中信建投证券 CHINA SECURITIES

证券研究报告·行业深度

国内 AI 行业蓄势待发，国产算力迈入自强新纪元

通信

维持 强大于市

发布日期： 2024 年 04 月 09 日

核心观点

政策持续大力推动国内 AI 产业发展，国产算力基础设施行业将快速增长。国产头部 AI 芯片单芯片算力或已接近 A100、或优于 H20，已基本满足大规模使用条件。模型和应用层面，国内领先的大模型基本实现能力边界的突破，应用端有望迎来加速落地。AI 产业发展，算力先行，尤其在美国对中国先进芯片进口限制持续升级的背景下，国产算力自立自强大势所趋，将直接拉动服务器、交换机、光模块、液冷、连接器/线束、PCB、IDC 建设等环节需求，建议重视。

1、近期，美国再次升级 AI 芯片和相关工具出口 措施，国产算力自立自强大势所趋。AI 发展，算力先行，此前国内 AI 发展掣肘于海外 AI 芯片禁运和国产 AI 芯片能力不足，目前 海思、寒武纪、平头哥、壁仞科技、百度昆仑芯、燧原科技、海光等国内 GPU 厂商均已经推出用于训练、推理场景的算力芯片，性能在不断提升，国产头部芯片单芯片算力或已接近 A100、或优于 H20，已基本满足大规模使用条件。

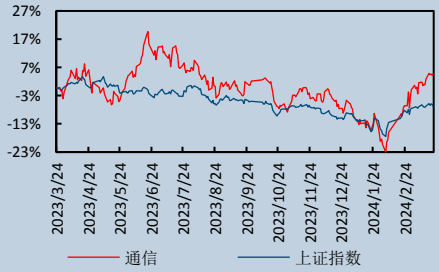
2、国产算力发展，将使更多价值量留存在国内产业链。在海外 AI 芯片主导的 AI 算力产业链中，AI 芯片、服务器、交换机等大价值量环节基本由海外公司主导，而国产算力产业链自身基本可以实现闭环，各环节的国内公司都将集中受益。

3、服务器：AI 服务器高增，芯片国产化渗透提升带来竞争格局变化。**交换机：** 网支撑高性能计算场景已经逐步得到验证，国内交换机厂商 400G、800G 相关订单预计将实现高速增长。**光模块：**2024 年国内预计 400G 需求大幅增长，部分头部 CSP 可能将采购 800G 产品。**液冷：**运营商新增 AI 服务器招标中液冷渗透比例已经达到大份额，2024 年进入实质性规模部署阶段。**连接器/线束：**AI 带动连接器系统向 112G/224G 等升级，拉动高速产品需求。**PCB：**AI 拉动高速 PCB 升级，利好头部厂商份额和盈利提升。**IDC 建设：**关注智算中心建设和存量改造机会。

4、优选以下公司建议重点关注：交换机环节锐捷网络、菲菱科思等；光模块环节中际旭创、天孚通信、新易盛、华工科技、源杰科技等；液冷环节英维克、润泽科技等；连接器环节华丰科技、鼎通科技等；IDC 建设环节科士达、科华数据等。

5、风险提示：国内算力芯片等关键器件供应不足；国内大模型发展和 AI 应用落地不及预期；资本开支投入不及预期；市场竞争加剧等。

市场表现



目 录

图目录

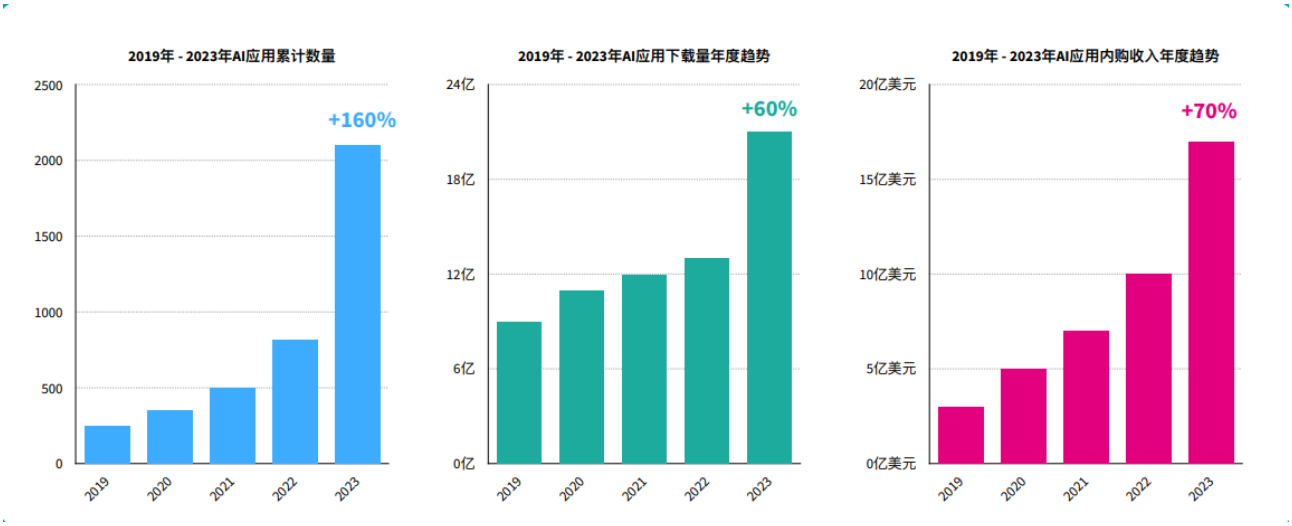
表目录

一、国内 AI 产业有望迎来跨越式发展

1.1 海外 AI 产业蓬勃发展

OpenAI 于 2023 年 3 月发布 GPT-4，谷歌于 2023 年 12 月发布 Gemini 大模型，并在近期推出 Gemini1.5 pro 以及开源模型 Gemma，大模型能力持续迭代升级。伴随大模型能力的提升，海外 AI 应用蓬勃发展，云大厂比如微软推出 copilot、bing AI 等，谷歌推出 workspece、聊天机器人 Gemini 等外，B 端垂直企业服务、C 端应用等层出不穷。据 SensorTower 数据显示，2023 年，AI 应用年度下载量和内购收入分别上涨 60%和 70%，超过 21 亿次和 17 亿美元（其中 2023H1 下载量突破 3 亿次）。英伟达提到，FY24 全年估计 40%收入来自 AI 推理端。近期 openai sora、谷歌 Genie 发布，视频应用领域 AI 能力边界大幅跃升，AI 向基础世界模型、AGI 领域迈进。

图 1: AI 应用下载量与收入高速增长



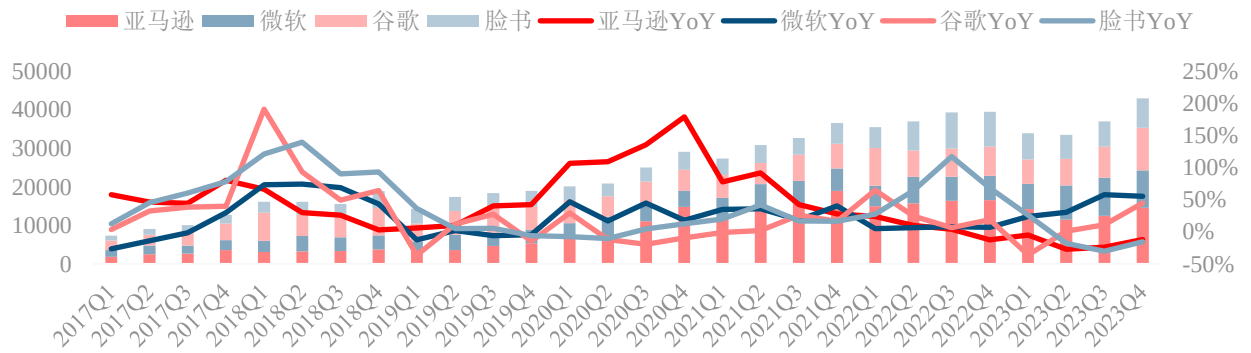
数据：SensorTower，中信建投

从支撑 AI 发展的基础设施角度，不管是从英伟达、超微电脑、台积电、AMD 等硬件厂商的业绩和指引，还是从海外云厂商的 capex 投入，都印证海外 AI 产业的持续提速。

英伟达 FY24Q3、FY24Q4 业绩持续超过分析师预期，主要来自 AI 带动数据中心业务超预期带动，英伟达对下一季度指引乐观，预计 FY25Q1 收入 240 亿美元，同样超过分析师预期的 219 亿美元。超微电脑 FY24Q2 营收超预期，FY24Q3 预计净销售额在 37 亿美元至 41 亿美元之间，远超市场预期，主要得益于 AI 系统强劲需求的驱动。台积电预计未来几年 AI 相关业务 CAGR 将达 50%，上调远期 AI 营收占比目标，预期 2027 年 AI 营收占比达到高双位数（high teen），此前预期为低双位数（low teen）。AMD 在 2023 年 12 月上调 规模预测，预计到 2027 年，人工智能 的整体市场规模将达 4000 亿美元，CAGR 达到 70%，此前 2023 年 8 月 AMD 预计 2027 年人工 行业规模为 1500 亿美元。

海外云厂商对 AI 投入展望持续乐观。谷歌指引 2024 年资本支出将明显增长。meta 指引 2024 年资本开支 300 亿美元-370 亿美元，上限上调 20 亿美元。微软表示，基于对云和人工智能基础设施的投资、第三方产能合同的交付转移到下一季度，预计下一季度资本支出将环比大幅增加。亚马逊预计 2024 年资本支出将同比增加。

图 2：海外云厂商资本开支情况（百万美元）



数据　：Bloomberg，亚马逊官网，微软官网，谷歌官网，中信建投

1.2 国内大模型基本实现能力边界突破

国内厂商也加快研发节奏，纷纷发布大模型产品，并不断持续迭代更新。2023 年 3 月-6 月间，包括百度、清华智谱、阿里巴巴、科大讯飞、百川智能等厂商相继发布自己的大模型产品，后续持续迭代更新，在 2023 年 9 月、10 月前后发布重要更新，提升模型能力。国内领先的大模型基本在 2023 年 10 月至 11 月实现了能力边界的突破，实现看齐甚至部分能力超越 ChatGPT，并且后续在持续的进一步迭代升级。随着国内大模型能力的提升，AI 应用预计 2024 年也将迎来加速落地。

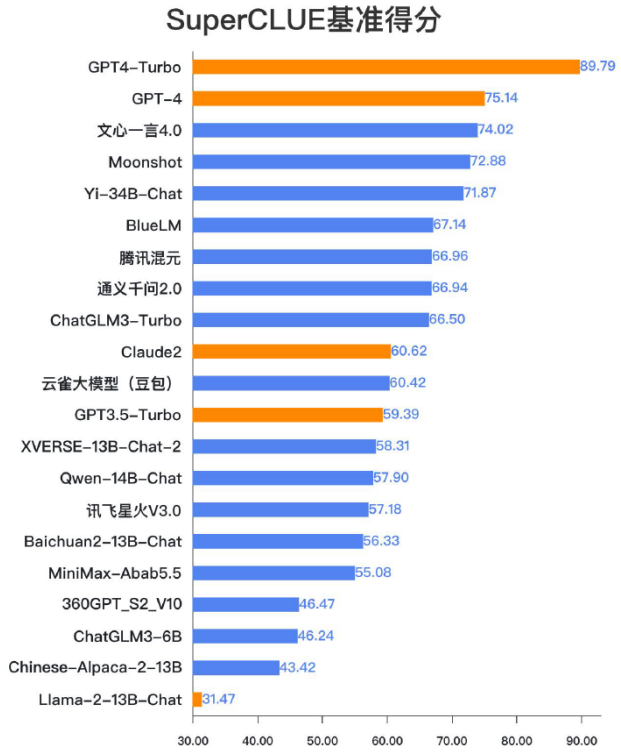
表 1：国内部分大模型厂商产品情况

厂商	大模型	发布时间
百度	文心一言	2023 年 3 月
	文心一言 4.0	2023 年 10 月
月之暗面	Kimichat	2023 年 10 月
阿里巴巴	通义千问	2023 年 4 月
	通义千问 2.0	2023 年 10 月
科大讯飞	讯飞星火 V1.0	2023 年 5 月
	讯飞星火 V1.5	2023 年 6 月
	讯飞星火 V2.0	2023 年 8 月
	讯飞星火 V3.0	2023 年 10 月
	讯飞星火 V3.5	2024 年 1 月
百川智能	baichuan-7B	2023 年 6 月
	baichuan-13B、baichuan-13B-chat	2023 年 7 月
	baichuan2-7B/13B、baichuan2-53B	2023 年 9 月
清华智谱	ChatGLM-6B	2023 年 3 月
	ChatGLM3-6B	2023 年 10 月
	GLM4、定制化大模型 GLMs	2024 年 1 月

资料　：新浪，亿欧网，网易，时代财经，清华大学官网，搜狐，中信建投

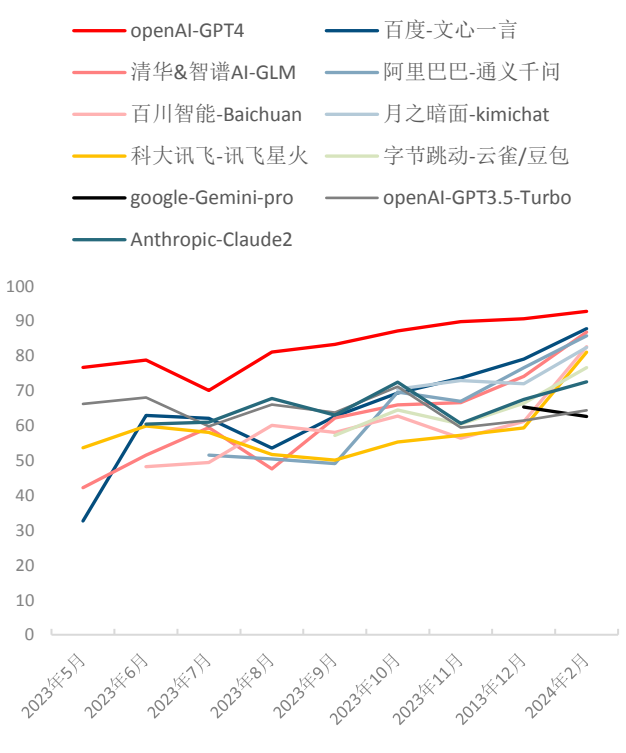
参考国内中文模型评测机构 SuperCLUE 发布中文大模型基准测评，对比来看，国内大模型厂商的能力在快速提升。2023 年 5 月，国内大模型总体与 GPT3.5 有约 20 分的差距，国产得分最高的星火认知大模型总分 53.58，而 GPT3.5 为 66.18。2023 年 11 月，国产头部大模型已基本完成对 GPT3.5 的总分超越，与 GPT4-Turbo 的差距也在快速缩小，74.02 分的文心一言 4.0、72.88 分的 Moonshot 等大模型超越了 59.39 分的 GPT3.5，与 89.79 分的 GPT4 仍有距离。SuperCLUE 最新的 2024 年 2 月测评结果显示，国产第一梯队大模型已将与 GPT4.0 的得分差距拉至 10 分以内，其中文心一言 4.0 总分 87.75，GLM-4 总分 86.77，通义千问 2.1 总分 85.7、Baichaun3 总分 82.59、Moonshot（kimichat）总分 82.37、讯飞星火 V3.5 总分 81.01，而 GPT4.0-Turbo 总分 92.71、GPT3.5 总分 64.34。

图 3: SuperCLUE 中文通用大模型基准测评（2023.12）



数据 ：SuperCLUE，智东西，中信建投

图 4：部分模型 SuperCLUE 中文通用大模型基准测评对比



数据 ：SuperCLUE，中信建投

近期 Kimi 支持 200 万字超长文本，用户数激增，国内模型的能力和应用的展望进一步乐观。2024 年 3 月 18 日，月之暗面宣布 Kimi 智能助手已支持 200 万字超长无损上下文（2023 年 10 月刚发布时，Kimi 可支持的无损上下文输入长度为 20 万字），在长文本处理能力上取得了突破性进展，并于即日起开启产品内测。Kimi 的月活用户从 2023 年底的 50 万左右增至接近 300 万，网页端的日活从 3 月 9 日的 12 万多增至 14 日的 35 万左右，3 月 21 日，Kimi 因访问量暴增而疑似宕机。

1.3 国家大力推动 AI 建设与应用落地

2024 年 2 月 19 日，国务院国资委召开中央企业人工智能专题推进会。会议认为，加快推动人工智能发展，是国资央企发挥功能使命，抢抓战略机遇，培育新质生产力，推进高质量发展的必然要求。中央企业要主动拥抱人工智能带来的深刻变革，把加快发展新一代人工智能摆在更加突出的位置，不断强化创新策略、应用示范

和人才聚集，着力打造人工智能产业集群，发挥需求规模大、产业配套全、应用场景多的优势，带头抢抓人工智能赋能传统产业，加快构建数据驱动、人机协同、跨界融合、共创分享的智能经济形态。会议强调，中央企业要把发展人工智能放在全局工作中统筹谋划，深入推进产业焕新，加快布局和发展人工智能产业。**要夯实发展基础底座，把主要资源集中投入到最需要、最有优势的领域，加快建设一批智能算力中心，进一步深化开放合作，更好发挥跨央企协同创新平台作用。**开展 AI+专项行动，强化需求牵引，加快重点行业赋能，构建一批产业多模态优质数据集，打造从基础设施、算法工具、智能平台到解决方案的大模型赋能产业生态。10 家头部中央企业签订倡议书，表示将主动向社会开放人工智能应用场景。

表 2：推动人工智能发展部分相关政策

时间	政策/会议	发布部门	内容
2022 年 7 月	《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》	科技部、教育部、工业和信息化部、交通运输部、农业农村部、国家	要求以人工智能与实体经济深度融合为主线，强化场景创新，加速人工智能技术应用，促进经济高质量发展。坚持企业主导、创新引领、开放融合、协同治理的原则，围绕智能经济、科研等领域打造人工智能重大场景，强化企业、高校等机构的场景创新能力，加快场景开放，扩大场景创新要素供给。
2022 年 8 月	《科技部关于支持建设新一代人工智能示范应用场景的通知》	科技部	围绕构建全链条、全过程的人工智能行业应用生态，加强研发上下游配合与新技术集成，打造一批可复制、可推广的标杆型人工智能应用场景。首批支持智慧农场、智能港口、智能矿山、智能工厂、智慧家居、智能教育、自动驾驶、智能诊疗、智慧法院、智能供应链共十个示范应用场景。
2023 年 7 月	《生成式人工智能服务管理暂行办法》	国家网信办、国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局	提出国家坚持发展和安全并重、促进创新和依法治理相结合的原则，规定生成式人工智能服务的基本规范，促使监管部门采取精细化措施进行监管，鼓励生成式人工智能创新发展和产业应用。
2024 年 2 月	“AI 赋能 产业焕新”中央企业人工智能专题推进会	国务院国资委	会议强调中央企业要更重视并主动拥抱人工智能变革，把发展人工智能放在全局工作中统筹规划，集中资源投入最需要、最有优势的领域，加快建设智能算力中心，促进跨央企协同创新，带头抢抓人工智能赋能传统产业，构建数据驱动、人机协同、跨界融合、共创分享的智能经济形态。

资料 ：国务院，工信部，科技部，中信建投

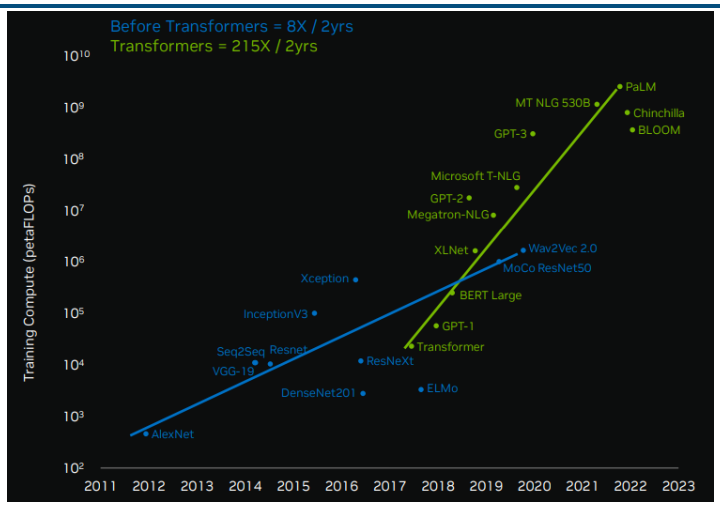
二、国产算力基础设施迎来发展机会

2.1 AI 发展需要强大算力基础设施支撑

大语言模型所使用的数据量和参数规模呈现“指数级”增长，带来智能算力需求爆炸式增长。OpenAI 在 2018 年推出的 GPT 参数量为 1.17 亿，预训练数据量约 5GB，而 GPT-3 参数量达 1750 亿，预训练数据量达 45TB，

而当前来看，大模型参数进一步提升，已经达到万亿级，并持续迭代发展。**训练阶段算力需求与模型参数数量、训练数据集规模等有关**，参考天翼智库测算信息，根据 OpenAI 发布的论文《Scaling Laws for Neural Language Models》数据，训练阶段算力需求=6×模型参数数量×训练集规模，GPT-3 模型参数约 1750 亿个，预训练数据量为 45TB，折合成训练集约为 3000 亿 tokens，GPT-3 的总算力消耗约为 3646PFLOPS-day，实际运行中，GPU 算力除用于模型训练，还需处理通信、数据读写等任务，对算力会有更大消耗。**面向推理侧算力需求，**参考天翼智库测算信息，根据 OpenAI 发布的论文《Scaling Laws for Neural Language Models》数据，平均每 1000 个 token 对应 750 个单词，推理阶段算力需求=2×模型参数数量×token 数。ChatGPT 上市仅 5 天就突破 100 万用户，两个月内用户就突破 1 亿大关，现在每周活跃用户维持在亿量级。假设按照 1 亿的 ChatGPT 的活跃用户数、日活跃用户 2000 万人，平均每位用户单次查询对应 1000 个 token，每天查询 10 次，GPT-3 模型每日对话产生推力算力需求为 810PFLOPS-day，同样考虑到有效算力比率，实际运行中需要更大算力支撑。

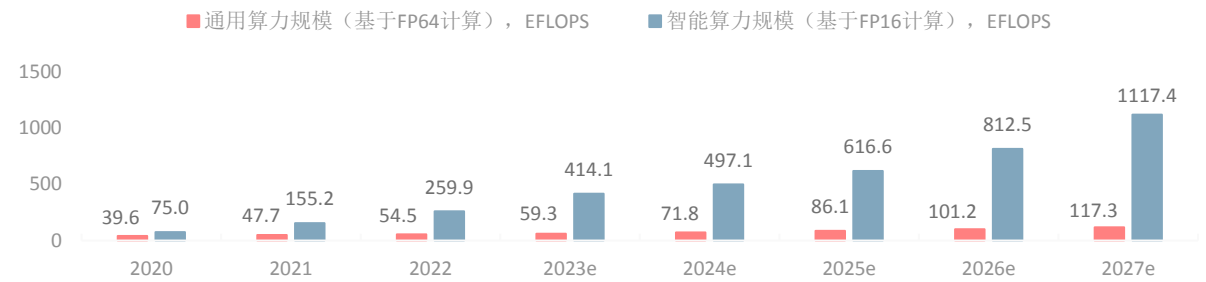
图 5: 大语言模型所使用的数据量和参数规模呈现“指数级”增长



数据　　：英伟达官网，中信建投

人工智能的发展将带动算力规模高速增长，继而刺激算力基础设施的需求。根据中国信通院数据，2022 年全球计算设备算力总规模达到 906EFLOPS，预计未来 5 年全球算力规模增速将超 50%。IDC 数据显示，2022 年中国通用算力和智能算力规模分别达 54.5EFLOPS（基于 FP64）和 259.9EFLOPS（基于 FP16），2027 年通用算力和智能算力规模将达到 117.3EFLOPS 和 1117.4EFLOPS，预估未来 5 年复合增长率 16.6%和 33.9%。

图 6: 2020-2027 中国通用和智能算力规模（EFLOPS）



数据　　：IDC，中信建投

运营商加大智能算力基础设施投入。中国移动 2024 年计划总体资本开支 1730 亿元同比下降 4%，用于算力

资本开支计划 475 亿元同比增长 21%。中国电信 2024 年计划总体资本开支 960 亿元同比下降 4%，用于产业数字化资本开支 370 亿元同比增长 4%，用于云/算力投资 180 亿元。中国联通 2024 年计划总体资本开支 650 亿元同比下降 12%，公司表示投资重点将由稳基础的联网通信业务转向高增长的算网数智业务。在相关 AI 服务器采购方面，2023 年 8 月，中国电信启动 2023-2024 年 AI 算力服务器集采，整体采购规模为 4175 台，中标总价超 84 亿元。2023 年 9 月，中国移动启动了 2023 年至 2024 年新型智算中心（试验网）采购项目，采购人工智能服务器（2454 台）、数据中心交换机（204 套）及其配套产品等，总价约 33 亿元（标包 4-12 总额，标包 1-3 采购失败）。2024 年 3 月，中国联通发布 2024 年人工智能服务器集中采购项目资格预审公告，涵盖人工智能服务器合计 2503 台，关键组网设备 RoCE 交换机合计 688 台。同时，三大运营商也在积极加快智算中心等基础设施的建设。

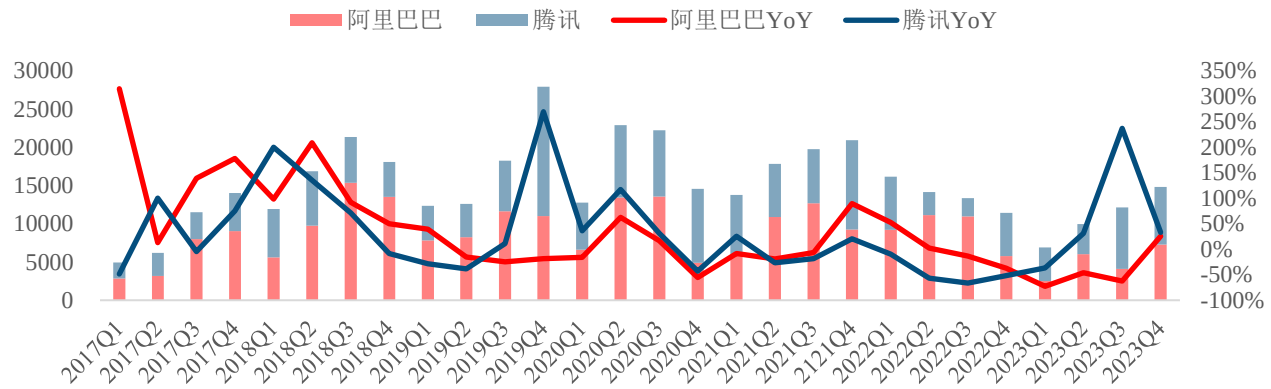
表 3：三大运营商公开在建/已投运智算中心汇总

建设主体	名称	建设规模	投运情况
中国电信	中国电信临港智算中心	位于上海，是全国首个万卡规模国产液冷集群，首批 1.5 万卡能力，预计将于 2024 年上半年逐步完成建设，其中国产算力卡将达万张。	正在建设中
	中国电信京津冀大数据智能算力中心	位于天津， 一期项目于 2021 年投产，2024 年全部建成后，能够提供约 4.2 万个机架，预计可带动大数据、人工智能等上下游产业链投资近 500 亿元。	建设完善中
	中国电信安徽智算中心	位于安徽省合肥市，总投资预计超过 100 亿元，可提供 16000 个中高密度机架，支持约 30 万台服务器运行，承载算力规模达到 2.2E FLOPS。	已启用
	中国电信长三角国家枢纽嘉兴算力中心	可提供约 1.56 万架机柜资源，建成后将成为“国家云”建设核心基地、长三角新型算力调度中心。	正在建设中
中国移动	中国移动呼和浩特智算中心	2023 年 6 月启 设，项目总投资超 20 亿，建成后算力规模达 5.5EFLOPS，国产化率超 80%，其中打造的 B07 机房楼建成后将成为全球运营商最大的单体智算中心，算力规模达 5.5EFlops，使用万片级 AI 加速芯片。	正在建设中 部分已启用
	中国移动长三角(芜湖)算力中心项目	总投资超 20 亿元，规划智算算力超 2000PFLOPS，由新华三、六尺科技和安徽移动芜湖分公司联合建设，“东数西算”芜湖集群首个大规模智算中心。	正在建设中
	重庆移动智算中心	位于两江新区水土新城，新增 2000 余台高性能服务器。	已揭牌
	中国移动智算中心(武汉)项目	预计 2024 年建成投运，首期规划 1000PFLOPS（A800、H800、昇腾 910 服务器），采用“风冷+液冷”散热模式。	已完成规划
中国联通	中国联通(青岛)智算中心	位于青岛市西海岸新区， 一期工程已建成 2 栋 IDC 机房，全部建成后将拥有 6 栋 IDC 机房，可提供 12000 个机柜。	已启用
	中国联通长三角(芜湖)智算中心	总投资 60 亿元，首期规划建设 3000P 智算能力，总机柜数不少于 10000 架，是中国联通在长三角地区等级最高、规模最大的智算中心。	正在建设中
	中国联通广东 Al 智算中心	位于广东省深圳市，由中国联通研究院、广东联通携手 建设的全栈自主创新 AI 智算中心，超过 2000 台 昇腾 AI 服务器。	已投运

资料 ：工联网 iitime，中信建投

国内互联网厂商资本开支呈现回暖态势。2023 年全年腾讯资本开支为 238.93 亿元，同比增长 32.6%，2023Q1、2023Q2、2023Q3、2023Q4 腾讯的资本开支分别为 44.11、39.35、80.05、75.24 亿元，分别同比-36.7%、+31.1%、+236.8%、+33.1%。阿里巴巴在 2023 年前三季度单季资本开支同比均呈现下滑，2023Q1、2023Q2、2023Q3 阿里巴巴的资本开支分别为 25.13、60.07、41.12 亿元，分别同比-72.7%、-46.0%、-62.5%，2023Q4 同比转为正增长，资本开支为 72.86 亿元，同比增长 25.8%。

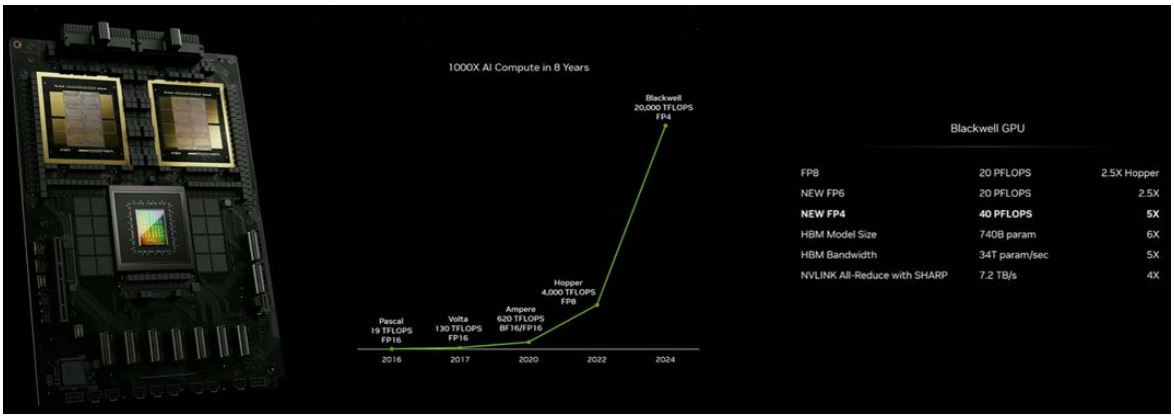
图 7：国内云厂商资本开支情况（百万元）



数据　　：Bloomberg，阿里巴巴官网，腾讯官网，，中信建投

英伟达持续升级 GPU，算力持续提升。2024 年 3 月 GTC 上，英伟达发布 GB200 超级芯片，采用 Blackwell 架构，采用台积电的 4 纳米（4NP）工艺，整合两个独立制造的裸晶（Die）形成一个 Blackwell GPU，两个 Blackwell GPU 与一个 GraceCPU 结合成为 GB200 superchip。Blackwell GPU 共有 2080 亿个晶体管，上一代 H100 只有 800 亿晶体管，整体性能明显提升。一个 GB200 NVL72 就最高支持 27 万亿参数的模型。英伟达表示，过去在 90 天内训练一个 1.8 万亿参数的 MoE 架构 GPT 模型，需要 8000 个 Hopper 架构 GPU，15 兆瓦功率，如今同样给 90 天时间，在 Blackwell 架构下只需要 2000 个 GPU，以及 1/4 的能源消耗。

图 8：英伟达 AI 计算能力过去 8 年提升 10000 倍



数据　　：英伟达 GTC，中信建投

2.2 禁运持续升级，国产化大势所趋

美国对中国先进芯片进口限制持续升级。2023 年 10 月，美国颁布新的半导体出口限制，对芯片算力和性能密度做了更严格的规定，A100/A800、H100/H200/H800、L4、L40s 均不满足出口条件。在 2022 年 8 月，美国首次针对中国实施大规模芯片出口制裁，停止出口 A100 和 H100 两款芯片和相应产品组成的系统。本次制裁主要限制总计算性能（算力*位宽）≥4800 且互联带宽≥600GB/s 的高端 AI 芯片出口，在制裁后，英伟达为中国重新设计了 A800 和 H800 两款“阉割版”芯片，主要在互联速率和双精度计算性能上做了限制。2023 年 10 月升级版本的芯片禁令加大了打击力度，性能满足以下条件均受出口　　：(1)总计算能力 TPP（算力*位宽）超

过 4800 的芯片；(2)TPP 超过 1600 且 PD（TPP/芯片面积）超过 5.92 的芯片；(3)2400≤TPP<4800，且 1.6≤PD<5.92 的芯片；(4) 1600≤TPP，且 3.2≤PD<5.92 的芯片。在此要求下，A100/A800、H100/H200/H800、L4、L40s 均不满足出口条件，英伟达只能全方位削弱芯片算力，向中国提供 H20、L20、L2 芯片。而近日美国政府再次升级对华半导体出口　　措施。参考钛媒体信息，北京时间 2024 年 3 月 30 日凌晨，美国商务部下属的工业与安全局（BIS）发布“实施额外出口　　”的新规措施，修订了 BIS 于 2022、2023 年 10 月制定的两次出口限制新规，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，此次新规中，BIS 删除和修订了部分关于美国、中国澳门等地对华销售半导体产品的限制措施，包括中国澳门和 D:5 国家组将采取“推定拒绝政策”，并且美国对中国出口的 AI 半导体产品将采取“逐案审查”政策规则，包括技术级别、客户身份、合规计划等信息全面查验。

国内算力自立自强是必然趋势。此前国内对英伟达芯片依赖度较高，2022 年，中国 AI 加速卡市场中，英伟达占据 85%的出货量，而国产芯片中，　　、百度昆仑、寒武纪、燧原各自占比 10%、2%、1%、1%。IDC 数据显示，2023 年上半年，中国加速芯片的市场规模超过 50 万张。从技术角度看，GPU 卡占有 90%的市场份额，从品牌角度看,中国本土 AI 芯片品牌出货超过 5 万张，占比整个市场 10%左右的份额。当前禁运持续升级，但是国内人工智能发展的趋势和力度并不会因此而发生变化，相反我们更需要重视人工智能的发展，美国对于中国先进芯片的限制升级可能将进一步推动我国高水平科技自立自强的步伐。

预计未来国产化比例将大幅提升，短期由于国内算力芯片供需的缺口，包括 H20 等在内的海外芯片也预计对国内算力行业进一步形成补充。H20 芯片在单卡性能上不具备突出优势，但利用 NVLINK 技术集群性能提升。

表 4：英伟达部分芯片性能对比

	FP64	FP32	FP16 Tensor Core	GPU 内存	GPU 内存带宽	TDP	interconnect
H200 SXM	34 TFLops	67 TFLops	1979 TFLops	141GB	4.8TB/s	700W	NVLink：900GB/s PCIe 5.0: 128GB/s
H100 SXM	34 TFlops	67 TFlops	1979 TFlops	80GB	3.35TB/s	700W	NVLink：900GB/s PCIe 5.0: 128GB/s
H800 SXM	1 TFlops	67 Tflops	1979 Tflops	80GB	3.35TB/s	700W	NVLink：400GB/s PCIe 5.0: 128GB/s
A100 80GB SXM	9.7 TFlops	19.5 TFlops	312 TFlops	80GB	2039GB/s	400W	NVLink：600GB/s PCIe 4.0: 64GB/s
A800 80GB SXM	9.7 TFlops	19.5 TFlops	312 TFlops	80GB	2039GB/s	400W	NVLink：400GB/s PCIe 4.0: 64GB/s
L40s	-	91.6 TFlops	362.05 TFlops	48GB	864GB/s	350W	不支持 NVLink；PCIe 4.0：64GB/s
H20	1 TFlops	44 Tflops	148 Tflops	96GB	4.0TB/s	400W	NVLink：900GB/s PCIe 5.0: 128GB/s
L20 PCIe	-	59.8 TFlops	119.5 TFlops	48GB	864GB/s	275W	不支持 NVLink；PCIe 4.0：64GB/s
L2 PCIe	-	24.1 Tflops	96.5 Tflops	24GB	300GB/s	TBD	不支持 NVLink；PCIe 4.0：64GB/s

资料　：英伟达官网，中信建投

和声明。

2.3 当前国产芯片性能或已接近 A100，或优于 H20

目前 海思、寒武纪、平头哥、壁仞科技、百度昆仑芯、燧原科技、海光等国内 GPU 厂商均已推出用于训练、推理场景的算力芯片，并且持续迭代升级，性能在不断提升。而生态方面，国内 GPU 厂商也推出软件开发包，支持 TensorFlow、Pytorch 等主流框架，并且基于自身的软件建立了开发平台，吸引更多的开发者建立完善生态体系。

国产头部芯片单芯片算力或已接近 A100，或优于 H20。以 FP16 精度为例，国产芯片中 昇腾 910 算力为 256TFLOPS，略低于 A100 的 312TFLOPS，相较于 H100 的 1513TFLOPS 有较大差距，但强于 H20 的 148TFLOPS。此外，平头哥含光 800 在 INT8 精度，壁仞科技 BR100 在 FP32 精度均超过 A100。在单颗芯片峰值算力上，国产芯片已经满足大规模使用条件。随着国产芯片能力的提升，国内算力产业发展将进一步提速。

表 5：国内主要 AI 训练芯片与英伟达主流训练芯片对比

厂商	加速卡	FP32(TFLOPS)	FP16(TFLOPS)	INT8(TOPS)	功耗(W)	显存	显存带宽
英伟达	A100 PCIe	19.5	312	624	400	80GB	1935GB/s
	A800 PCIe	19.5	312	624	400	80GB	1935GB/s
	H100 PCIe	51	1513	3026	700	80GB	3.35TB/s
	H20	44	148	296	400	96GB	4TB/s
海思	昇腾 910	-	256	512	350	-	-
壁仞科技	BR100	256	1024	2048	550	64GB	1.64TB/s
平头哥	含光 800	-	-	825	276	-	-
寒武纪	MLU370-X8	24	96	256	250	48GB	614.4GB/s
摩尔线程	MTT S3000	15.2	-	-	250	32GB	448GB/s
燧原	云燧 T20	32	-	-	300	32GB	1.6TB/s

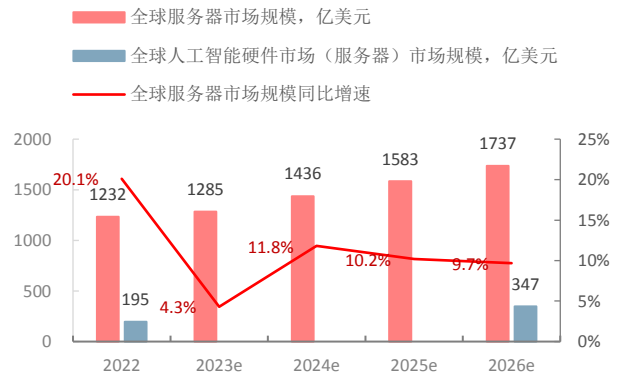
资料 ：英伟达官网，壁仞科技官网，平头哥官网，摩尔线程官网，燧原官网，昇腾官网，人民网，中信建投

三、国产算力产业链环节梳理

3.1 服务器：AI 高增，国产算力芯片发展或带来格局生变

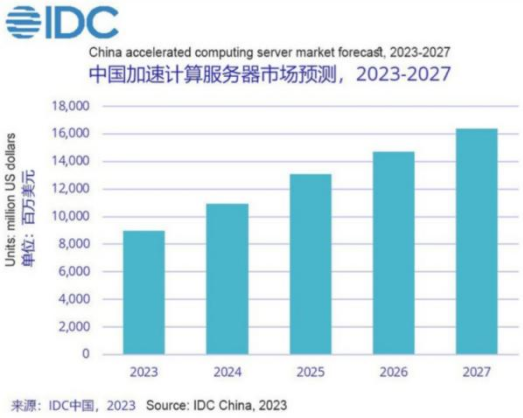
通用服务器相对疲软，AI 服务器高增。受经济持续疲软、高通胀、企业资本支出缩减、去库存等影响，2023 年服务器市场整体出货量不及预期。IDC 数据显示，2023 年第三季度，全球服务器销售额为 315.6 亿美元，同比增长 0.5%；出货量为 306.6 万台，同比下降 22.8%；预计 2023 年全球服务器市场规模微幅增长至 1284.71 亿美元，增长率 4.26%；预计未来四年的年增长率预计分别为 11.8%、10.2%、9.7%和 8.9%。到 2027 年，市场规模预计将达到 1891.39 亿美元。Trendforce 数据显示，预计 2023 年中国服务器需求将同比下降 9.7%。通用服务器受 AI 需求暴涨、全球整机支出向 AI 倾斜影响，通用服务器市场被进一步压缩。IDC 数据，2023 年上半年全球通用服务器市场和 CPU 市场规模均出现下滑，其中二季度 CPU 市场同比下滑 13.4%。AI 服务器高增。IDC 预计，全球人工智能硬件市场（服务器）规模将从 2022 年的 195 亿美元增长到 2026 年的 347 亿美元，年复合增长率达 17.3%。IDC 预计，2023 年中国人工智能服务器市场规模将达到 91 亿美元，同比增长 82.5%，2027 年将达到 134 亿美元，五年年复合增长率达 21.8%。

图 9：全球服务器和 AI 服务器规模预测（亿美元）



数据 ：IDC，中信建投

图 10：中国 AI 服务器市场市场规模预测（百万美元）

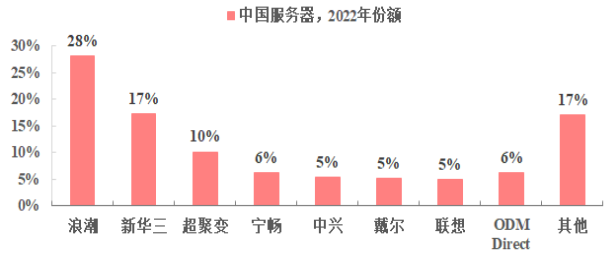


数据 ：IDC，中信建投

AI 服务器占比提升和国产化率提升，国内服务器厂商竞争格局或存变数。此前服务器竞争格局中，浪潮、新华三等厂商份额较高。2022 年中国服务器市场份额来看，浪潮、新华三、超聚变、宁畅、 位列前五，份额分别为 28%、17%、10%、6%、5%。2022 年我国 AI 服务器市场份额来看，浪潮、新华三、宁畅、安擎、坤前、 位列前六，份额分别为 47%、11%、9%、7%、6%、6%。随着国产 AI 芯片占比的提升，AI 服务器供应商格局或存在变化。当前昇腾在国产 GPU 中性能较为领先，国内深度参与 昇腾算力服务器供应的厂商有望更为受益，具体可参考中国电信、中国移动等中标候选人情况。**未来随着国内其他厂商 GPU 新产品的推出以及推理等场景的丰富，国内 GPU 生态也有望更加丰富，进一步可能存在新的变化。**

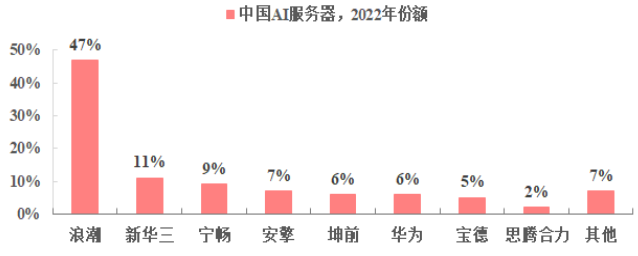
国产化的大趋势下，通用服务器市场格局或同样产生变化。国产化 CPU 服务器中，X86 解决方案目前有海光、兆芯、澜起，并以海光为主；ARM 解决方案有 鲲鹏、飞腾等。中国移动 2021-2022 年 PC 服务器集采中，采用海光芯片的服务器 59982 台占比 20.90%，采用鲲鹏芯片的服务器 58901 台，占比 20.53%，合计整体国产服务器占比高达 41.43%。近期中国移动 2024 年 PC 服务器产品集采项目中，ARM 服务器对比 X86 服务器的招投标数量达 1.71:1，ARM 服务器份额超越 X86。

图 11：中国服务器市场份额情况（2022 年，%）



数据 ：IDC，中信建投

图 12：中国 AI 服务器市场份额情况（2022 年，%）



数据 ：IDC，中信建投

和声明。

表 6：中国电信 2023-2024 年 AI 服务器集采中标情况

标包	排名	企业名称	报价（含税），元
训练型风冷服务器(I 系列)	1	超聚变数字技术有限公司	5,342,130,820.87
	2	浪潮电子信息产业股份有限公司	5,376,127.950.02
	3	紫光华山科技有限公司	5,365,504,587.24
	4	宁畅信息产业(北京)有限公司	5,384,213,138.44
	5	通讯股份有限公司	5,285,535,434.18
	6	烽火通信科技股份有限公司	5,040,734,142.75
	7	联想(北京)信息技术有限公司	5,208,044,679.61
训练型液冷服务器(I 系列)	1	超聚变数字技术有限公司	341,995,767.86
	2	浪潮电子信息产业股份有限公司	344,327,286.69
	3	紫光华山科技有限公司	342,200,460.58
	4	宁畅信息产业(北京)有限公司	342,729,177.41
训练型风冷服务器(G 系列)	1	四川华鲲振宇智能科技有限责任公司	1,304,993,691.62
	2	河南昆仑技术有限公司	1,301,333,577.55
	3	烽火通信科技股份有限公司	1,301,966,880.40
	4	宝德计算机系统股份有限公司	1,308,854,254.13
	5	新华三信息技术有限公司	1,301,879,376.59
	6	湖南湘江鲲鹏信息科技有限责任公司	1,300,296,513.27
	7	北京神州数码云科信息技术有限公司	1,301,661,686.61
	8	黄河科技集团信息产业发展有限公司	1,307,306,069.38
训练型液冷服务器(G 系列)	1	四川华鲲振宇智能科技有限责任公司	1,477,099,321.33
	2	河南昆仑技术有限公司	1,475,413,614.45
	3	烽火通信科技股份有限公司	1,475,489,300.72
	4	新华三信息技术有限公司	1,475,901,834.34
	5	宝德计算机系统股份有限公司	1,484,725,722.97
	6	湖南湘江鲲鹏信息科技有限责任公司	1,474,476,346.12
	7	北京神州数码云科信息技术有限公司	1.476,210,004.55
	8	黄河科技集团信息产业发展有限公司	1,484,431,732.00

资料　：中国电信，c114 通信网，中信建投

表 7：中国移动 2023 年至 2024 年新型智算中心（试验网）采购项目（部分招标情况）

标包	产品	中标\位序	中标候选人	报价（不含税），元	份额
4	特定场景 AI 训练服务器(PCIe 风冷)，52 台	1	新华三技术有限公司	14,723,820	70%
		2	烽火通信科技股份有限公司	14,389,626	30%
5	通用 AI 推理服务器(PCIe 风冷)，16 台	1	通讯股份有限公司	3,629,945	100%
6	特定场景 AI 推理服务器(PCIe 风冷)，64 台	1	新华三技术有限公司	10,661,093	70%
		2	河南昆仑技术有限公司	10,149,566	30%

和声明。

7	数据中心交换机(接入交换机)，10 套 数据中心交换机(出口交换机)，2 套	1	技术有限公司，	3,862,671	100%
8	数据中心交换机(出口交换机)，64 套	1	锐捷网络股份有限公司	67,483,077	25%
	数据中心交换机(接入交换机)，128 套	2	技术有限公司	85,235,400	75%
9	分布式文件存储-性能型典配，96 套	1	技术有限公司	91,458,730	70%
	分布式文件存储-高性能典配，23 套	2	曙光信息产业股份有限公司	83,793,247	30%
10	虚拟化软件，608 许可	1	趋动科技(上海)有限公司	25,182,163	100%
11	特定场景 AI 训练服务器(扣卡液冷)，356 台	1	河南昆仑技术有限公司	490,956,052	72%
		2	四川华鲲振宇智能科技有限责任公司	490,491,527	28%
12	特定场景 AI 训练服务器（扣卡风冷)，106 台 特定场景 AI 训练服务器（扣卡液冷)，1144 台	1	河南昆仑技术有限公司	2,473,721,502	41%
		2	四川华鲲振宇智能科技有限责任公司	2,473,721,365	30%
		3	烽火通信科技股份有限公司	2,473,729,237	20%
		4	神州数码（中国）有限公司	2,473,722,754	8%

资料　：中国移动采购与招标网，中信建投

表 8：中国移动 2024 年 PC 服务器产品集中采购

	产品	中标位序	中标候选人	报价（不含税），元
标包 6	公有云服务器 PC4，2000 台	第一	通讯股份有限公司	77,007,440.00
		第二	浪潮电子信息产业股份有限公司	77,179,600.00
标包 9	PC6，5000 台	第一	浪潮电子信息产业股份有限公司	351,221,050.00
标包 11	C1-Z-ARM，2617 台 C12-Z-ARM，11194 台	第一	四川虹信软件股份有限公司	724,862,246.93
		第二	河南昆仑技术有限公司	724,868,185.66
		第三	黄河科技集团信息产业发展有限公司	724,863,075.59
		第四	武汉长江计算科技有限公司	725,375,739.91
		第五	神州数码(中国)有限公司	724,863,351.81
标包 12	C1-Z-x86，762 台 C12-Z-x86，6584 台	第一	通讯股份有限公司	343,638,579.78
		第二	中移(杭州)信息技术有限公司	353,355,888.96
标包 13	C3-Z，10166 台 C4-Z，193 台	第一	河南昆仑技术有限公司	993,236,142.46
		第二	黄河科技集团信息产业发展有限公司	993,242,410.82
		第三	四川虹信软件股份有限公司	993,236,527.21
		第四	武汉长江计算科技有限公司	993,239,765.84
标包 14	B1-Z-ARM，14384 台 B2-Z-ARM，23520 台 B3-Z-ARM，4039 台	第一	河南昆仑技术有限公司	2,897,396,503.49
		第二	四川虹信软件股份有限公司	2,938,659,339.69
		第三	武汉长江计算科技有限公司	2,938,665,727.88
		第四	宝德计算机系统股份有限公司	2,903,287,752.18
		第五	同方股份有限公司	2,900,340,878.46
		第六	湖南湘江鲲鹏信息科技有限责任公司	2,938,665,035.47
标包 16	S4-Z-ARM，4918 台	第一	河南昆仑技术有限公司	398,087,460.82
		第二	四川虹信软件股份有限公司	405,397,182.58

标包 20	S5-Z，7171 台 S2-Z，3170 台 S3-Z，5066 台 S1-Z，6465 台	第一	四川虹信软件股份有限公司	1,372,327,651.28
		第二	河南昆仑技术有限公司	1,368,964,774.01
		第三	中科可控信息产业有限公司	1,389,929,706.76

资料 ：中国移动采购与招标网，中信建投

3.2 交换机： 网高速产品逐步成熟，高端产品预计实现快速增长

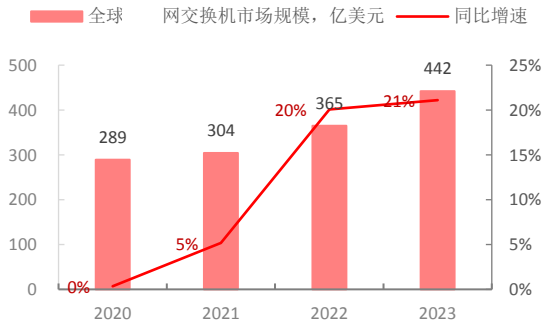
AI 部署需要更大的网络容量，数据中心交换带宽当前处于每两年翻一番的速度快速增长。2022 年 8 月，博通发布 Tomahawk 5，交换带宽提升至 51.2T，serdes 速率达到 100Gb/sec，单通道速率最高达到 800G，可以支持 800G、1.6T 网络部署。下一代交换机带宽将向 102.4T 升级，进一步为 1.6T、3.2T 网络奠定基础。

图 13：光通信行业光口和电口升级迭代示意图

数据 ：Naddod，中信建投

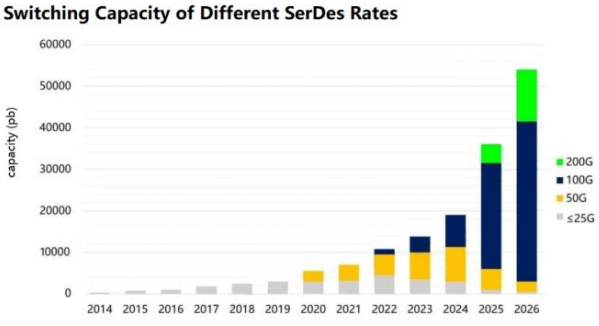
受益于 AI 发展的带动，高端交换机需求快速增长。根据 IDC 数据，2023 年全球 网交换机收入达到 442 亿美元，同比增长 20.1%，其中数据中心部分的市场收入同比增长 13.6%，占整个市场收入的 41.5%，2023 年全年，数据中心部分 200/400 GbE 交换机的收入同比增长 68.9%。工业富联 2023 年年报显示，800G 高速交换机已进行 NPI，预计 2024 年将开始上量并贡献营业收入，预计 2024 年将是 800 Gbps 端口部署的重要一年，预计到 2027 年 400 Gbps/800 Gbps 的端口数量渗透率将达到 40%以上。国内由于政企等需求较弱、高端 AI 算力芯片供应短缺等影响，IDC 数据，2023 年中国 网交换机收入同比下降 4%（2022 年规模近 50 亿美元），但 2023 年四季度同比增长了 9.1%，随着国内算力建设，预计国内高端交换机渗透提升将加速，拉动整体需求。

图 14：全球交换机市场规模（亿美元）



数据 ：IDC，中信建投

图 15：交换机不同 SerDes 速率情况



数据 ：fibermall，中信建投