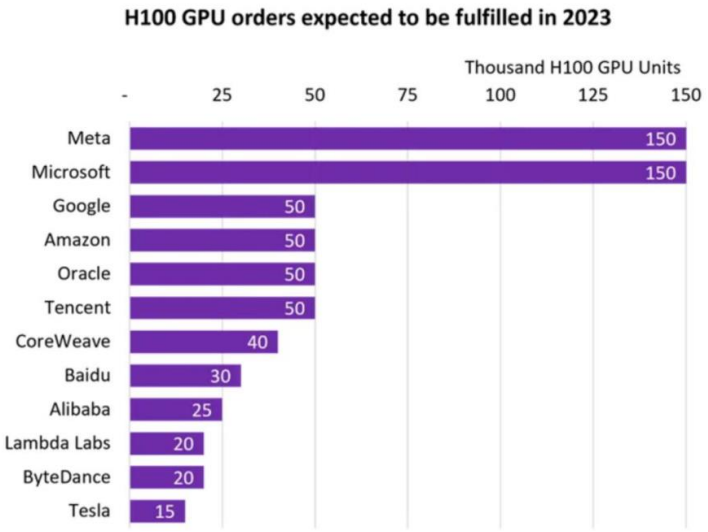
美国互联网大厂资本开支情况（亿美元）



国内互联网大厂资本开支情况（亿元）



H100客户情况@2023



数据来源：Wind，各公司公告，Omdia Research，半导体产业纵横公众号，平安证券研究所

1.4 科技巨头资本支出维持在高位，加速布局万卡集群

- **科技巨头加速布局万卡算力集群。**基于人工智能的广阔前景，全球科技巨头纷纷加大对AI基础设施布局以维持行业竞争力，其高额资本支出为万卡、十万卡集群建设奠定了基础。根据中国信通院&浪潮信息报告，国际上，Meta、微软&OpenAI等多家AI巨头陆续宣布或者完成10万卡集群建设，国内通信运营商、头部互联网、大型AI研发企业等均发力超万卡集群的布局。

全球科技巨头万卡智算集群布局情况（部分）

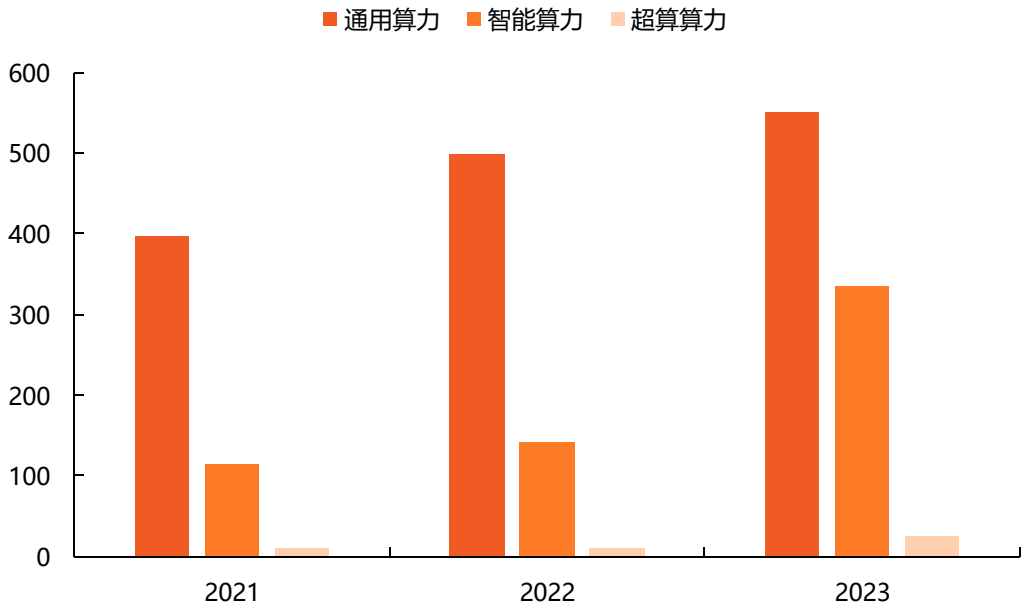
科技巨头	万卡智算集群布局情况
谷歌	2023年5月，推出AI超级计算机A3，搭载了约26000块H100GPU，为其在机器学习和深度学习研究中的应用提供强大的算力支持
Meta	2024年初，Meta建成了两个各含24576块GPU的集群
微软	早在2020年，微软便构建了一个覆盖1万块GPU的超级计算机，加速其在云计算和AI服务领域的发展
亚马逊	AmazonEC2Ultra集群采用了2万个H100TensorCoreGPU，为用户在处理大规模数据分析和机器学习任务方面提供强大算力支持
特斯拉	2023年8月，特斯拉上线集成1万块H100GPU的集群，将极大提升特斯拉在自动驾驶和车辆智能化方面的研发速度
腾讯	推出的星脉高性能网络能够支持高达10万卡GPU的超大规模计算，网络带宽高达3.2T，为未来的AI和大数据应用提供了广阔的发展空间
字节跳动	提出的MegaScale生产系统，支撑12288卡Ampere架构训练集群，为字节跳动在内容推荐、图像处理等AI应用方面提供了强大的算力保障
中国移动	计划商用哈尔滨、呼和浩特、贵阳三个万卡集群，总规模接近6万张GPU卡
中国联通	计划在上海临港国际云数据中心建成中国联通首个万卡集群，集群建成后将为中国联通在数据中心和云计算市场提供新的竞争优势

数据来源：浪潮信息、中国信通院《人工智能算力高质量发展评估体系报告》，平安证券研究所

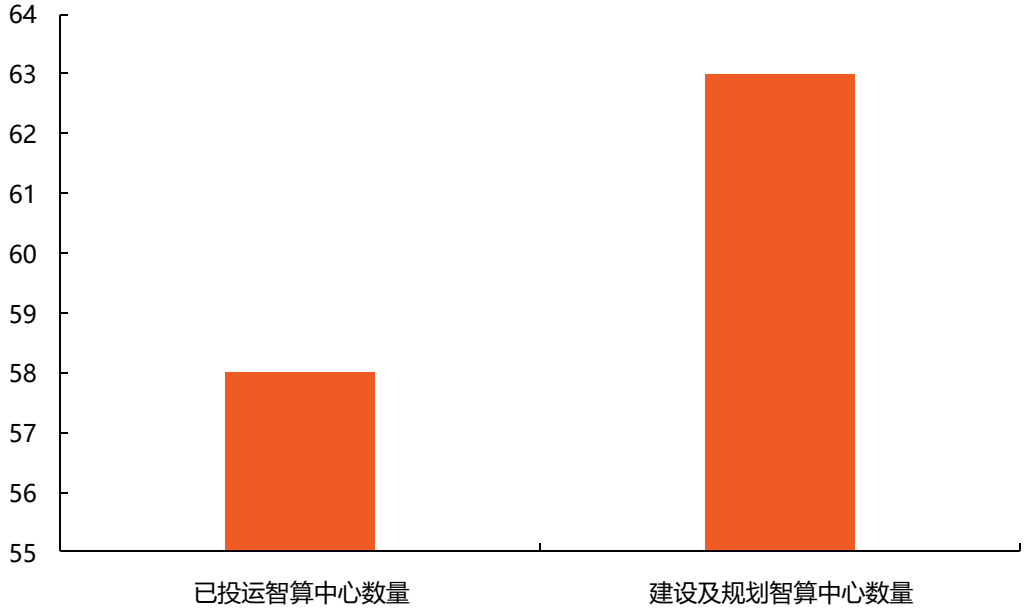
1.5 全球智能算力规模高速增长，国内智算中心建设方兴未艾

- **全球智能算力增速远高于算力整体规模增速。**根据中国信通院&浪潮信息报告数据，截至2023年底，全球算力总规模约为910EFLOPS，同比增长40%，智能算力规模为335EFLOPS，同比增长136%，智能算力增速远超算力整体规模增速。
- **国内智算中心建设如火如荼。**据第一新生研究院统计，截至2023年底，我国已投运（58家）、在建及规划（63家）的智算中心数量已超120家，其中绝大多数由地方政府及电信运营商主导规划建设。

全球算力规模（EFLOPS）



2023年全国智算中心建设情况（家数）



数据来源：浪潮信息、中国信通院《人工智能算力高质量发展评估体系报告》，第一新声研究院，大信创圈公众号，平安证券研究所

1.6 国家统筹布局，各省市加快推动地方智能算力发展

➢ **国家统筹布局，各省市积极响应，出台大量政策引导推动地方智能算力发展。**智能算力发展备受关注，智算中心建设是各省市算力布局的重点。2024年，北京、广东、河北等多地提出2025年智能算力目标。

我国算力相关支持政策汇总（部分）

发布时间	发布部委/省份	文件名称	主要内容
2024.10	国家发改委	《国家数据标准体系建设指南》	要强化基础设施互联互通、算力保障和流通利用标准建设，为数据资源、数据技术、数据流通、融合应用提供支撑。
2024.09	国务院办公厅	《国务院办公厅关于加快公共数据资源开发利用的意见》	聚焦算力网络和可信流通，支持数据基础设施企业发展。落实研发费用加计扣除、高新技术企业税收优惠等政策。
2024.03	中央人民政府	《政府工作报告》	适度超前建设数字基础设施，加快形成全国一体化算力体系，培育算力产业生态。
2023.12	国家发改委	《关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见(发改数据[2023]1779号)》	到2025年底，普惠易用、绿色安全的综合算力基础设施体系初步成型，东西部算力协同调度机制逐步完善，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。
2023.10	工信部	《算力基础设施高质量发展行动计划》	结合人工智能产业发展和业务需求，重点在西部算力枢纽及人工智能发展基础较好地区集约化开展智算中心建设，逐步合理提升智能算力占比。
2022.08	科技部、财政部	《企业技术创新能力提升行动方案(2022-2023年)》	推动国家超算中心、智能计算中心等面向企业提供低成本算力服务。
2024.06	山东	山东省算力基础设施高质量发展行动方案	强化多元算力协同部署，引导通用算力、智能算力、高性能算力中心等合理梯次布局，支持重点企业建设智算中心，适度超前提高智能算力占比。
2024.05	河北	关于进一步优化算力布局推动人工智能产业创新发展的意见	到2025年全省算力规模达到35百亿亿次/秒（EFLOPS）以上，智能算力占比达到35%左右，新增算力基础软硬件设施自主可控比例60%以上。
2024.04	北京	《北京市算力基础设施建设实施方案》	到2025年，智算供给规模达到45EFLOPS，2025-2027年根据人工智能大模型发展需要和国家相关部署进一步优化算力布局。
2024.03	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	2025年，在计算方面，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。

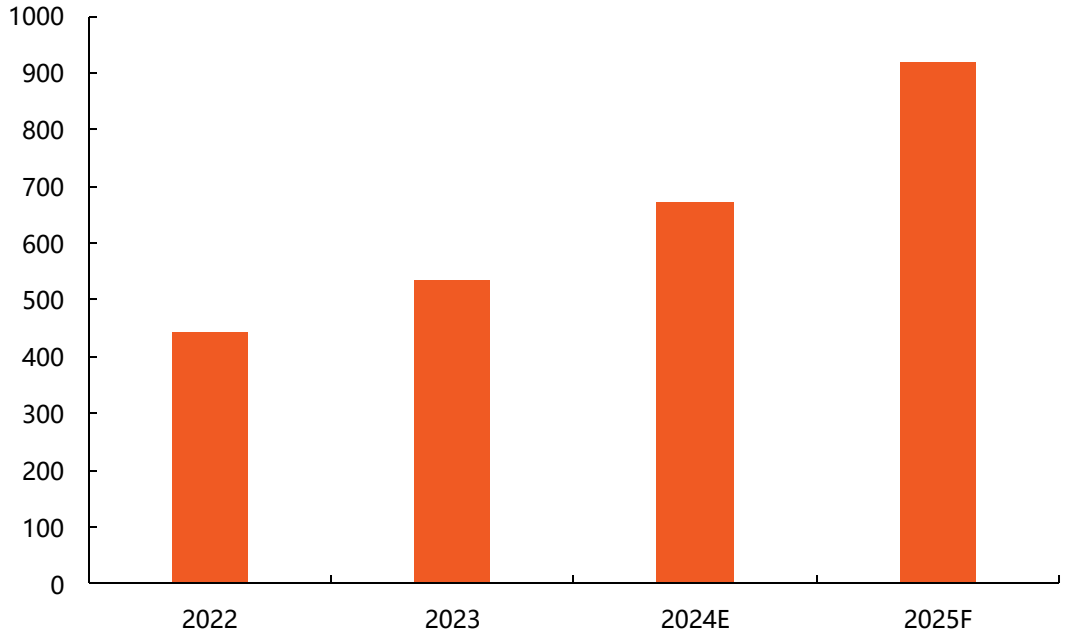
数据来源：政府官网，平安证券研究所

目录CONTENTS

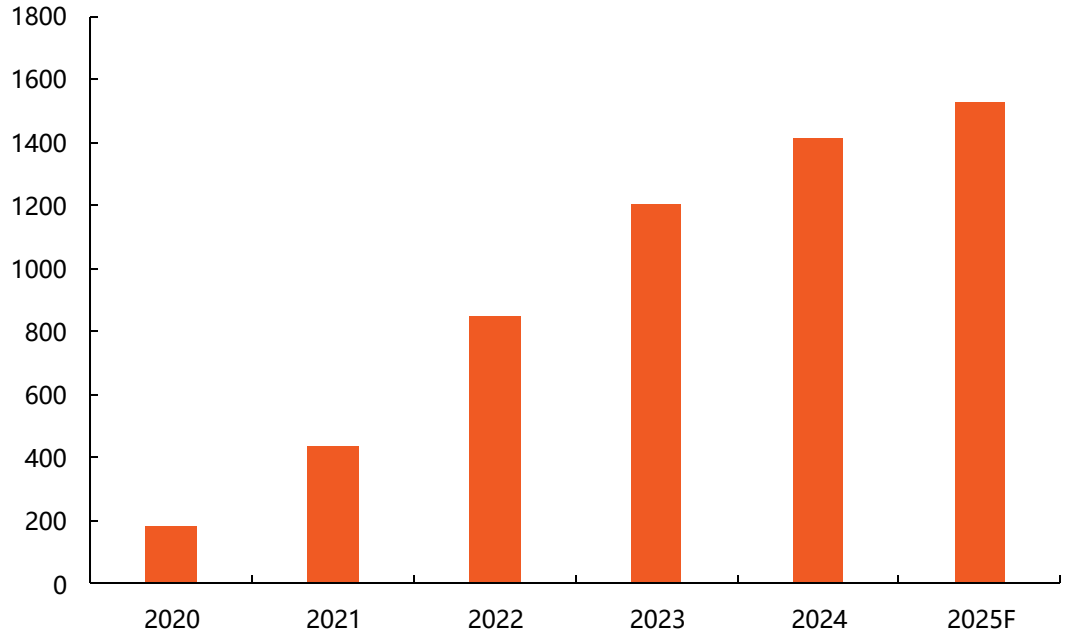
2 作为最底层的算力基座，AI算力芯片迎来发展机遇

- **AI方兴未艾，作为底层的算力基座，AI算力芯片将迎来快速发展。**全球市场，根据中投产业研究院数据，2022年全球AI芯片市场规模约442亿美元，预计2025年将达到约920亿美元，期间CAGR约27.7%；国内市场，根据中商产业研究院数据，2020年国内AI芯片市场规模约184亿元，预计2025年将增长到1530亿元，期间CAGR约52.7%。

2022-2025年全球AI芯片市场规模预测（亿美元）



2020-2025年中国AI芯片市场规模预测（亿元）

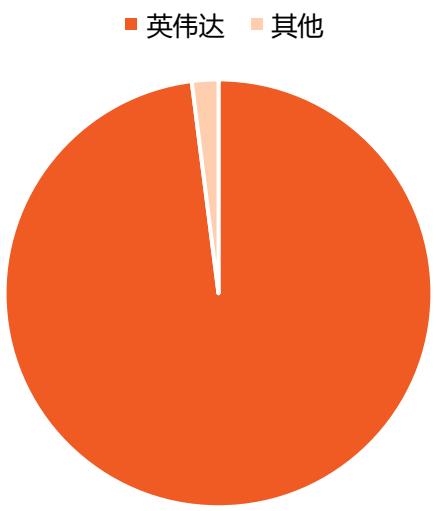


数据来源：中投产业研究院，中商产业研究院，平安证券研究所

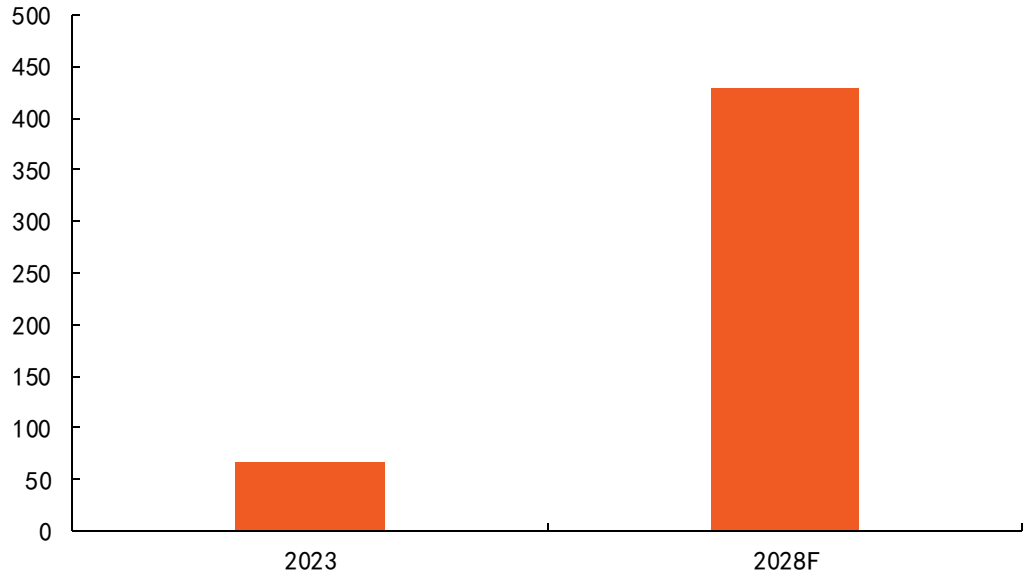
2 英伟达GPU主导AI算力芯片市场，ASIC芯片多元化发展势头迅猛

- **英伟达通用GPU执全球AI算力芯片市场之牛耳，ASIC多元化发展趋势显著。**
- AI算力芯片主要包括GPU和ASIC，GPU是通用算力卡，是目前AI算力芯片的主流选择，英伟达是该领域的领导者（数据中心GPU市占率98%@2023），AMD奋力直追，国内海光信息的DCU也属于通用GPU；ASIC是专用定制芯片，其物理设计严格匹配算法逻辑，面对专项任务，其计算性能和效率可能比通用GPU更强，博通、MARVELL是ASIC领域的领先者，大型云服务厂商多与其合作，比如谷歌的TPU便是与博通合作开发的，国内华为昇腾也属于ASIC路线。

全球数据中心GPU市场格局@2023（%）



数据中心ASIC加速芯片快速增长（亿美元）

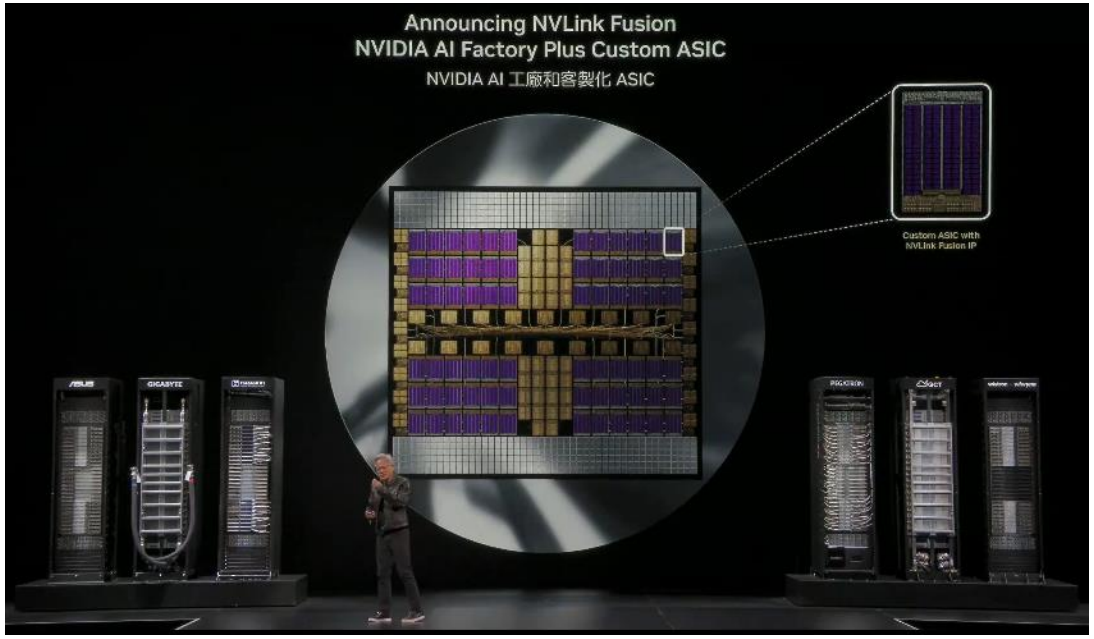


数据来源：TechInsights，北京半导体行业协会，MARVELL，平安证券研究所

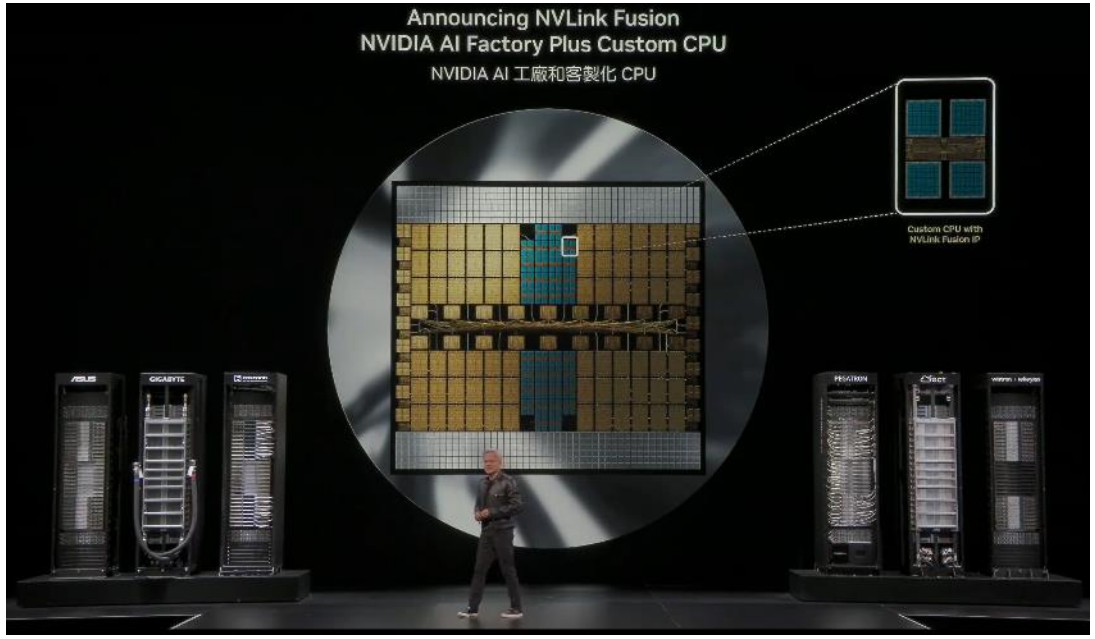
2 NVLink Fusion打破异构融合计算壁垒，ASIC芯片有望受益

- **NVLink Fusion支持第三方ASIC芯片接入英伟达计算体系，AI算力基础设施在异构融合计算领域的障碍得到一定缓解，客观上有望推动ASIC芯片的发展与繁荣。**2025年5月19日，英伟达在COMPUTEX重磅发布NVLink Fusion，通过提供IP或接口，支持第三方ASIC加速芯片、CPU等接入英伟达计算体系。 NVLink Fusion的发布意味着英伟达在保持自身架构主导地位的基础上，选择主动拥抱第三方组件，异构融合计算壁垒得到一定程度的缓解，客观上对ASIC芯片的发展将起到推动作用。

NVLink Fusion支持ASIC融入英伟达计算体系



NVLink Fusion支持第三方CPU融入英伟达计算体系



数据来源：英伟达官网，平安证券研究所

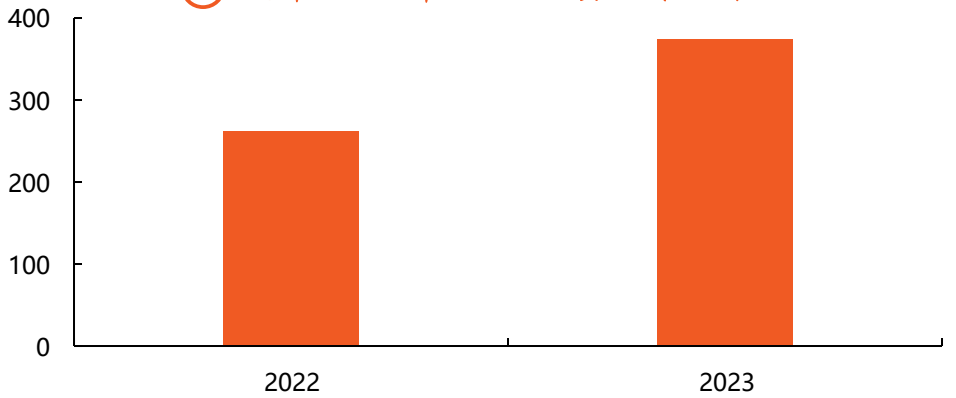
2.1 英伟达|AI算力芯片引领者，Blackwell架构芯片性能大幅提升

- **英伟达是全球GPU引领者，Blackwell架构芯片性能大幅提升。**根据TechInsights数据，2022年，英伟达数据中心GPU出货量264万块，2023年增长42.4%至376万块，是GPU市场的领导者。2024年GTC大会，英伟达推出全新Blackwell架构GPU芯片，4nm工艺，晶体管数量达2080亿个，并展示了GB200超级芯片，将两块Blackwell GPU与一块Grace CPU相连，可提供10 PFLOPS的FP16算力，性能得到大幅提升。

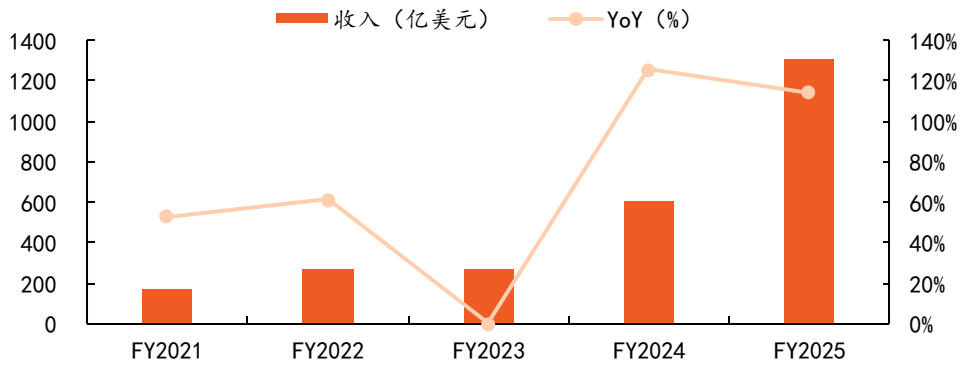
英伟达主要GPU产品性能参数对比

性能参数	H100 SXM	H200 SXM	GB200
FP16	1979 TFLOPS	1979 TFLOPS	10 PFLOPS
FP32	67 TFLOPS	67 TFLOPS	180 TFLOPS
FP64	34 TFLOPS	34 TFLOPS	90 TFLOPS
GPU 显存	80GB HBM3	141GB HBM3e	384 GB HBM3e
GPU 显存带宽	3.35TB/s	4.8TB/s	16 TB/s
最大热设计功耗（TDP）	700W	700W	2700W

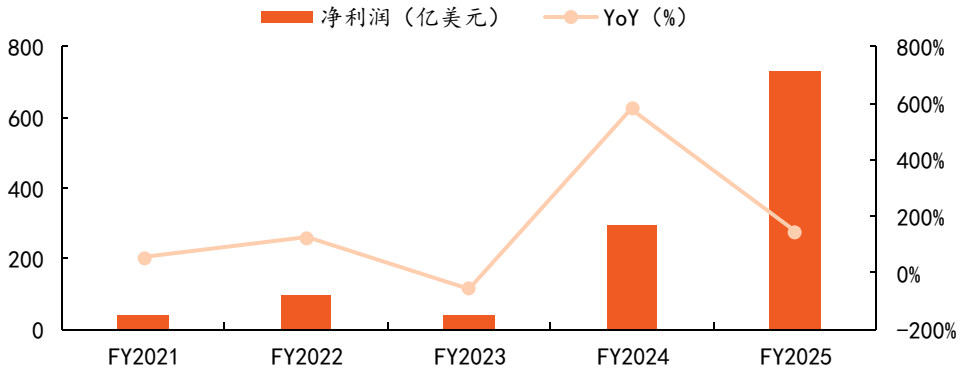
英伟达数据中心GPU出货量（万块）



英伟达近财年收入情况（亿美元）



英伟达近财年净利润情况（亿美元）



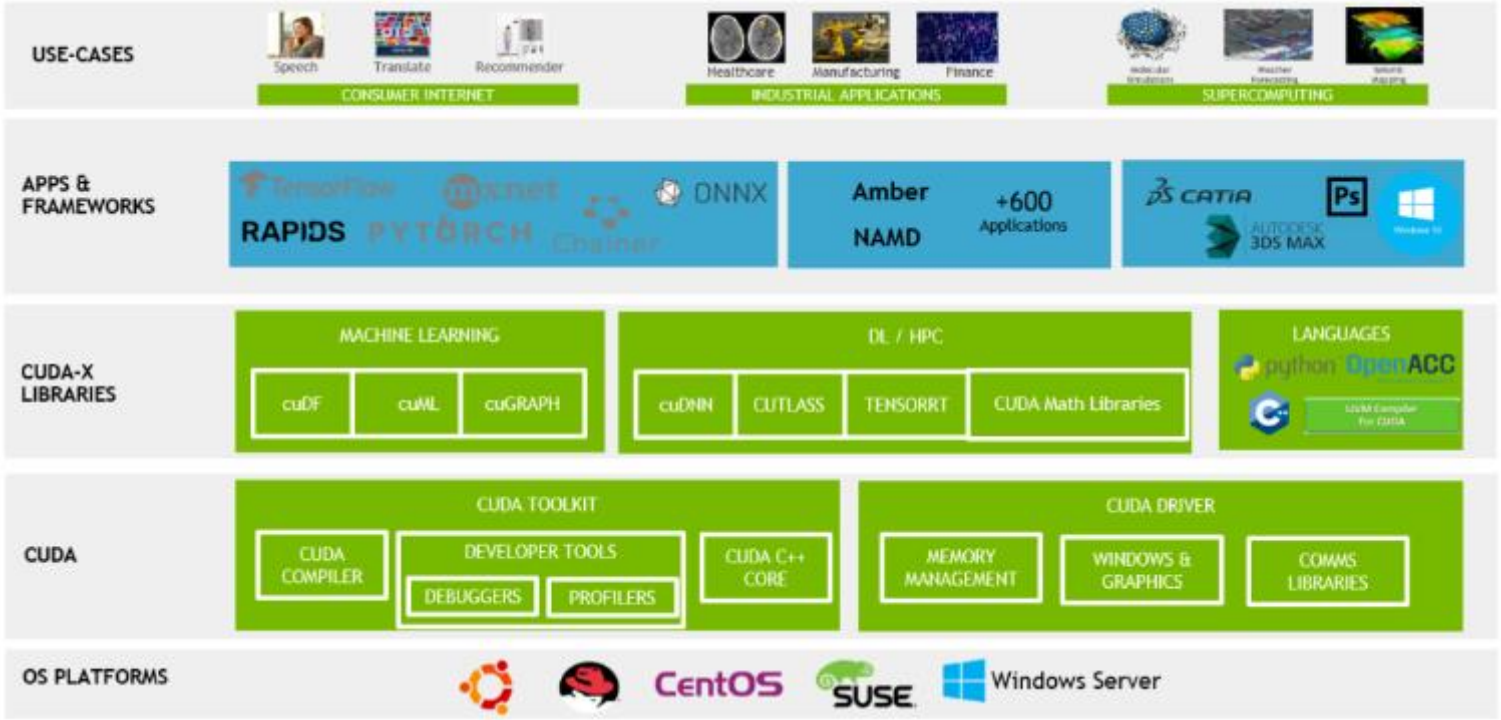
备注：英伟达按2021-2025财年统计

数据来源：英伟达官网，TechInsights，半导体产业纵横公众号，iFind，平安证券研究所

2.1 英伟达|CUDA助力英伟达构建强大的生态护城河

- **CUDA平台与GPU硬件紧密结合、协同设计，助力英伟达构建强大的生态护城河。**CUDA是英伟达为其GPU提供的并行计算平台和编程模型，开发者可利用其充分挖掘GPU的并行计算能力，并应用至通用计算领域。CUDA根植于英伟达GPU架构中，专为英伟达硬件设计，英伟达每代GPU发布时，CUDA平台通常也会同步更新，以满足新架构GPU发挥顶级效率。英伟达CUDA和GPU软硬件紧密结合，成本低、兼容性好，吸引了大量开发者参与其中，生态繁荣、用户粘性强，助力英伟达构建了强大的生态护城河壁垒。

CUDA加速计算解决方案

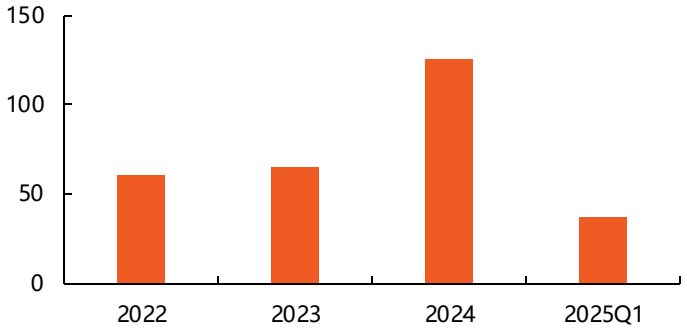


数据来源：5G创见公众号，平安证券研究所

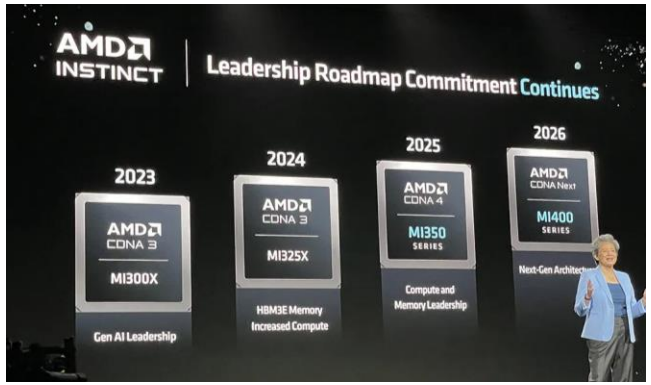
2.2 AMD|对标英伟达持续追赶，下一代MI335将与英伟达正面交锋

- **AMD是全球GPU市场的重要玩家，对标英伟达持续追赶。**2023年，AMD数据中心收入64.96亿美元，2024年增长93.64%至125.79亿美元。
- 2024年10月，AMD发布MI325X GPU，较MI300X主要增强了HBM，经AMD实测，当跑Meta Llama-2模型时，MI325X单卡在ROCm加持下训练效率超过了英伟达H200。同时，AMD宣布下一代MI350系列的首款产品“MI355X”将于2025H2推出，推理性能将有35倍提升，提供288GB的HBM3E内存，峰值算力提升1.8倍，与英伟达B200的算力持平。

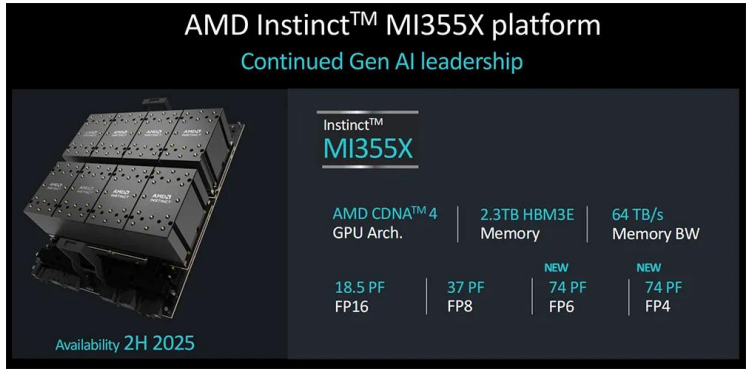
AMD近年数据中心业务收入情况（亿美元）



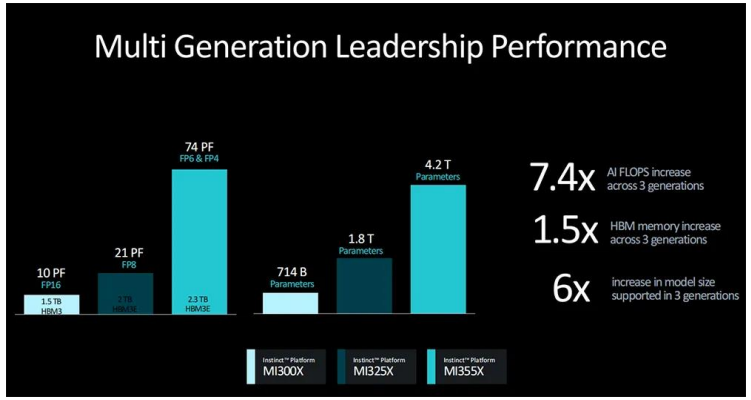
AMD AI算力芯片路线图



AMD MI355平台性能



AMD MI355芯片平台与前代性能对比

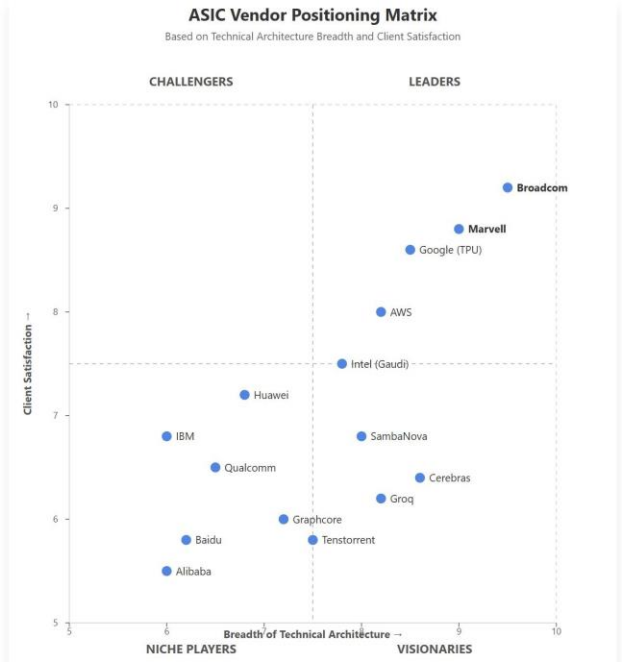


数据来源：iFind，智东西公众号，36氪公众号，平安证券研究所

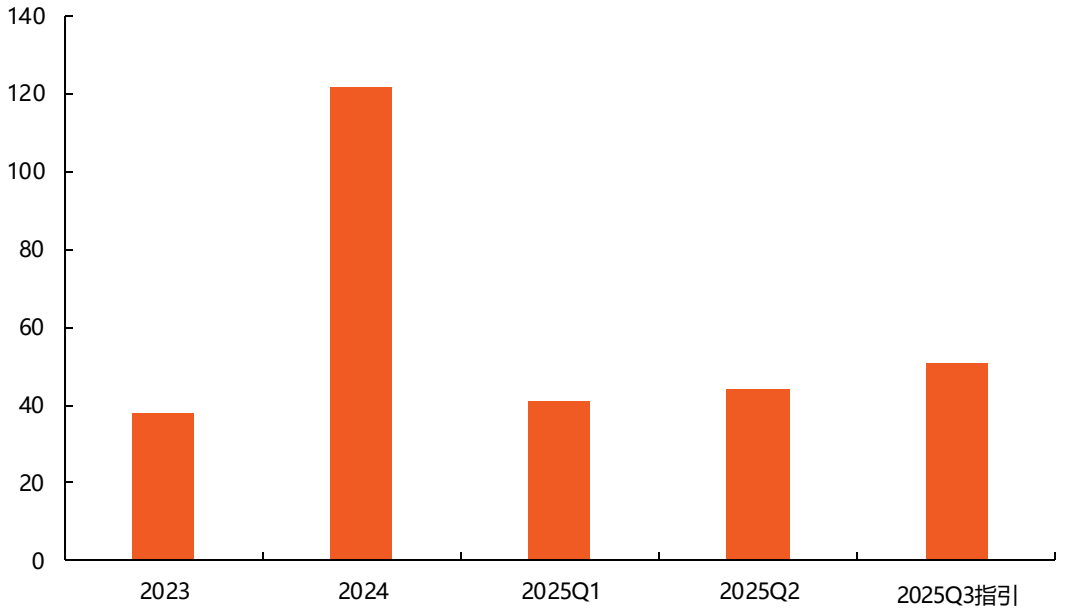
2.3 博通|大型云服务厂商定制ASIC芯片的首选合作商

- **博通在ASIC领域处于主导地位，特别是在数据中心和人工智能加速方面。**博通拥有卓越的互联设计、片上网络以及专用加速器设计能力，占据全球ASIC市场55%-60%的份额；在特定的AI工作负载上，博通ASIC解决方案效率较竞争解决方案高出40%，是拥有大量特定工作负载的大型云服务厂商的首选合作商，获得了谷歌、微软、Meta等科技巨头的青睐。
- **博通AI业务收入快速增长。**根据Wind数据，2024财年，博通AI业务收入达到122亿美元，同比大幅增长220%；2025Q1，博通AI业务收入41亿美元，同比增长77%，Q2 AI业务收入44亿美元，Q3 AI业务收入指引51亿美元，稳定增长。

ASIC厂商市场地位分布图



博通近年AI收入情况（亿美元）

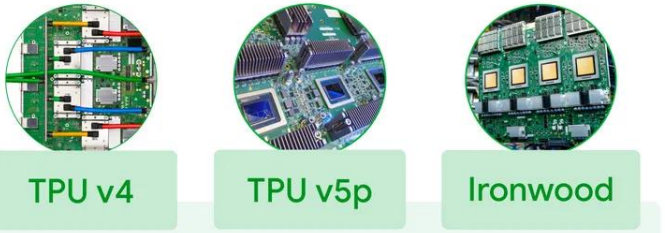


数据来源：Fourester Research，Wind，平安证券研究所

2.3 博通|合作开发谷歌第七代TPU Ironwood，对标英伟达B200

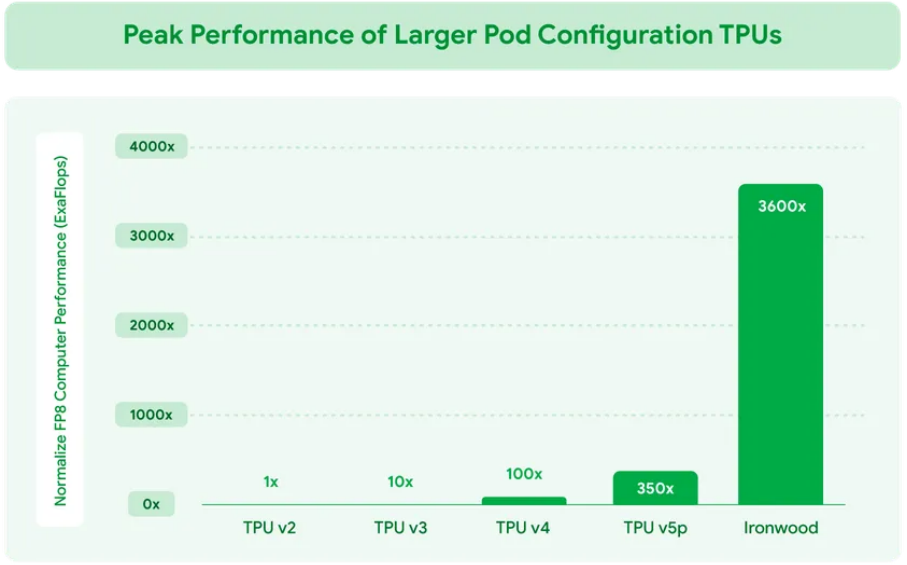
- TPU（Tensor Processing Unit）是谷歌与博通为加速机器学习任务而联合设计的ASIC芯片，采用低精度计算、脉动阵列和专用硬件设计等技术实现高效的矩阵运算加速。谷歌与博通在TPU领域合作已久，2016年便推出第一代TPU，至今已迭代至7代-Ironwood。TPU主要是谷歌自用，并不对外销售，根据TechInsights数据，2023年，谷歌自用TPU芯片量达到200万颗。
- 2025年4月，谷歌发布第七代TPU Ironwood，也是谷歌第一代专为推理设计的AI加速芯片，FP8峰值算力4614 TFLOPS，带宽192GB，整体性能直逼英伟达B200。

谷歌最新Ironwood与TPU v4、TPU v5p性能对比



	TPU v4	TPU v5p	Ironwood
	2022	2023	2025
Pod Size (chips)	4096	8960	9216
HBM Bandwidth/ Capacity	32 GB @ 1.2 TBps HBM	95 GB @ 2.8 TBps HBM	192 GB @ 7.4 TBps HBM
Peak Flops per chip	275 TFLOPS	459 TFLOPS	4614 TFLOPS

谷歌最新Ironwood FP8算力是TPU V2的3600倍



数据来源：谷歌官网，平安证券研究所

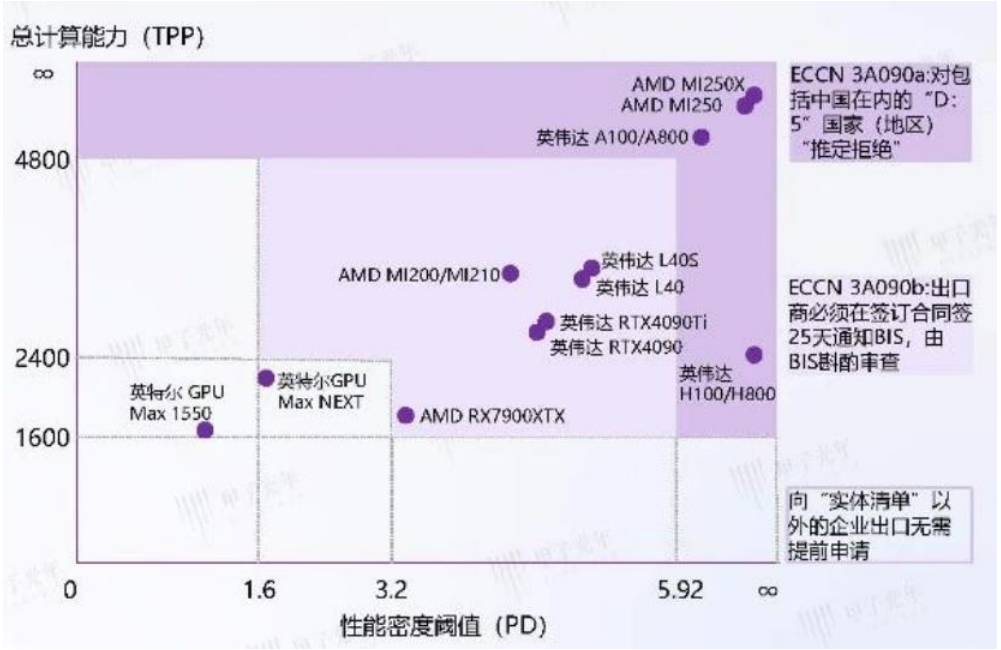
2.4 美国对中国AI算力芯片围追堵截

➢ **AI算力芯片是美国对华科技制裁的重灾区。**美国通过限制高性能AI算力芯片出口、中国先进AI算力芯片晶圆代工、中国先进制程流片所需的关键设备等方式，对国内AI算力芯片围追堵截，倒逼国内AI算力芯片从最底层的设计、制造等环节实现国产替代，自主可控主旋律长期坚挺。

美国对中国AI算力芯片的限制措施（部分）

时间	限制措施
2025.04	英伟达H20向中国出口需申请许可证
2023.10	美国商务部将壁仞科技、摩尔线程等公司列入实体名单。
2023.03	美国商务部将浪潮信息、龙芯中科等公司列入实体名单。
2022.10	BIS对中国实体超级计算机计算芯片和包含此类芯片的计算机商品的禁令，对收到许可证要求限制的外国生产项目的范围扩大到实体名单上中国境内的28家现有实体；针对<=18nm的DRAM>=128层的NAND存储芯片增加了新的许可证要求；限制美国人在没有许可证的情况下支持中国某些半导体制造设施的研发和集成电路的制造；将包括长江存储、中国科学院大学等科研院校在内的31家实体列入未经核实名单(UVL)。
2022.08	美国通知英伟达向中国和俄罗斯出口A100和H100芯片需新的许可证要求。
2022.08	BIS公告美国准备对EDA等四项技术实行出口管制。
2022.07	美国众议院通过《芯片与科学法案》，主要内容包括：分5年提供527亿美元用于半导体制造激励计划、研发投资、税收抵免，其中美国芯片基金共500亿美元，390亿美元用于鼓励半导体制造企业，110亿美元补贴芯片研发；法案要求获得补贴的半导体企业未来10年内不得在中国大陆新建或扩建先进制程的半导体工厂。
2020.10	中芯国际被纳入实体名单，对用于<=10nm技术节点的产品或技术，美国商务部采取“推定拒绝”的审批政策进行审核。

美国商务部工业和安全局对华算力芯片出口限制



数据来源：甲子光年公众号，英伟达公告，平安证券研究所

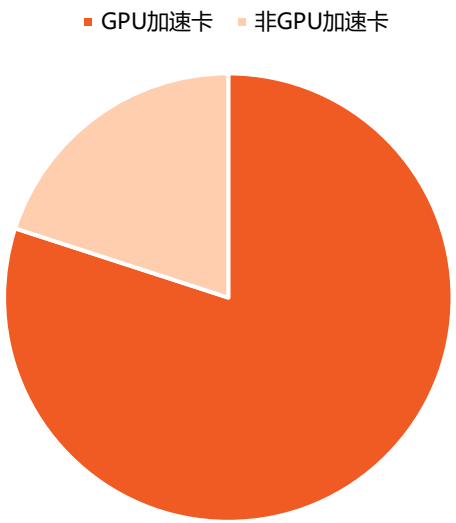
2.5 减配版H20优势削弱，国产AI算力芯片蓬勃发展

- **国产AI算力芯片奋起直追，已占据一定的市场份额。**国内AI算力芯片市场参与者主要有英伟达H20、华为昇腾系列、寒武纪思元系列、海光信息DCU系列等，为符合美国此前出口管制要求，英伟达已对其GPU产品多次减配，较国产AI算力芯片的性能优势逐渐削弱，国产AI芯片竞争力增强，已占据一定的市场份额。根据IDC数据，2024H1，中国AI加速芯片市场规模超过90万张，其中本土品牌出货量接近20万张，市占率约20%。

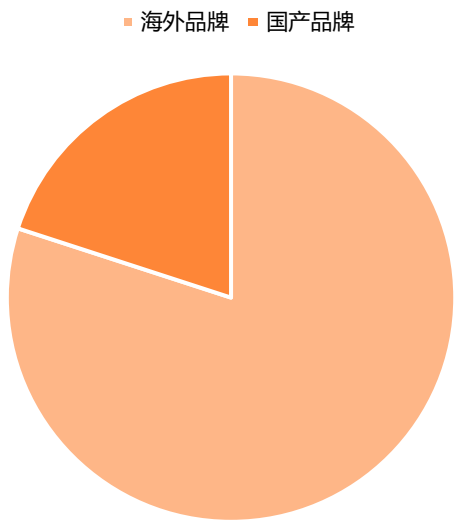
国内AI算力芯片市场产品对比

厂商	英伟达	海光信息	华为	寒武纪
芯片型号	HGX H20	深算一号	昇腾910B	思元370
INT8（TOPS）	296		-	256
BF16 | FP16（TFLOPS）	148	-	376	96
FP32（TFLOPS）	44	-	94	24
FP64（TFLOPS）	1		-	-
GPU 显存	96GB HBM3	32GB HBM2	64GB HBM2e	48GB
GPU 显存带宽	4.0TB/s	1024GB/s	-	614.4 GB/s
最大热设计功耗（TDP）	400W	350W	400W	250W

国内AI加速芯片市场份额@2024H1（出货量）



国内AI加速芯片市场结构@2024H1（出货量）

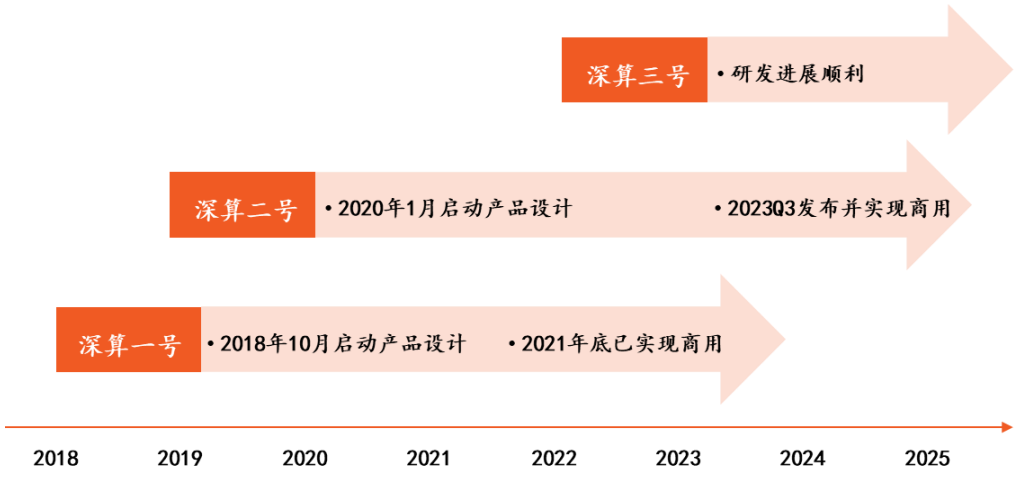


数据来源：IT之家，海光信息招股说明书，极智视界公众号，寒武纪官网，IDC，平安证券研究所

2.6 海光信息|DCU已迭代多次，可全面适配国内主流大模型

- **海光DCU是国产AI算力芯片的重要参与者，已迭代多次，发展势头良好。**海光DCU系列包括深算一号、二号、三号产品。根据公司公告，深算二号于2023Q3发布，具有全精度浮点数据和各种常见整型数据计算能力，能够充分挖掘应用的并行性，发挥其大规模并行计算的能力，性能相对于深算一号实现了翻倍增长，已在大数据处理、人工智能、商业计算等领域实现商用。
- **海光DCU深算系列属于通用GPU，兼容“类CUDA”生态，能够支持全精度模型训练。**海光DCU已实现LLaMa、GPT、Bloom、ChatGLM、悟道、紫东太初等为代表的大模型的全面应用，与国内包括文心一言、通义千问等大模型全面适配，性能达到国内领先水平。

海光信息深算系列DCU产品迭代节奏



海光信息DCU产品与行业同类可比产品参数对比

项目	海光	NVIDIA	AMD
品牌	深算一号	Ampere 100	MI100
生产工艺	7nm FinFET	7nm FinFET	7nm FinFET
核心数量	4096（64 CUs）	2560 CUDA processors 640 Tensor Core	120 CUs
显存容量	32GB HBM2	80GB HBM2e	32GB HBM2
显存位宽	4096 bit	5120 bit	4096 bit
显存频率	2.0 GHz	3.2 GHz	2.4 GHz
显存带宽	1024 GB/s	2039 GB/s	1228 GB/s
TDP	350 W	400 W	300 W
CPU to GPU互联	PCIe Gen4×16	PCIe Gen4×16	PCIe Gen4×16
GPU to GPU互联	xGMI×2，Up to 184 GB/s	NVLink，Up to 600 GB/s	Infinity Fabric×3，Up to 276 GB/s

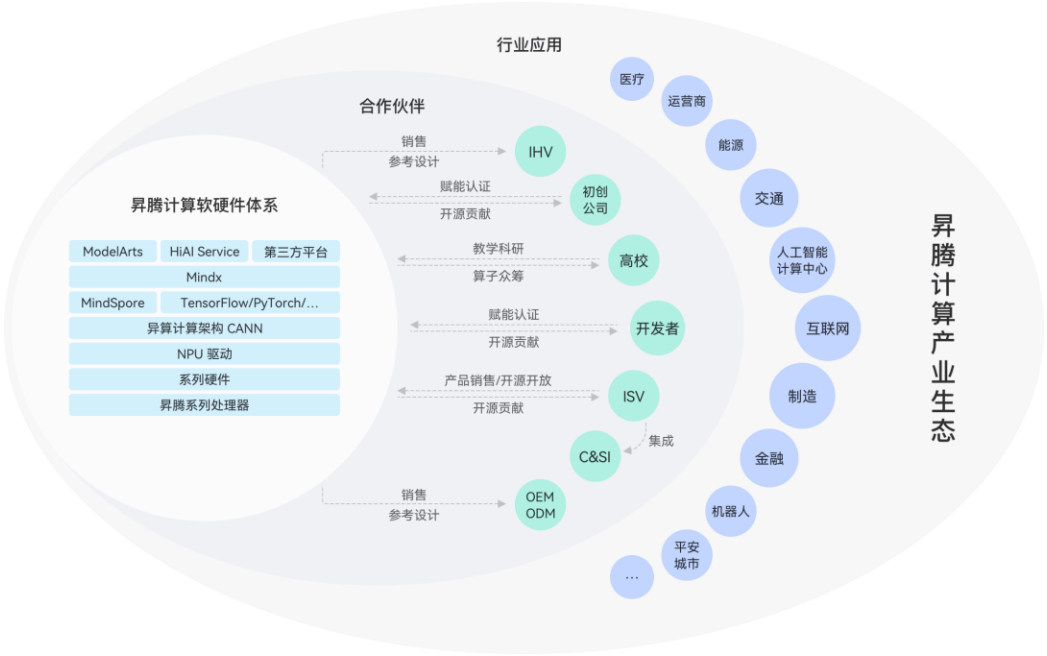
备注：数据来源于中国计量科学研究院出具的《测试报告》（报告编号：CLzn2020-01190）

数据来源：海光信息招股说明书，海光信息公告，平安证券研究所

2.7 华为昇腾|910B对标英伟达A100，原生生态持续完善

- **昇腾910B可对标英伟达竞品，昇腾AI原生生态持续完善。**昇腾910B采用达芬奇架构，性能基本可对标英伟达A100。华为表示，910B在训练大型语言模型时表现出色，其效率可达英伟达A100的80%，而在某些特定测试中，其性能甚至超越A100达20%。910B已可实现万卡规模量级集群。根据科大讯飞公告，2023年10月，科大讯飞与华为联合发布我国首个全国产支持万亿参数大模型训练的万卡国产算力平台“飞星一号”，基于该平台训练完成的“讯飞星火V3.5”于2024年1月30日正式发布，“飞星一号”即采用了华为昇腾910B。
- 根据华为计算公众号信息，截至2024年9月，昇腾已经累计培养3万+原生贡献者，20+伙伴及客户原生打造100+核心大算子、孵化了40+原生大模型，以及50+大模型应用，昇腾生态已经走向原生驱动。

华为昇腾计算产业全景



数据来源：昇腾社区官网，平安证券研究所

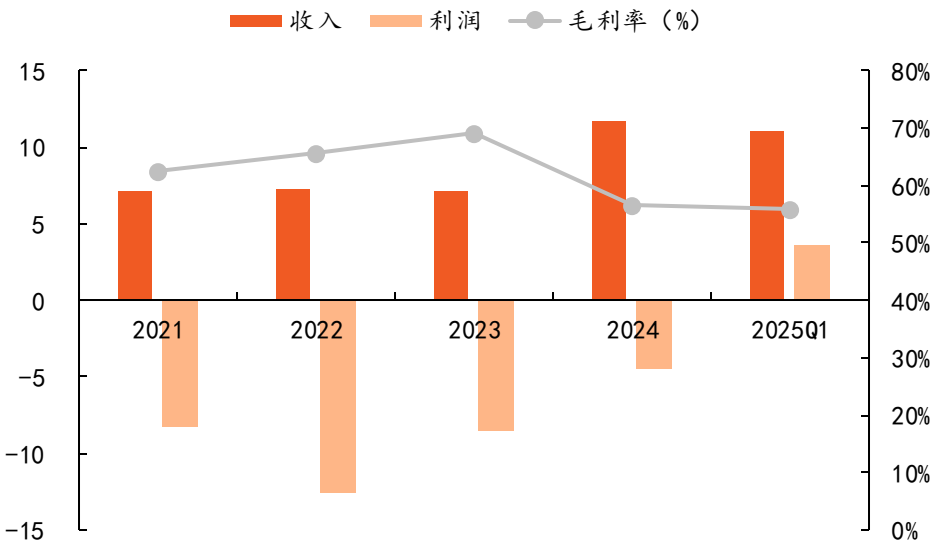
2.8 寒武纪|思元590性能大幅提升，公司迎来盈利拐点

- **寒武纪思元590训练性能大幅提升。**思元370基于7nm制程工艺，采用chiplet技术，集成了390亿个晶体管，最大算力256TOPS（INT8），是寒武纪第二代产品思元270的2倍。思元590是寒武纪新一代云端智能训练芯片，采用MLUarch05架构，能够提供更大的内存容量和带宽，IO和片间互联接口也较上代实现大幅升级。
- **寒武纪迎来盈利拐点。**根据iFind数据，2024Q4，寒武纪实现收入9.89亿元，归母净利润2.72亿元，实现单季扭亏为盈；2025Q1,公司实现收入11.11亿元，归母净利润3.55亿元，盈利得以持续且净利率进一步提升至31.98%。

寒武纪思元370系列产品性能指标

性能参数	MLU370-S4/S8	MLU370-X4/X8
制程工艺	7nm	
INT8	192 TOPS	256TOPS
FP16	72 TFLOPS	96 TFLOPS
FP32	18 TFLOPS	24 TFLOPS
内存容量	24GB/48GB	24GB/48GB
内存带宽	307.2GB/s	307.2GB/s，614.4 GB/s
系统接口	x16 PCIe Gen4	
最大热设计功耗	75W	250W

寒武纪近年业绩情况（亿元）



数据来源：寒武纪官网，iFind，平安证券研究所

2.9 燧原、沐熙、天数、壁仞等在AI算力芯片领域百花竞艳

- **燧原、沐熙、天数、壁仞等公司在国内AI算力芯片市场中也有一定的竞争力，燧原是ASIC路线，沐熙、天数、壁仞是通用GPU路线。**燧原科技，腾讯是第一大股东，基于其S60的庆阳智算中心万卡推理集群已投入使用；沐熙， 高性能GPU产品研发经验丰富，超讯通信与中特新联、星航智算采购合同中涉及采购其曦云C500-P Pcie；天数智芯，联合无问苍穹在天垓150千卡集群上将70B LLaMA模型训练性能提升至国际领先水平；壁仞科技，其通用GPU参与中国移动呼和浩特项目。

燧原、沐熙、天数、壁仞等也是国内AI算力芯片市场的重要参与者

Enflame 燧原科技

燧原® S60

- 庆阳智算中心燧原S60万卡推理集群，2025年1月开始对外提供推理算力服务。
- 太湖亿芯（无锡）智算中心S60推理算力集群。

METAX 沐曦

曦云C系列

- 超讯通信与中特新联、星航智算采购合同中涉及采购曦云C500-P Pcie，合同金额合计约14.88亿元。

天数智芯 Iluvatar CoreX

天垓100 智铠100

- 无问芯穹联合天数在天垓150千卡集群上将70B LLaMA模型训练性能提升至国际领先水平，并将一个千亿级参数的MoE模型性能提升了53.5%。

壁仞科技 BIREN TECHNOLOGY

壁砺™110E 壁砺™106B、106C

- 壁仞科技通用GPU参与中国移动呼和浩特智算中心项目。

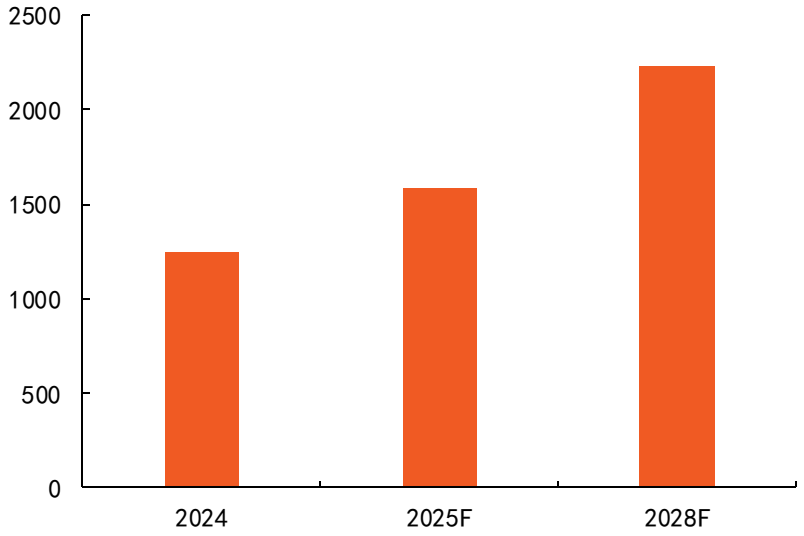
数据来源：燧原科技Enflame公众号，超讯通信公告，上海经信委公众号，壁仞科技Birentech公众号，平安证券研究所

目录CONTENTS

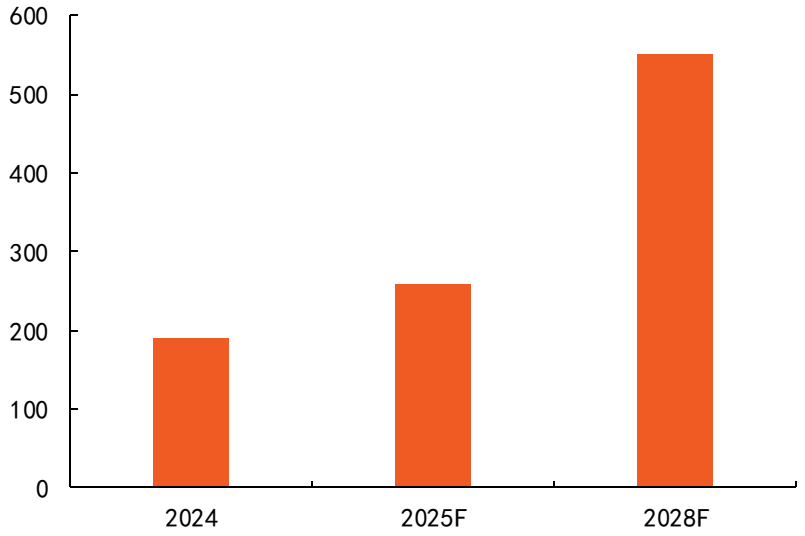
3.1 AI服务器|需求高涨，推理端增长潜力高于训练端

- **AI算力需求旺盛，AI服务器市场景气度将持续高企。**根据IDC&浪潮信息报告数据，2024年，全球AI服务器市场规模为1251亿美元，预计2025年将增长到1587亿美元，同比增速26.9%，到2028年有望突破2227亿美元，2024-2028年CAGR约15.5%。国内来看，2024年，我国AI服务器市场规模约190亿美元，同比增长86.9%，预计2028年将达到552亿美元，2024-2028年CAGR约30.6%。
- **推理端将成为AI服务器工作负载的主力需求。**根据IDC&浪潮信息报告数据，2028年，预计中国AI服务器推理工作负载将占到73.0%，而训练工作负载仅占27.0%，推理端需求远高于训练端。

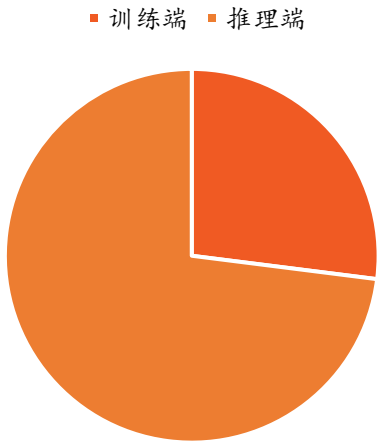
全球AI服务器市场规模预测（亿美元）



我国AI服务器市场规模预测（亿美元）



中国AI服务器工作负载情况预测（%）@2028

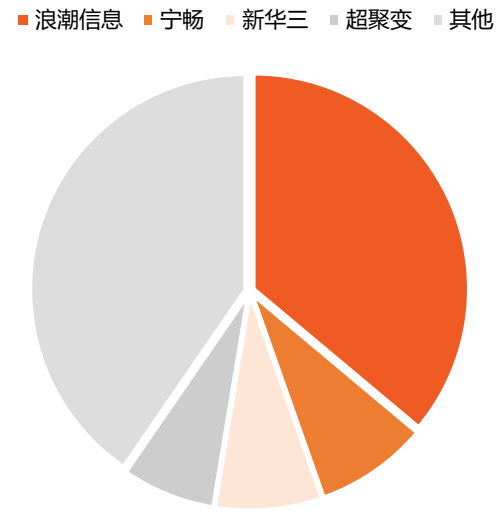


数据来源：IDC、浪潮信息《2025年中国人工智能计算力发展评估报告》，平安证券研究所

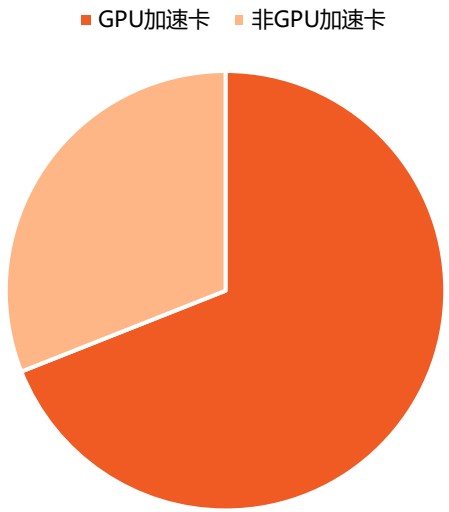
3.1 AI服务器|集中度高，本土ASIC服务器蓬勃发展

- **我国AI服务器市场集中度较高。**根据IDC数据，2024年，浪潮信息、宁畅、新华三位居我国AI服务器市场Top3，合计市占率52.6%；互联网是我国AI服务器市场最大的需求方，占整体市场超65%的份额，同时来自其他行业的采购需求也均有不同幅度的增长。
- **国内GPU服务器仍占主导，ASIC等其他AI服务器快速发展，且对本土化芯片更为青睐。**根据IDC数据，2024年，我国GPU服务器占AI服务器总市场的份额为69%，仍占据市场优势地位，同时ASIC和FPGA等非GPU架构的AI服务器也在高速增长，华为昇腾系列、寒武纪思元系列以及阿里、百度等我国CSP厂商自研AI芯片等都是ASIC架构，2024年占比超过30%，IDC预计到2029年，非GPU服务器市占比将提升到接近50%。此外，根据TrendForce集邦咨询“AI芯片自主化进程加速，云端巨头竞相自研ASIC”研报，我国AI服务器市场外购英伟达、AMD等芯片的比例，预计会从2024年约63%下降至2025年约42%，而我国本土芯片供应商在国有AI芯片政策支持下，预期2025年占比将提升至40%，几乎与外购芯片比例平分秋色。

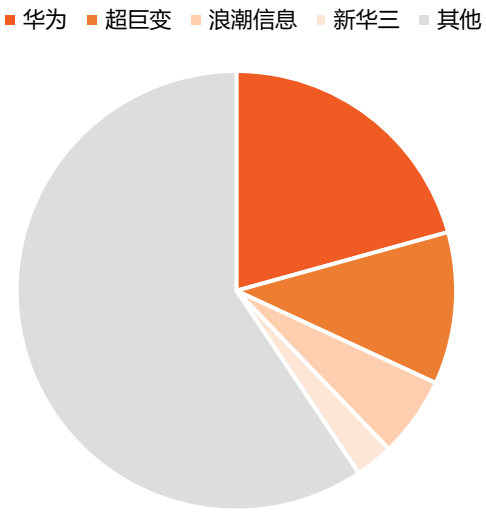
国内AI服务器行业格局@2024



国内加速计算服务器市场份额@2024



国内非GPU架构AI服务器行业格局@2024



数据来源：IDC，AI人工智能行业动态公众号，TrendForce《AI芯片自主化进程加速，云端巨头竞相自研ASIC》研报，平安证券研究所

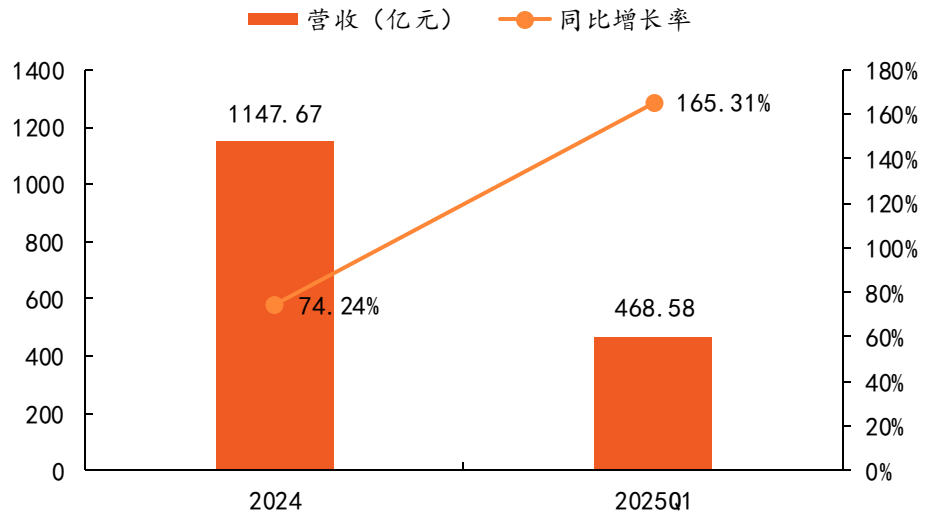
3.1 浪潮信息|全球AI服务器行业领先企业，新产品持续迭代

- **浪潮信息是全球AI服务器领先企业。**根据浪潮信息2023年半年报，截至2022年，浪潮信息人工智能服务器市场份额长期居于全球前列，连续6年保持中国第一；根据新黄河信息，2023年，浪潮信息AI服务器全球市场占有率位列第一。
- **浪潮信息持续迭代新产品，行业领先地位进一步巩固。**公司推动 AI 算力融入各类计算平台之中，并将计算品牌全面升级为“元脑”。2024 年，公司发布了业界首个仅靠 4 颗CPU运行千亿参数大模型的AI通用服务器NF8260G7，能够灵活满足基于大模型的AI应用及云计算、数据库等通用场景；还重磅发布了元脑服务器第八代新品，基于开放架构设计，业界率先实现“一机多芯”，引领多元算力生态共进，具备更全面的智能能力和更高能效；推出了国内首款42kW智算风冷算力仓，单机柜可部署AI服务器的数量是传统风冷机柜的6倍以上，相比传统风冷数据中心整体节能25%以上。受益于我国AI服务器市场的高景气度，公司2024年和2025年一季度营收持续高速增长。

浪潮信息元脑人工智能服务NF5688G8技术规格

型号	NF5688-A8-A0-R0-00
高度	6U
GPU	1块NVIDIA HGX-Hopper-8GPU模组
处理器	2颗AMD第五代EPYCTM处理器，最大cTDP 500W
内存	24条DDR5 DIMMs内存，速率最高支持6400MT/S
存储	最多支持24块2.5英寸SSD硬盘，其中最大支持16块NVMe

浪潮信息2024和2025Q1营收高速增长



数据来源：浪潮信息官网，Wind，平安证券研究所

3.1 昇腾服务器|协同合作伙伴拓展市场，发展势头良好

- 昇腾计算产业是基于昇腾系列处理器和基础软件构建的全栈AI计算基础设施、行业应用及服务，包括昇腾系列处理器、系列硬件、CANN（Compute Architecture for Neural Networks，异构计算架构）、AI计算框架、应用使能、开发工具链、管理运维工具、行业应用及服务等全产业链。昇腾计算产业已形成较为完善的生态链，协同合作伙伴拓展市场，昇腾服务器是我国国产AI芯片（ASIC架构）服务器的典型代表，发展势头良好。如前所述，根据IDC数据，2024年，我国非GPU架构的AI服务器占比已超过30%。

昇腾芯片与14家AI服务器厂商建立了生态伙伴关系

数据来源：昇腾社区网站，平安证券研究所

3.1 昇腾服务器|华为云推出超节点，在AI基础设施领域实现突破

- **华为云推出CloudMatrix 384超节点，并已在芜湖数据中心规模上线。**根据新京报消息，4月10日-11日，华为云生态大会2025在芜湖召开，华为公司常务董事、华为云计算CEO张平安公布了AI基础设施架构突破性进展，推出CloudMatrix 384超节点，并宣布已在芜湖数据中心规模上线。CloudMatrix 384是目前国内唯一正式商用的大规模超节点集群，实现从服务器级到矩阵级的资源供给模式转变。CloudMatrix 384具备高密、高速、高效的特点，通过全面的架构创新，在算力、互联带宽、内存带宽等方面实现全面领先。
- **CloudMatrix 384在多项关键指标上实现对英伟达GB200 NVL72的超越。**根据人民网引用SemiAnalysis信息，CloudMatrix 384基于384颗昇腾芯片构建，通过全互连拓扑架构实现芯片间高效协同，可提供高达300PFLOPs的密集BF16算力，接近达到英伟达GB200 NVL72的两倍，同时，CM384总内存容量超出英伟达方案3.6倍，内存带宽也达到2.1倍，为大规模AI训练和推理提供了更高效的硬件支持。CloudMatrix 384的发布，标志着我国在AI计算系统领域已具备与国际巨头正面竞争的实力，我国在AI基础设施领域实现里程碑式突破。

CloudMatrix 384具备高密、高速、高效的特点

CloudMatrix 384具备六大技术优势

技术优势	简介
MoE亲和	通过高速互联总线，能够实现一卡一专家高效分布式推理，单卡的MoE计算和通信效率都大幅提升
以网强算	通过MatrixLink服务将单层网络升级为两层高速网络
以存强算	首创了EMS弹性内存存储，打破传统GPU算力与显存绑定的关键障碍
长稳可靠	开发了昇腾云脑运维“1-3-10”标准，即1分钟感知、3分钟定界、10分钟内恢复
朝推夜训	通过“训推共池”“灵活调度”两大关键技术实现朝推夜训
即开即用	华为云已经在全国三大枢纽数据中心——乌兰察布、贵安和芜湖完成了超节点规模布局，支持百TB级的带宽互联，10毫秒时延圈覆盖全国19个城市群

数据来源：新京报，人民网，SemiAnalysis，华为云，央广网，平安证券研究所

3.2 AI大模型一体机|大幅拉低AI应用门槛

- **根据星环科技网站信息，AI大模型一体机是将大型人工智能模型与专用硬件设备整合在一起的综合性解决方案。**AI大模型一体机通常包含高性能计算硬件、预训练好的大型AI模型以及配套的软件系统，所有组件都经过优化设计，能够协同工作。与传统AI部署方式不同，一体机采用“开箱即用”的设计理念，用户无需自行搭建复杂的技术栈，大大降低了AI应用的门槛。
- 相比传统的AI大模型部署架构，AI大模型一体机在便捷性、性能优化、数据安全性等方面有较为明显的优势。

AI大模型一体机的技术架构的核心组成部分

- **计算硬件**：包括高性能GPU或TPU等专门为AI计算优化的处理器
- **软件系统**：包含操作系统、驱动程序和AI框架等基础软件
- **预装大模型**：预装的大规模预训练模型，这些模型可能专注于自然语言处理、计算机视觉或多模态任务

AI大模型一体机相比传统AI大模型部署架构的优势

优势	简介
便捷性	传统AI模型部署需要企业具备专业的技术团队，解决从硬件选型到软件配置的一系列复杂问题。而一体机将这些工作提前完成，用户只需连接电源和网络，就能快速获得AI能力。这种设计特别适合那些希望快速应用AI但又缺乏专业技术团队的中小企业和机构。
性能优化	在性能表现方面，一体机通常经过深度优化。硬件和软件的协同设计意味着计算资源能够被充分利用，避免了一般部署中常见的资源浪费问题。同时，许多一体机还针对特定场景进行了优化，比如金融领域的风险模型或医疗领域的影像分析，这使得它们在专业任务上的表现往往优于通用型AI服务。
数据安全性	与将数据上传至公有云AI服务不同，一体机可以在企业本地运行，敏感数据无需离开组织内部网络。这一特点对政府机构、医疗机构和金融机构等对数据安全要求严格的用户尤为重要。

数据来源：星环科技官网，平安证券研究所

3.2 AI大模型一体机|轻量化DeepSeek本地化部署之最佳选择

- **DeepSeek大模型的开源、低成本和高性能将大幅降低大模型的获得、部署和应用成本，很好的解决了之前大模型本地部署成本高的痛点，DeepSeek浪潮将加快AI大模型一体机的发展。**在DeepSeek横空出世之前， AI大模型一体机虽然在便捷性、性能优化、数据安全性等方面有较为明显的优势，但大模型高昂的部署成本（包括软硬件成本以及大模型的授权使用费用等），仍然极大限制了大模型一体机在本地部署的普及和推广。2024年12月和2025年1月，DeepSeek V3和R1相继正式发布，其开源、低成本、高性能特点引发了全球的广泛关注。DeepSeek在开源DeepSeek-R1-Zero和DeepSeek-R1这两个660B模型的同时，通过 DeepSeek-R1蒸馏了6个小模型开源给社区，其中32B和70B模型性能表现良好，多项能力可对标OpenAI o1-mini，蒸馏后的小模型需要的硬件部署成本更低。根据云轴科技ZStack智塔DeepSeek一体机的配置要求，部署DeepSeek R1 671B满血版，最小规模只需要1台8卡H20（141G的版本）服务器或2台8卡昇腾910B（910B_4版本）即可，大模型本地部署门槛得到大幅降低，强力推动AI大模型一体机的发展。

DeepSeek R1蒸馏的小模型性能表现良好

	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820.0**
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633.0

ZStack智塔AI一体机DeepSeek版4种配置机型

	满血版		性价比版			基础版	轻量版
支持DeepSeek参数版本	DeepSeek R1 671B		DeepSeek R1 70B			DeepSeek R1 32B	DeepSeek R1 14B/8B/7B/1.5B
最小规模	1台	2台	1台	1台	1台	1台	1台
适用场景	可支持大规模商业化应用部署		可支持大规模商业化应用部署			适合中小规模应用部署	轻量级部署及初级商用的高性价比
GPU	英伟达 H20_141G ×8	昇腾 910B_4 ×8	英伟达 L20 ×8	海光 K100_AI ×4	昇腾 300I DUO ×4	英伟达 L20 ×2	英伟达 RTX4090D ×1
CPU	英特尔	鲲鹏	英特尔	海光	鲲鹏	英特尔	英特尔
软件	ZStack AIOS智塔平台 GPU算力调度/GPU虚拟化技术/模型仓库/模型精调/推理服务/AI云原生模块/第三方AI应用市场……						
模型支持	内置DeepSeek模型、向量模型、精排模型，可部署Qwen等其他文本及多模态模型						
服务	完善的技术支持服务\软硬件安装部署\网络环境规划\性能调优测试\模型精调评估						

数据来源：DeepSeek官网，北京软件和信息服务业协会，ZStack云计算，平安证券研究所

3.2 DeepSeek大模型一体机|供应端呈现“百机大战”格局

➢ **基于DeepSeek V3/R1大模型的特点与本地部署模式良好的契合，DeepSeek一体机成为我国AI一体机的主流发展趋势。**根据《中国AI大模型一体机市场分析与品牌推荐2025》信息，从DeepSeek R1正式发布以来，我国市场已有100多家厂商推出AI（DeepSeek）一体机。其中，数字基础设施软硬件提供商、云服务厂商、GenAI初创厂商、GenAI模型/工具厂商、企业应用开发商、IT咨询/系统集成商纷纷涉足。我国DeepSeek大模型一体机在供应端呈现“百机大战”格局。

2025年1月以来推出AI(DeepSeek)一体机的部分厂商及产品简介

厂商	产品简介
联想集团、沐曦股份	2月5日，联想集团与国产GPU领军企业沐曦股份联合发布基于DeepSeek大模型的首个国产一体机解决方案。该方案以“联想服务器/工作站+沐曦训推一体国产GPU+自主算法”为核心架构，配合联想AI Force智能体开发平台，推出智能体一体机与训推一体服务器双产品形态，率先实现从千亿参数大模型训练到场景化推理落地的全链条覆盖。截至3月7日，该解决方案首月累计发货量已突破千台，配备沐曦国产GPU卡近万张，覆盖医疗、教育、制造等十余个核心行业。
天翼云	2月10日，天翼云正式推出息壤智算一体机-DeepSeek版，为各行各业提供性能卓越、安全可控的智能算力解决方案，息壤智算一体机-DeepSeek版集国产算力（鲲鹏+昇腾）、国产模型和国产云服务于一身，深度融合了DeepSeek-R1/V3系列大模型，实现了从芯片、推理引擎到模型服务的全栈国产化
深信服	根据深信服网站2月17日信息，深信服现已打造「HCI+AICP新一代超融合」解决方案，只需在原集群基础上增加一台GPU节点，就能基于本地集群快速部署并承载DeepSeek在内的企业级大模型。除了支持英伟达GPU，深信服AICP算力平台和多家国产厂商开展了广泛的软硬件兼容测试，可适配天数智芯、昇腾、海光、沐曦、燧原等多款国产卡，为用户实现算力异构管理。根据深信服2025年一季报，公司2025年一季度实现营业收入12.62亿元，同比增长21.91%，主要是因为，公司云业务订单增速较好，带动公司整体收入实现增长。
百度	根据百度智能云网站2月24日信息，百度智能云相继推出了百舸、千帆、以及一见DeepSeek一体机解决方案。百舸DeepSeek一体机是一款专为DeepSeek私有化部署场景设计的具有极致效价比的大模型一体机，搭载国产高性能芯片昆仑芯P800，单机支持DeepSeek满血版，充分满足高性能推理需求。百度智能云发布四款千帆DeepSeek一体机，国产单机8卡即可轻松承载DeepSeek满血版和蒸馏版模型，包含昆仑芯P800（OAM版）、昆仑芯P800（PCIe版）、以及昇腾Atlas 800两款机型，除了满足训推需求之外，千帆DeepSeek一体机还为企业提供了一站式的模型应用解决方案，能够满足企业全链路模型开发应用工具链需求。百度智能云一见DeepSeek一体机，除了满足昆仑芯P800 DeepSeek模型推理需求之外，结合多模态大模型、视觉专家模型，可为能源、制造、连锁等行业提供从视觉感知到智能决策的全链路闭环解决方案。
星环科技	根据星环科技网站2月28日信息，星环科技正式推出新一代高性能大模型一体机TxData-LM（LLMops for DeepSeek一体机版本），TxData-LM以“满血版”DeepSeek 671B大模型为核心，依托星环自研的Sophon LLMops平台，打通语料开发、模型训练、知识融合、应用部署等全链路流程，支持企业高效构建智能体与应用。
神州数码	2月28日，神州鲲泰问学一体机DeepSeek版在第三届北京人工智能产业创新发展大会上正式发布，神州鲲泰问学一体机DeepSeek版提供DeepSeek全系列模型选型服务及全栈部署方案，基于鲲鹏、昇腾的全栈适配服务，通过深度优化的私有化部署方案，为企业提供高效、稳定的AI服务。
云轴科技ZStack	根据云轴科技网站3月3日信息，云轴科技ZStack推出ZStack智塔AI一体机DeepSeek版，这是一款专为大模型部署与应用设计的软硬协同产品，集成多样化硬件配置（英伟达、昇腾、海光等主流GPU机型）与AI Infra平台ZStack AIOS智塔软件，支持DeepSeek全尺寸模型（1.5B参数-671B参数）等多种模型，为企业提供从模型精调到推理再到应用的全流程支持。

数据来源：中国日报网，各公司官网，平安证券研究所

3.2 DeepSeek大模型一体机|需求端快速落地

- **DeepSeek大模型一体机在政务、金融、医疗、教育、物流等多个行业快速落地，未来发展潜力较大。**根据中国日报网信息，联想集团与沐曦股份联合发布的基于DeepSeek大模型的首个国产一体机解决方案，首月累计发货量已突破千台，配备沐曦国产GPU卡近万张，覆盖医疗、教育、制造等十余个核心行业。根据新浪财经信息，5月20日，京东在京东云城市大会上表示，过去三个月，“开箱即用”的京东云大模型一体机快速发展，全国规模化落地已突破500台。

DeepSeek大模型一体机在政务、金融、医疗、教育、物流等多个行业的落地案例示例

落地单位	行业	项目简介
广州市中级人民法院	政务	根据广州日报2月26日信息，广州市中级人民法院联合广州数据集团正式上线“法院DeepSeek应用体验中心”，标志着广州市政务领域首个DeepSeek一体机私有化部署项目成功落地。该一体机部署了DeepSeek-R1-Distill-Llama-70B模型推理服务，有效解决司法系统内网环境下DeepSeek模型运用的难题，确保“司法数据不出域”。
扬州市大数据管理中心	政务	根据科创板日报5月3日信息，扬州市大数据管理中心已部署10台昇腾AI一体机，成功运行DeepSeek-R1-671B大模型，为各政务部门的人工智能应用创新提供算力支持。
安徽省马鞍山市雨山区	政务	3月14日，雨山区在全市首个引进新华三DeepSeek一体机，完成DeepSeek-R1满血版（671B）本地化部署，下一步，雨山区将依托DeepSeek一体机高性能算力底座，先行先试开发、精调、适配专用于雨山区政务服务的“雨小智”AI政务模型。
国信证券	金融	5月14日，中国招标投标公共服务平台发布国信证券DeepSeek一体机（含高速组网）采购项目候选人公示，采购需求包括2台DeepSeek一体机服务器和1套高速组网设备。
中信建投证券	金融	5月21日，中国招标投标公共服务平台发布中信建投证券deepseek一体机租赁采购项目征集公告，采购内容即为中信建投证券deepseek一体机租赁。
黑龙江数智产业投资有限公司	金融	4月29日，中国招标投标公共服务平台发布DeepSeek一体机及配套设备采购项目-公开招标公告，招标人为黑龙江数智产业投资有限公司，采购需求包括大模型一体机2台、通用服务器2台等。
爱尔眼科	医疗	根据联想网站3月7日信息，2月底，爱尔眼科与联想合作，率先在行业内实现了本地部署DeepSeek满血版大模型并引入联想AI一体机方案。爱尔眼科数字人“爱科”（Eyecho）正式升级接入DeepSeek R1推理模型，实现智能问答、医学科普及患者互动的精准性与效率提升，用户体验全面优化。
广州医科大学附属脑科医院	医疗	4月21日，中国政府采购网发布广州医科大学附属脑科医院DeepSeek本地部署一体机服务器采购项目（CX2025-0002）结果公告，中标供应商为中国电信股份有限公司广东分公司，中标金额为46万元。
监利市人民医院	医疗	3月37日，监利市人民医院网站发布DEEPSEEK一体机院内采购公告，监利市人民医院拟对DEEPSEEK一体机进行院内采购，项目预算为29万元。
浙江师范大学	教育	3月14日，浙江政府采购网发布浙江师范大学2025年4月一体机集群政府采购意向，采购单位为浙江师范大学，预算金额为350万，采购需求包括deepseek一体机2台、相关网络交换设备必要存储设备等。
山东港口物流集团	物流	5月19日，中国招标投标公共服务平台发布山东港口物流集团发展公司（中欧“铁运达”项目）DeepSeek一体机采购项目标段1中标结果公示，成交价为1678962元。

数据来源：中国日报网，广州日报，科创板日报，雨山发布公众号，中国招投标网，联想官网，中国政府采购网，监利市人民医院网，浙江政府采购网，平安证券研究所

目录CONTENTS

4.1 投资要点

- **投资建议：**DeepSeek火爆出圈，轻量化、低成本、高性能，推理场景逐渐打开，推理端算力需求将逐步超过训练端。当前，AI算力是美国对华科技制裁的重灾区，先进的AI算力芯片无法出口至国内，反向倒逼国内AI算力从设计到制造到整机的全面国产替代，国内通用GPU、ASIC芯片蓬勃发展，同时服务器和一体机厂商也逐渐向国产AI芯片倾斜，全产业链合力，国内AI算力自主可控已取得不菲成果。推荐海光信息、浪潮信息、龙芯中科、芯原股份、紫光股份、深信服、神州数码，建议关注寒武纪、华勤技术、软通动力。

股票简称	股票代码	2025/6/10	EPS（元）				PE（倍）				评级
		收盘价（元）	2024A	2025F	2026F	2027F	2024A	2025F	2026F	2027F	
海光信息	688041	141.98	0.83	1.27	1.86	2.65	170.9	111.6	76.5	53.6	推荐
龙芯中科	688047	130	-1.56	-0.48	0.13	0.73	-83.4	-271.5	1022.2	177.3	推荐
浪潮信息	000977	51.39	1.56	2.18	2.74	3.31	33.0	23.6	18.8	15.5	推荐
芯原股份	688521	84	-1.20	-0.53	-0.17	0.15	-70.0	-160	-489.2	568.5	推荐
紫光股份	000938	24.74	0.55	0.71	0.86	1.08	45.0	34.8	28.6	23.0	推荐
深信服	300454	96.51	0.47	0.97	1.25	1.65	206.8	99.5	77.0	58.5	推荐
神州数码	000034	38.29	1.06	1.72	2.06	2.48	36.2	22.3	18.6	15.4	推荐
寒武纪	688256	619.86	-1.08	2.94	5.01	7.81	-572.1	210.9	123.6	79.4	未评级
华勤技术	603296	71.09	2.88	3.57	4.34	5.24	24.7	19.9	16.4	13.6	未评级
软通动力	301236	54.06	0.19	0.41	0.57	0.77	285.6	130.6	94.2	70.2	未评级

数据来源：iFind，平安证券研究所

4.2 风险提示

- 1）大模型应用落地不及预期的风险。大模型逐渐成熟，落地节奏致关重要，若下游推理的应用场景落地不及预期，可能导致智能算力需求疲软。
- 2）国产AI芯片开发不及预期的风险。AI芯片设计难度大，更新迭代较快，且国内起步稍晚，若国内AI芯片设计公司产品研发速度缓慢，可能导致国内AI算力供给不足。
- 3）美国对华科技制裁的风险。AI芯片通常使用较先进的制程，若国内晶圆制造突破迟缓或产能不足，可能导致AI算力芯片供不应求，成为制约国内AI整体发展的瓶颈。

股票投资评级：

强烈推荐（预计6个月内，股价表现强于市场表现20%以上）

推　　荐（预计6个月内，股价表现强于市场表现10%至20%之间）

中　　性（预计6个月内，股价表现相对市场表现±10%之间）

回　　避（预计6个月内，股价表现弱于市场表现10%以上）

行业投资评级：

强于大市（预计6个月内，行业指数表现强于市场表现5%以上）

中　　性（预计6个月内，行业指数表现相对市场表现在±5%之间）

弱于大市（预计6个月内，行业指数表现弱于市场表现5%以上）

公司声明及风险提示：

负责撰写此报告的分析师（一人或多人）就本研究报告确认：本人具有中国证券业协会授予的证券投资咨询执业资格。

平安证券股份有限公司具备证券投资咨询业务资格。本公司研究报告是针对与公司签署服务协议的签约客户的专属研究产品，为该类客户进行投资决策时提供辅助和参考，双方对权利与义务均有严格约定。本公司研究报告仅提供给上述特定客户，并不面向公众发布。未经书面授权刊载或者转发的，本公司将采取维权措施追究其侵权责任。

证券市场是一个风险无时不在的市场。您在进行证券交易时存在赢利的可能，也存在亏损的风险。请您务必对此有清醒的认识，认真考虑是否进行证券交易。

市场有风险，投资需谨慎。

免责条款：

此报告旨为发给平安证券股份有限公司（以下简称“平安证券”）的特定客户及其他专业人士。未经平安证券事先书面明文批准，不得更改或以任何方式传送、复印或派发此报告的材料、内容及其复印本予任何其他人。

此报告所载资料的来源及观点的出处皆被平安证券认为可靠，但平安证券不能担保其准确性或完整性，报告中的信息或所表达观点不构成所述证券买卖的出价或询价，报告内容仅供参考。平安证券不对因使用此报告的材料而引致的损失而负上任何责任，除非法律法规有明确规定。客户并不能仅依靠此报告而取代行使独立判断。

平安证券可发出其它与本报告所载资料不一致及有不同结论的报告。本报告及该等报告反映编写分析员的不同设想、见解及分析方法。报告所载资料、意见及推测仅反映分析员于发出此报告日期当日的判断，可随时更改。此报告所指的证券价格、价值及收入可跌可升。为免生疑问，此报告所载观点并不代表平安证券的立场。

平安证券在法律许可的情况下可能参与此报告所提及的发行商的投资银行业务或投资其发行的证券。