

计算机行业深度报告

算力服务：从“东数西算”到“东推西训”

方正证券研究所证券研究报告

分析师

张初晨

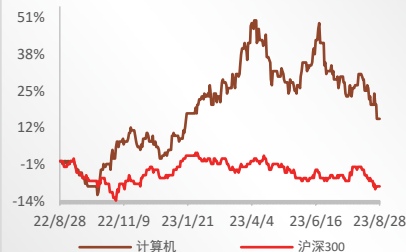
登记编号：S1220523070001

行业评级：推荐

行业信息

上市公司总家数	382
总股本(亿股)	3,439.43
销售收入(亿元)	7,948.78
利润总额(亿元)	-130.30
行业平均 PE	137.62
平均股价(元)	24.82

行业相对指数表现



数据来源：wind 方正证券研究所

相关研究

《具身智能浪潮将至，重视国内产业链机遇》
2023.08.24

《数据库：行业景气百舸争流千帆竞，细分赛道长坡厚雪奋者先》2023.08.03

《AI 赋能之下，电力 IT 投资机会在何处？》
2023.07.24

《ChatGPT Plugins：从超级 APP 到开放平台》2023.04.02

资源整合与统一调度是国家高性能网络计算的重点。当前“东数西算”网络初具形态，规划增量、盘活存量，重建设、重长期有效运营，满足新时期下海量计算、传输、存储需求，是当下国家算力网建设的重点。6月发布的全国一体化算力网调度平台实现跨厂商、跨架构、跨资源池的异构算力调度，对算力网建设具有重大意义。智算网络建设对无线基站、路由器、交换机及光传输接入设备等设备的数据吞吐量和传输速率提出更大的要求，通信设施升级是算力网建设的关键环节。

人工智能向大模型、多模态、多场景发展，智算算力成为当下重要战略资源。三大趋势：

- 1) 下游数智化发展叠加数据要素化，东西部算力新基建的投资价值进一步提升。AIGC 训练侧，总算力需求=模型参数量*训练词数*每个词的运算量，随着国内外千亿级参数量模型不断迭代，所需算力指数级别上升。推理计算算力需求空间相较训练更大。2022 年各行业人工智能指数环比提升明显，预计 23 年底，中国 50%的制造业供应链环节将采用人工智能。同时，数据要素化也将促进算力需求提升。
- 2) 算力租赁释放中小实体算力使用需求，算力报价可作供需刻画的晴雨表。算力需求群体将由小众扩展到千行百业。算力租赁降低众多个体的算力使用门槛和成本，提升总体算力使用效率和社会效益。多地发放算力券，形成政策与行业供需共振的局面。同时，算力报价作为实时性高、透明度强的数据，将作为投资者观察算力供需冷暖的重要指标。
- 3) 智算中心由“资源服务”向“应用服务”转变，产业进一步丰富。人工智能推理及训练是系统性的工作，需要连接、算力、模型开发形成一个全方位的解决方案。各大 IaaS 厂商致力于提供更全面的开发工具及套件，打造开放共赢的智算软件生态。

智算产业链受益公司的范围更广，国产化重要性不断提升。国外对华在超算领域的封锁进一步加剧，国内加大鼓励算力国产化。具体到产业链环节的竞争格局来看：

- 1) 英伟达生态占据主要地位，落地速度较快。英伟达芯片 AI 计算性能当前处于明显领先地位。本季度将出货用于高性能 AI 计算的 GH200、用于通用数据中心的 L40s、服务器参考设计方案 MGX、加速网络平台 Spectrum-X、RTX 工作站等，拥有较强的技术领先性和解决方案丰富度。同时凭借软件 CUDA 生态，软硬协同能力突出。当前英伟达高端芯片仍处于明显供不应求状态，公司具有较高的产业链话语权。国内产业链受益环节主要为高速通信光模块，有较好合作的 AI 服务器厂商，以及 PCB、先进封装等电子元器件企业。
- 2) 国产芯片厂商快速崛起，生态加速丰富。智能计算芯片占训练型服务器成本的 70%，是算力产业链的核心。参数对比看，国内智算芯片硬件性

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

能不断提升，华为海思芯片有超 200 项自主知识产权，研发实力优秀。华为 GPU 单卡算力对标英伟达 A100。同时，以华为昇腾计算生态为代表的国产开放式算力开发生态快速发展。截至 2023 年 5 月，华为鲲鹏和昇腾 AI 开发者已经超过 350 万，合作伙伴超过 5600 家，解决方案认证超过 15500 个。

- 3) 生态完整及供应链能力强的 AI 服务器厂商更加占优。看好擅长整合供应链上游硬件资源，保证关键零部件的供应稳定，同时更好匹配下游市场需求变化的平台型 AI 服务器厂商。华为在云、计算、存储、网络、能源方面较为领先，在企业级路由器市占率 50%，交换机市占率第一，使其在计算集群建设方面更具优势。此外，中科曙光作为中科体系内唯一高性能服务器交付平台，在资源和生态方面亦有较明显优势。

关注标的：中际旭创、工业富联、浪潮信息、紫光股份、中科曙光、海光信息、拓维信息、优刻得、神州数码、华丰科技等。

风险提示：人工智能技术发展不及预期，行业竞争加剧风险，人工智能行业监管细则落地不及预期，公司市场拓展不及预期等。

正文目录

1 产业发展基建先行，东数西算网络初成.....	6
1.1 资源整合与统一管理调度系国家高性能网络计算的重点.....	7
1.1.1 全国一体化算力网调度平台实现多元异构算力调度.....	8
1.1.2 通信技术及网络拓扑结构进步逐渐打破数据时空局限.....	9
2 智算算力成为人工智能发展的重要战略资源.....	10
2.1 技术革命带来算力需求指数上升，新基建投资价值毋庸置疑.....	11
2.1.1 人工智能向大模型、多模态、多场景发展，算力需求增长确定性高.....	12
2.1.2 数据要素化促进算力需求释放.....	15
2.1.3 国内智能算力发展提速，东西部将进一步发挥协同优势.....	15
2.2 算力租赁释放中小实体算力使用需求，算力报价可作供需刻画的晴雨表.....	17
2.2.1 政策利好算力服务增长，行业供给及需求抬升共振.....	17
2.2.2 算力服务报价可作为跟踪行业算力供需情况的重要指标.....	18
2.3 智算中心由“资源服务”向“应用服务”转变，产业进一步丰富.....	19
2.3.1 连接+算力+能力相辅相成，大厂积极布局.....	19
3 产业链受益范围更广，国产化重要性不断提升.....	21
3.1 英伟达生态占据主要地位，落地速度较快.....	22
3.2 算力侧国产化率预计将进一步提升.....	23
3.3 国产芯片厂商快速崛起，生态加速丰富.....	24
3.3.1 国产智能芯片硬件性能不断提升，海思研发实力优秀.....	24
3.3.2 国产芯片厂商软件生态仍在不断丰富，华为较为领先.....	28
3.4 国产 AI 服务器从生态完整性和供应链能力两维度考量.....	29

图表目录

图表 1: 东数西算网络建设规划	6
图表 2: 东数西算“五个一体化”	7
图表 3: 多样化的数据中心决定算力资源网络异构的特点	7
图表 4: 2022 年全国 IDC 市场活跃热力图	8
图表 5: 算力产业发展指数区域分布	8
图表 6: 高效算力的三个要素	8
图表 7: 数据中心算力网络发展趋势	8
图表 8: 国家高性能网络计算的构建	9
图表 9: 不同算力使用场景对时延的要求不同	10
图表 10: 未来计算场景将无处不在	11
图表 11: 人工智能大模型标志第四次工业革命进入快车道	11
图表 12: 智能算力是发展人工智能的核心生产力	12
图表 13: 大模型将 ML 所需算力提升到新高度	13
图表 14: 大模型所需算力指数增长	13
图表 15: 代表性大模型所需算力	13
图表 16: 妙鸭相机微信搜索指数	14
图表 17: 妙鸭相机等待时长仍较长	14
图表 18: 人工智能正快速渗透到千行百业	14
图表 19: 产业 AI 研发不断走高	15
图表 20: 我国数据产量及全球占比情况	15
图表 21: 中国数据交易机构数量明显增长	15
图表 22: 我国在用数据中心机架规模	16
图表 23: 智能算力在中国算力结构占比预计进一步提升	16
图表 24: 中国智能算力规模预测	16
图表 25: 2021-2025 中国 AI 服务器市场规模预测 (亿元)	16
图表 26: 算力调度、算力路由、算力交易	17
图表 27: 各地方政府纷纷发放算力券支持算力普惠	18
图表 28: 智算芯片订单相关新闻枚举	19
图表 29: 商汤 AIDC 意在提供更高性能、更多拓展的 AI 技术系统	20
图表 30: 华为昇思基于昇腾硬件提供从训练到推理的技术支持	20
图表 31: 算力服务生态将进一步丰富	21
图表 32: AI 算力产业链	21
图表 33: 英伟达 H100 训练性能提升约 9 倍	22
图表 34: 英伟达 H100 可将推理性能推高约 30 倍	22
图表 35: 英伟达数据中心业务保持高速增长	23
图表 36: 美国对华科技封锁不断升级	23
图表 37: 寒武纪中标浙东南数字基础设施项目	24
图表 38: 华为联合科大讯飞发布推训一体解决方案	24
图表 39: 服务器成本构成	24
图表 40: 2022Q2-Q4 全球独显 GPU 市场份额	24
图表 41: 华为海思有较优秀的研发实力	25
图表 42: 华为达芬奇 3D CUBE 计算引擎提升 AI 计算效率和灵活性	25
图表 43: 昇腾可提供云、边、端全面的硬件支持	26
图表 44: 海光信息系统集成商客户	26
图表 45: 海光信息产品支持国内外主流操作系统	26

图表 46: 国内外厂商 AI 芯片主要产品性能对比.....	27
图表 47: 昇腾具有较为完整的软硬件生态.....	28
图表 48: 华为昇思开源社区成就生态共赢.....	29
图表 49: 2022 年中国服务器市场份额.....	29
图表 50: 全球 AI 服务器下游出货占比.....	29
图表 51: 中国计算机网络设备上市企业竞争格局.....	30
图表 52: 2022 年中国云服务投入厂商分布.....	30
图表 53: 华为在网络、计算、存储等方面的优势赋能中国算力集群建设.....	30
图表 54: 神州鲲泰获华为计算优秀整机合作伙伴奖.....	31
图表 55: 神州鲲泰服务器入选鲲鹏 2022 年最佳整机奖.....	31
图表 56: 华为昇腾整机硬件伙伴.....	31
图表 57: 曙光致力于打造开放共赢的智算生态.....	33
图表 58: 中科曙光算力产品体系较完善.....	33
图表 59: 中科院投资上市企业枚举.....	33
图表 60: 浪潮信息致力于构建元脑生态.....	34

1 产业发展基建先行，东数西算网络初成

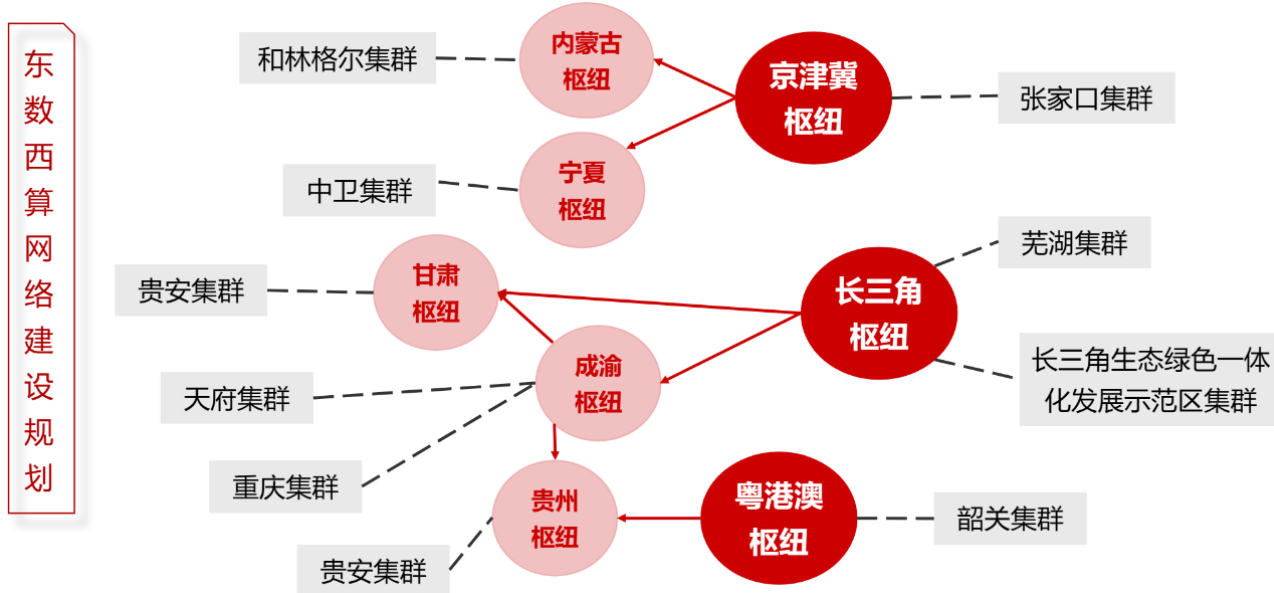
“东数西算”正式启动于2022年2月，由发改委、网信办、工业和信息化部、国家能源局联合印发相关文件，算力作为像水利、电力等生产资源被提升为重要的国家战略发展。依据规划，算力网络包含8大国家算力枢纽节点，10个国家数据中心集群。规划增量、盘活存量，重建设、重长期有效运营。本质是为了满足新时期各行业数智化加速推进所带来的海量计算、传输、存储需求，促进数据中心产业绿色集约发展。

“东数西算”促进东西部资源优势互补、数据流通、价值传递。东部地区经济相对发达，需要处理及运算的数据量更大。西部地区可再生能源丰富，数据中心就近消纳绿色能源，对承接东部算力需求有较大的潜力。

东数西算网络初具形态，节点算力扩充仍在加速。经过一年多的建设，“8+10”的网络形态初具，通过跟踪研究，各节点数据中心的建设仍在快速推进，产业消息催化较为密集。数据中心建设具有投资规模大、链条长、带动效应强的特点，对产业链上下游的投资拉动作用较为明显。

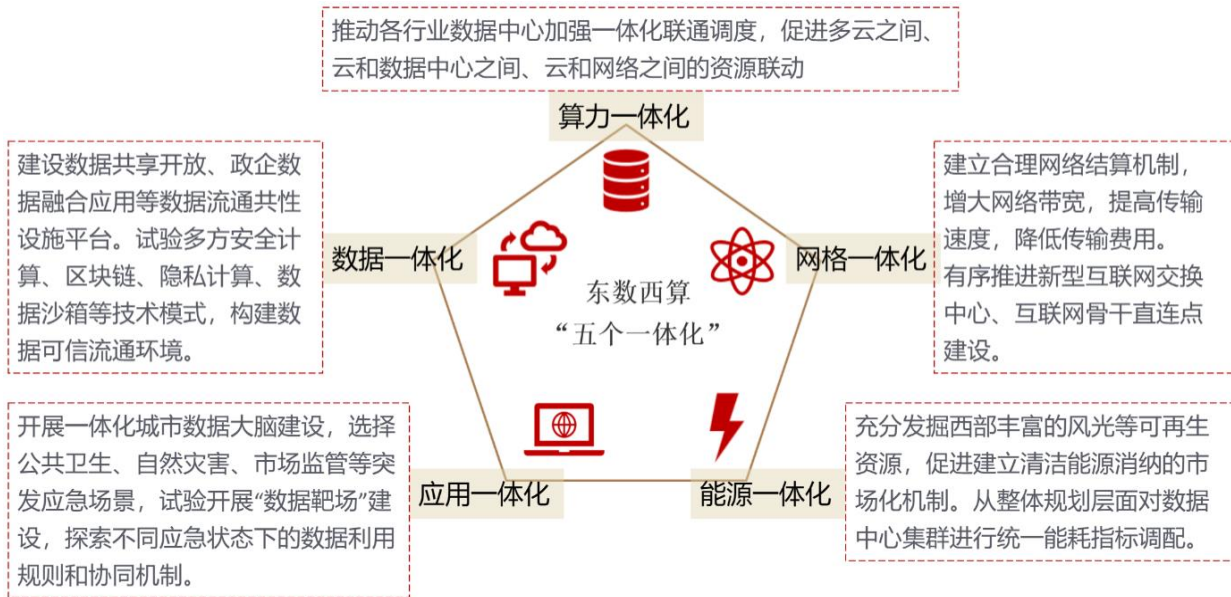
提升节点联动性，实现算力统筹调度亦为当前重要建设方向。当前部分算力资源的使用效率仍不高，算力基建的投入产出比有较大的提升空间。实现全国算力资源总体布局，供求双方更加精准的对接，成为当前阶段重要发力的方向。后续相关政策及实施规划预计将更加密集落地。

图表1:东数西算网络建设规划



资料来源：新华网，华为官网，方正证券研究所

图表2:东数西算“五个一体化”



资料来源：发改委高技术司《“东数西算”工程系列解读之七》，方正证券研究所

1.1 资源整合与统一管理调度系国家高性能网络计算的重点

当前算力网资源呈现分散、异构、动态自治的特点。随着过去多年的算力能力建设，我国存量的算力已经具有一定规模，据2022年中国算力大会，中国智能算力总规模约占全球的26%，服务器规模约2000万台，算力总规模超过150EFlops（每秒15000京次浮点运算次数），排名全球第二。但由于建设主体较为分散，主要由各级政府、不同企业建设，相互之间的协调即联动较为不足。同时，服务器类型、适用的训练种类的差异导致了异构的特点，算力存在度量困难，进而影响构建算力使用的结算体系。

图表3:多样化的数据中心决定算力资源网络异构的特点

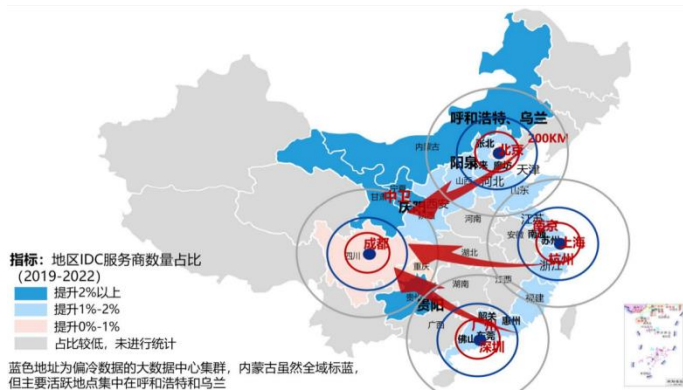
传统的超算中心	主要面向科学与工程计算	支持以双精度浮点运算为核心的计算密集型应用
新型的智算中心	主要面向人工智能应用	计算类型主要是单精度/半精度浮点和定点运算，配置深度学习处理器/加速器
云计算中心	以数据密集型应用为主	配置大容量内存和网络虚拟化设备

资料来源：北京航空航天大学学报《从网络到东数西算：构建国家计算基础设施》，方正证券研究所

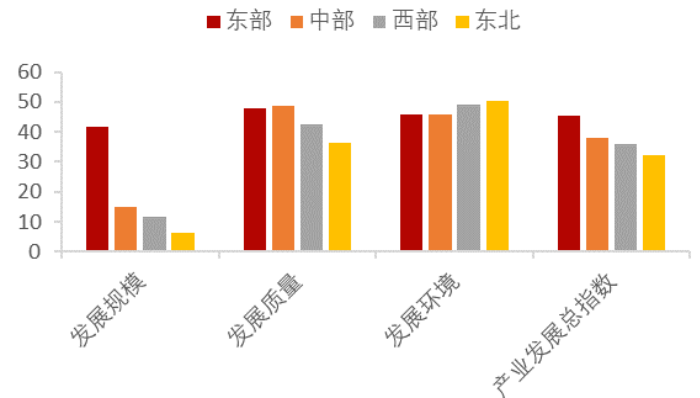
算力布局集群化趋势明显，利用效率仍有较大提升空间。当前全国八大枢纽节点投资约占总数据中心投资的80%，新建约超过90%为大型数据中心。主要原因：1）人工智能时代来临，对大规模并行计算的需求上升；2）头部云厂商新增机柜的60-80%用于云承载，主要集中在核心区域；3）中大型行业客户的专有云需求增长明显，对数据中心集群化有明显需求。同时，当前已建成的算力基础设施使用率仍较低。根据工信部《全国数据中心应用发展指引》，数据中心利用率约为50-60%，且随数据中心规模增大而递减。

中西部地区数据中心发展质量不占下风，规模有较大提升潜力。据国家新型数据中心发布的算力产业发展指数，东部由于发展较早，规模远大于中西部，中西部地区凭借占优的地理环境因素，以及后发优势，在发展质量上不输东部地区。中西部后续进一步扩展算力的潜力较大。

图表4:2022 年全国 IDC 市场活跃热力图



图表5:算力产业发展指数区域分布

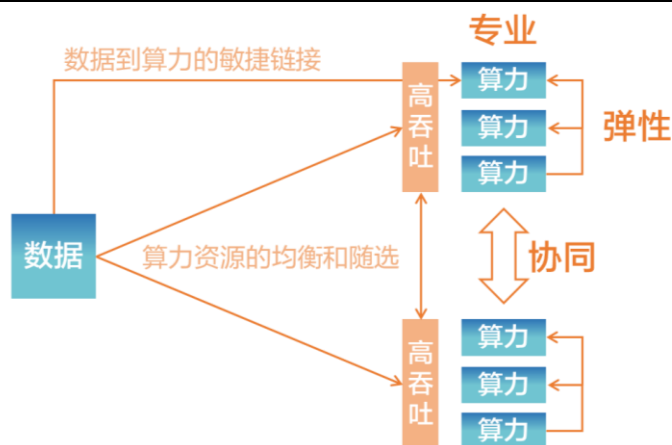


资料来源：中国通服数字基建产业研究院，方正证券研究所

资料来源：国家新型数据中心网站，方正证券研究所

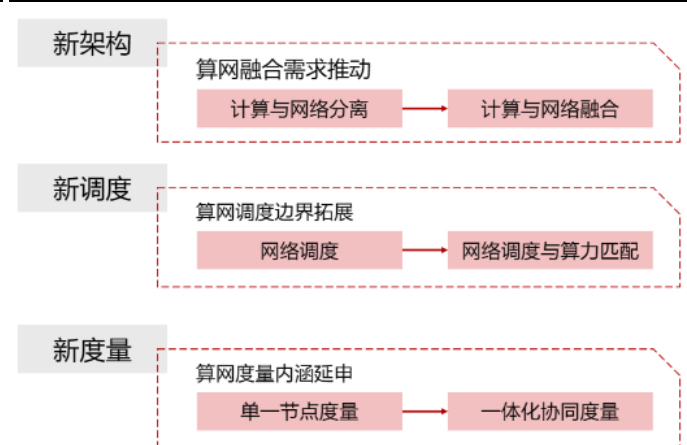
做好算力感知、算力建模及算力评估成为产业及政策制定者关注的重点。算力感知、建模、评估是解决当前算力资源闲置、算力成本较高、算力投资回报不足的关键，同时由于算网资源的天然特点，在实际操纵中并不容易。算力调度亦涉及跨区域、多层次、可视监测、智能运维等，需要各行业以及各级政府协同。鉴于新一届政府将发展数字经济，推动算力建设提升到较高的战略高度，未来相关产业及政策的催化预计也将更加密集。未来各种规模的企业获取来源多样、需求匹配的算力将更加容易。

图表6:高效算力的三个要素



资料来源：华为官网，方正证券研究所

图表7:数据中心算力网络发展趋势



资料来源：中国信通院《数据中心白皮书》，方正证券研究所

1.1.1 全国一体化算力网调度平台实现多元异构算力调度

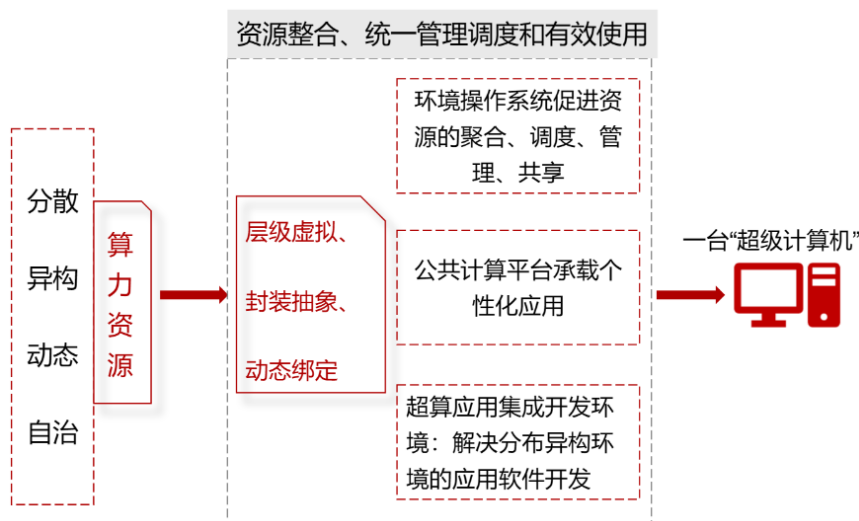
中国首个实现多元异构算力调度的全国性平台落地，直击算网发展痛点。2023年6月发布的算网平台1.0版本在三个方面有明显的突破：1）可以汇聚多种算

力资源，包括通用算力、智能算力、高性能算力、以及边缘算力；2) 可以根据用户的实际需求，通过异构资源池调度引擎，调用通用、智算、超算等不同类型的算力；3) 已接入天翼云、华为云、阿里云等，后续平台资源预计将快速增长。

各地方政府亦开始自发建设算力平台或调度中心，如北京算力互联互通验证平台、上海市人工智能公共算力服务平台、粤港澳大湾区算力调度平台、甘肃算力资源统一调度平台。

异构资源池实现跨厂商、跨架构、跨资源池调度，有效促进智算发展。跨厂商是指不同厂商的异构资源池的算力动态感知与作业智能分发调度。因此，平台将使得算力应用更加充分，实现西部算力使用的集约化。服务器架构根据使用场景不同而存在差异，且处在不断发展的过程中，导致新旧服务器架构的差异，常见架构演进包括使用应用服务集群、数据库读写分离、反向代理和 CDN 加速，分布式数据库和服务等，实现跨架构的资源调度意义重大。同时，以 AI 训练为例，算力需求方可以在智算资源池上进行推理训练，在通用算力资源池进行部署，发挥各类型算力的更大效用。

图表8: 国家高性能网络计算的构建



资料来源：中国计算机学会《从网格到东数西算：构建国家计算基础设施》，方正证券研究所

1.1.2 通信技术及网络拓扑结构进步逐渐打破数据时空局限

受限于传输数据效率，绝大部分“热数据”仍只能就近处理。根据数据的调用频率及时延要求，产业界通常将数据分为“热数据”和“冷数据”。所谓热数据是指需要经常调用，对实时可快速获得性的要求较高，冷数据则反之。因此在网络交换传输仍有待提升的情况下，热数据只能就近部署，当前大部分的热数据仍主要集中在东部地区。

数据网络交换传输时延问题预计将进一步改善。随着通讯协议、标准、网络建设进一步推进，数据网络交换的实时性预计仍有较大提升空间，算力市场供需双方对未来实现更长距离的数据交换传输抱有较大期待。

未来将有更多温数据，以致热数据在西部存储和计算。随着 5G、光通信等新一代通讯技术的普及，算力调度平台进一步完善，接下来数据传输交换读取的便

捷性将大大提高，当前行业对数据传输的困扰将进一步解决，逐步打破数据的时空局限。

图表9: 不同算力使用场景对时延的要求不同

数据类型	平均网络时延	部署建议	平均算力需求	典型业务场景		
热数据（低延时业务）	< 10ms	城区部署	5-10%	金融交易 车联网	直播 物联网	游戏 AI 推理
温数据（时延相对敏感业务）	< 30ms	区域部署	55-60%	智慧城市	智超政务	工业互联网
温冷数据（时延不敏感业务）	< 100ms	算力枢纽 （东数西算）	20-30%	灾备 基因测序	视频转播 大数据	医疗影响 云会议
冷数据（时延不敏感&数据读写频率低业务）	> 100ms	算力枢纽 （东数西算）	10%	数据备份 AI 训练	归档	邮件

资料来源：艾瑞咨询《中国算力网络行业全景洞察白皮书》，方正证券研究所

2 智算算力成为人工智能发展的重要战略资源

AIDC 的兴起就是为了更好配合数字经济和人工智能发展的需要。与以往的“新基建”略有不同，本轮政策层面积极推动算力“新基建”，产业层面人工智能“百模大战”，产业数字化、数字产业化快速推进，预计将呈现政策与产业互为动力，共振加速的局面，算力的需求落地节奏预计比以往，以及很多人想象的要快。

“东推西训”不是完全字面意义上理解将 AI 推理计算放在东部地区，将模型参数训练放在西部地区，更多的关注点在于，在算力网络不断成熟的背景下，各个行业快速增长的算力需求将进一步与智能算力的新增建设形成正向反馈，形成不断的价值循环，算力硬件及软件可以预期增长斜率的进一步提升。

图表10:未来计算场景将无处不在

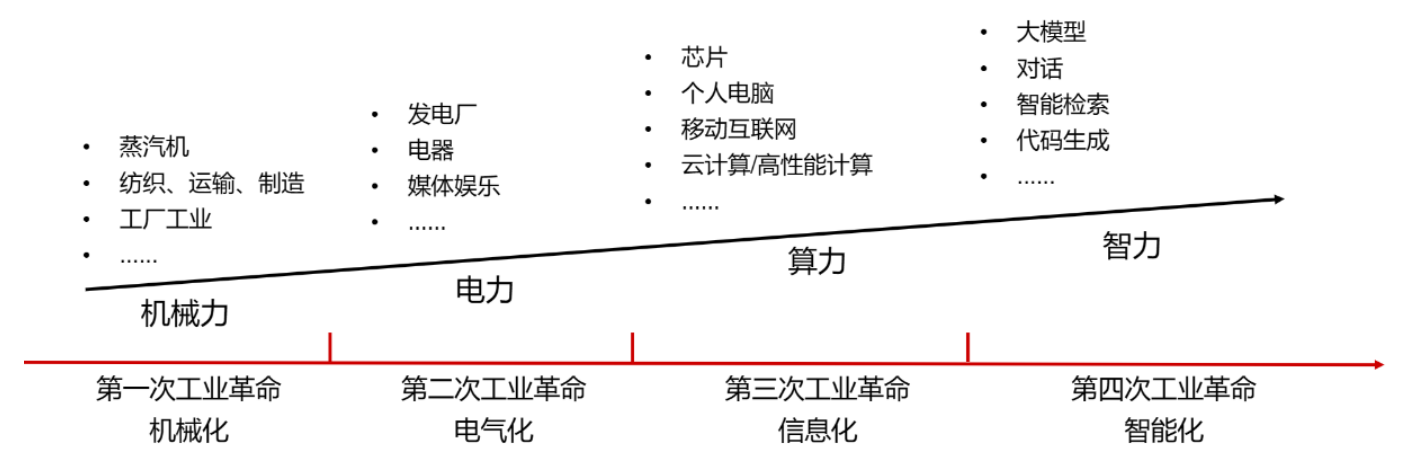


资料来源：华为官网，方正证券研究所

2.1 技术革命带来算力需求指数上升，新基建投资价值毋庸置疑

数字经济与人工智能技术正以前所未及的速度推进与发展，在产业内被视作继信息化革命之后的第四次大规模的科技浪潮。信息化以及人工智能的发展都离不开底层算力的支撑，带来算力需求的井喷式增长。透过众多产业新闻，互联网厂商及运营商牵头正大规模推进智算中心建设，政府层面对于算力基础设施的“新基建”成为新时期各实体行业及资本市场期待与关注的焦点。

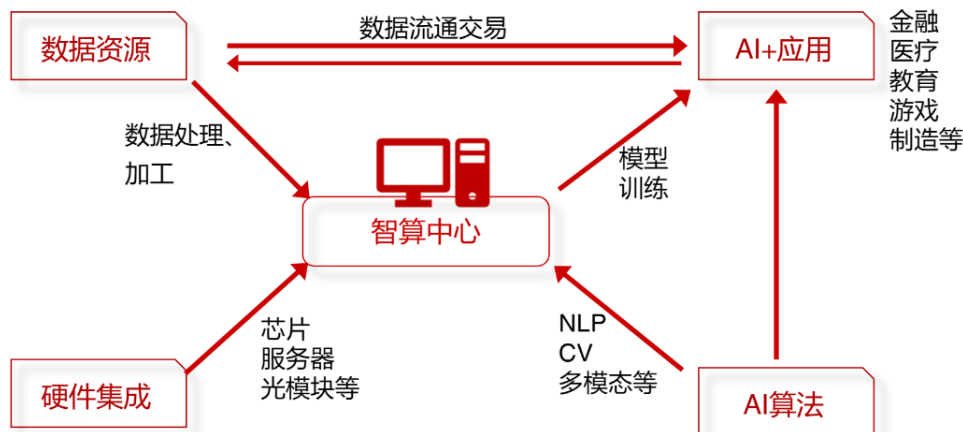
图表11:人工智能大模型标志第四次工业革命进入快车道



资料来源：昇思官网，方正证券研究所

智算中心作为盘活数据资源，整合 AI 算法的载体，向上拉动智算硬件巨大的需求，向下促进 AI 应用繁荣，是 AI 时代重要的战略资源。智能算力是发展人工智能的核心生产力，也是大国竞争的核心竞争力。正如国务院发展研究中心副主任在 2023 中国国际大数据产业博览会上表示，中国市场规模大、应用场景丰富、有制度优势，要加快算力网络建设，在全球竞争中赢得主动。根据 IDC 及浪潮信息联合发布的《2021-2022 全球算力指数评估报告》，国家算力指数与 GDP 走势呈现显著正相关，算力指数平均每提高 1 点，国家的数字经济和 GDP 将分别增长 3.5% 和 1.8%。

图表12: 智能算力是发展人工智能的核心生产力



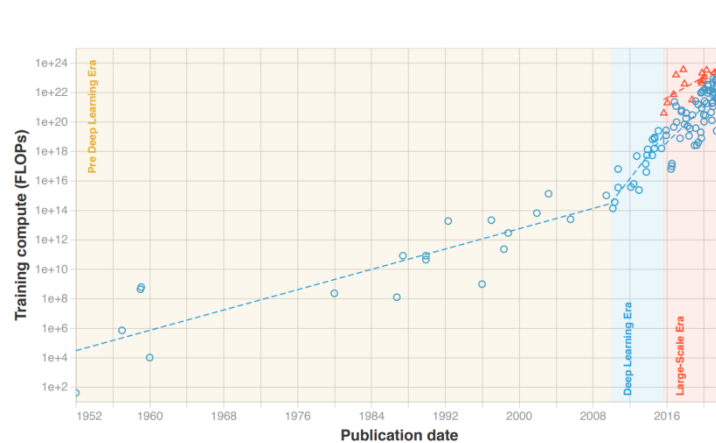
资料来源：方正证券研究所

AIDC 新型算力的投入与产出比，预计相较传统算力进一步提升，投资价值明显。从投资角度看，新型算力所能产生的经济价值预计进一步提升，预计来自于三方面：1) 通过智能调度带来的算力使用效率的提升。2) 基于硬件平台而衍生的一系列附加价值，这可能包括算力平台运营收入，平台软件工具按需收费等。3) 随着算力使用客群结构的改变，单位算力使用带来的效益也预计逐步上升。

2.1.1 人工智能向大模型、多模态、多场景发展，算力需求增长确定性高

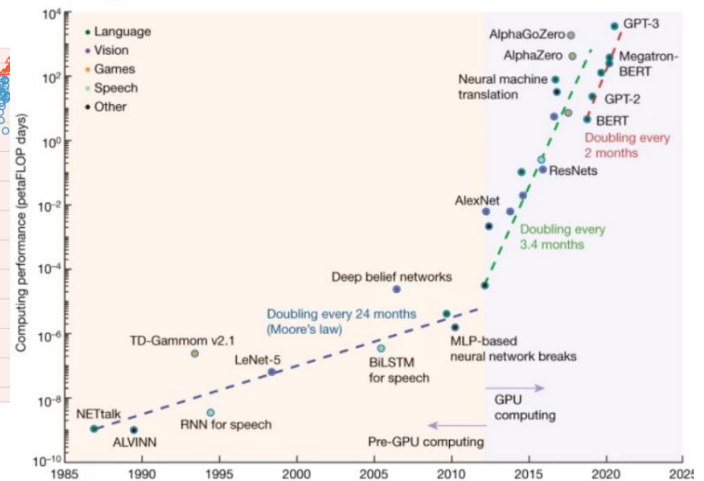
训练侧，人工智能模型参数量不断增长，所需算力指数级别上升。由于每输入一个词都会导致模型所有参数的更新，并且每个词需要消耗浮点算力，因此，**总算力需求=模型参数量*训练词数*每个词的运算量**。1 个词更新 1 个参数，需要进行惩罚及加法各 1 次，总共 2 次浮点运算。模型训练需要反向传播算法，反向传播需要的运算次数是正向传播的 2 倍，因此，训练时每个词的运算次数为 3 次（模型推理每个词的训练量为 1 次），也就是 6 次浮点运算。以 1750 亿参数的 GPT3 为例，带入上述公式计算可得，模型训练所需要消耗的算力就达 $3.14 * 10^{23}$ 次浮点运算（FLOPs）。如果训练使用计算能力为 19.5 TFLOPs 的英伟达 A100，所需 GPU 数量 = $1.8 * 10^{23}$ FLOPs / ($19.5 * 10^{12}$ FLOPs/s * 训练时间秒数)，如果期望训练时间为 15 天，则需要 2048 张 A100 GPU。

图表13:大模型将 ML 所需算力提升到新高度



资料来源:《compute trends across eras of machine learning》, 方正证券研究所

图表14:大模型所需算力指数增长



资料来源:《Brain-inspired computing needs a master plan》, 方正证券研究所

图表15:代表性大模型所需算力

	参数	数据	训练算力	时长
鹏城盘古	200B	300B token	512PAscend910	41 天
紫东太初	1B	1.3 亿图文对	16PAscend910	10 天
紫东太初	100B	300 万图文对	128PAscend910	30 天
GPT3	175B	300B token	2048 卡 A100	15 天
GPT3	175B	300B token	1024 卡 A100	34 天
ChatGPT	175(预训练)+6B(强化)	300B8 token 估算	2048 卡 A100	15.25 天

资料来源:《Language Models are Few-Shot Learners》, CSDN, 方正证券研究所

推理侧市场空间更大, 国产算力芯片有望更早替代。推理侧算力需求取决于问答及模型的复杂程度, 相对较难以一概之。以 NLP 模型来说, 所需 FLOPs $\approx L * \text{模型维度} * \text{模型层数}$, L 是指问题长度加上回答长度。推理侧应用场景广, 算力需求相较训练侧更大。同时由于推理芯片的可定制化程度较高, 芯片性能要求相对训练较低, 因此, 国产芯片厂商有希望率先在推理侧推进国产化替代。

推理侧算力支撑仍为约束应用厂商规模推广及提升用户体验的关键因素。目前智能应用的推出或内测正快速增加, 我们同时看到由于算力的制约, 会出现排队等待耗时较长, 无法大规模推广的问题。OpenAI 在今年 5 月就表示, GPU 算力是制约其向更多用户开放精调 API 的最主要因素。7 月 17 日发布的妙鸭相机作为 AIGC 应用较引起较大的关注, 微信搜索热度最高达 300 万次。根据我们的使用体验, 由于算力的不足, 等待时间长仍为需要重点解决的问题, 8 月 8 日午间等待时长超 4 小时, 7 月 25 日等高峰时段等待时长一度超过 10 小时。

图表16: 妙鸭相机微信搜索指数



资料来源: 微信指数, 方正证券研究所

图表17: 妙鸭相机等待时长仍较长

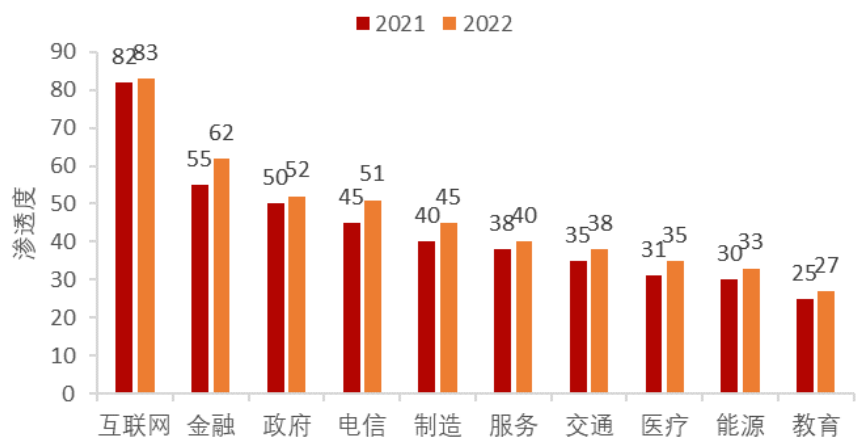


资料来源: 妙鸭相机微信小程序, 方正证券研究所

算力需求更先落地, 订单及业绩可追踪性较强。作为发展大模型及数智化转型的基础, 算力布局及建设具有先导性, 算力厂商在订单和业绩上更早落地, 更容易追踪, 产业链条及业绩的驱动因素更加清晰。

政策与产业需求共振, 智算落地速度存在超预期可能。根据 IDC 的报告, 2022 年国内主要行业智能化渗透度均稳步提升, 2023 年在大模型的推动下, 这一趋势有望加速。智能客服、实体机器人、智慧网点等预计更改广泛应用于服务行业, 预计到 2023 年年底, 中国 50% 的制造业供应链环节将采用人工智能。同时, 据斯坦福 HAI 研究, 当前产业机器学习系统研发成果数量不断走高, 企业层面发力人工智能趋势明显。当前智能算力发展, 上有国家政策大力支持, 下有产业快速增长的算力需求, 我们认为这种共振的格局下, 智能算力产业的增长存在较大超预期的可能。

图表18: 人工智能正快速渗透到千行百业

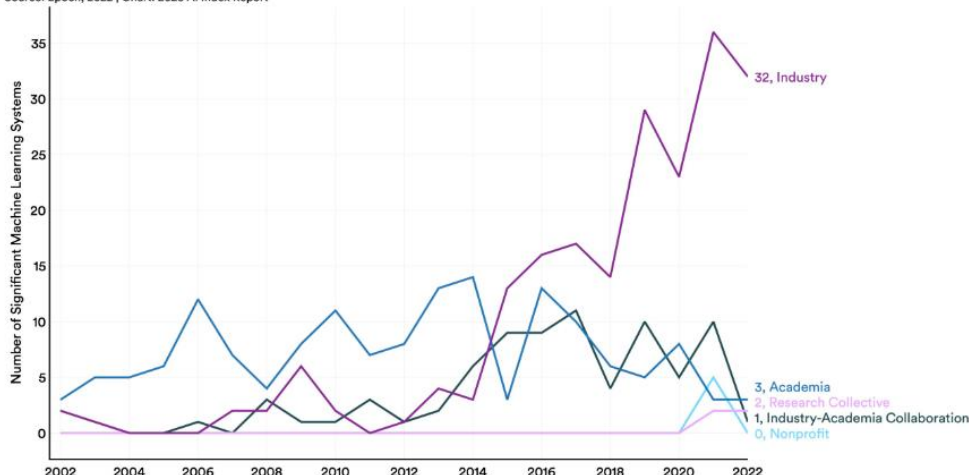


资料来源: IDC 《2022—2023 中国人工智能算力发展评估报告》, cnii, 方正证券研究所

图表19: 产业 AI 研发不断走高

Number of Significant Machine Learning Systems by Sector, 2002-22

Source: Epoch, 2022 | Chart: 2023 AI Index Report



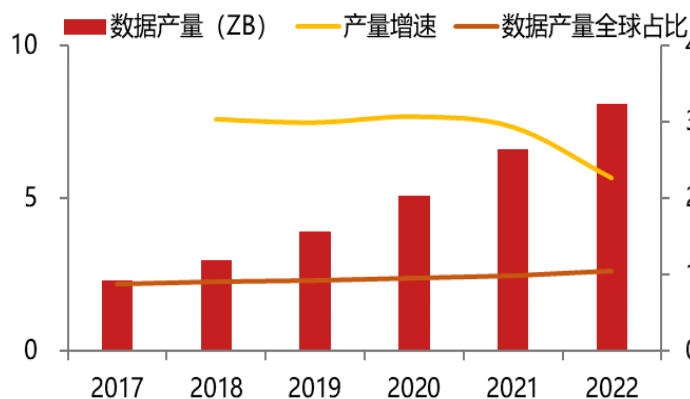
资料来源：斯坦福大学 HAI《Measuring trends in Artificial Intelligence》，方正证券研究所

2.1.2 数据要素化促进算力需求释放

国内优质数据供给量正在进一步加大。根据信通院数据，2022 年我国数据产量 8.1ZB，全球占比为 10.5%，占比相较 2021 年增长 0.6 个百分点。国内人口基数大、市场规模广、产业丰富，随着政府及产业个体对数字化的重视程度提高，未来优质数据的供给量将进一步增大。同时，中国数据交易机构数量今年来保持显著的增长，为数据的高效流通打下基础。

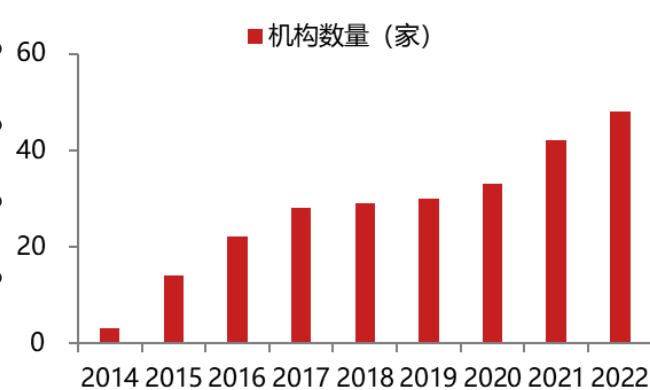
数据要素化推动智能算力需求的提升。主要基于以下原因：1) 数据作为模型发展的原料，数据质量直接影响模型质量。随着社会对数据采集与流转重视程度提高，将有更多高质量数据来发展模型，同时促进算力需求。2) 企业数智化转型中，要求企业基于更大量的内外部数据进行分析来辅助决策，也会产生更大层面的企业侧算力需求。

图表20: 我国数据产量及全球占比情况



资料来源：中国信通院，网络空间研究院，方正证券研究所

图表21: 中国数据交易机构数量明显增长

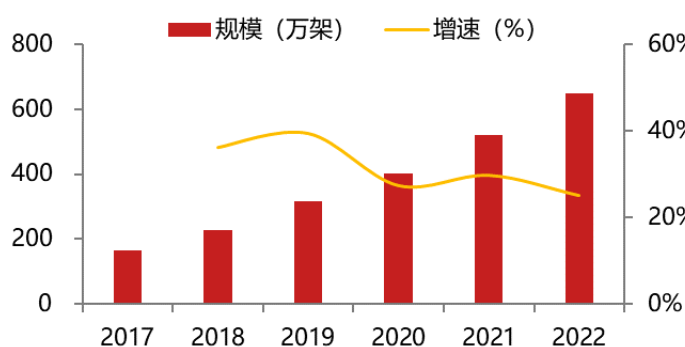


资料来源：中国信通院，方正证券研究所

2.1.3 国内智能算力发展提速，东西部将进一步发挥协同优势

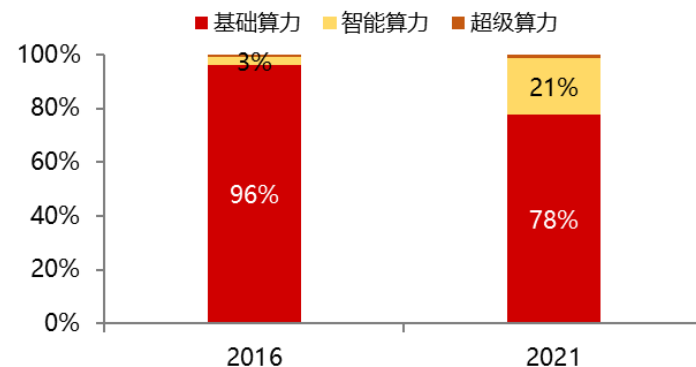
算力节点扩容建设将进一步加速，智算算力占比逐步提升。根据国家网信办《数字中国发展报告（2022年）》，新开工数据中心项目超60个，新建数据中心规模超130万标准机架，西部数据中心占比稳步提高。通过产业信息跟踪可知，当前不论东西部，算力节点建设仍都在加速，包括扩容、提速、现有架构升级，以应对当前算力需求的快速上升。根据IDC预计，中国智能算力规模2022至2026年复合增速为48%。由于当前人工智能应用仍处早期阶段，算力需求仍在快速释放，我们预计增速存在超预期的可能。

图表22:我国在用数据中心机架规模



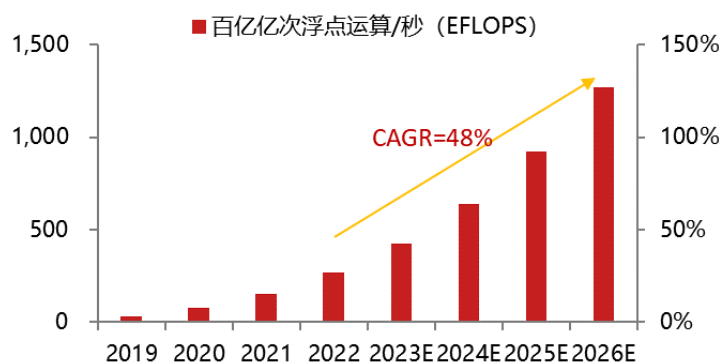
资料来源：国家工信部，方正证券研究所

图表23:智能算力在中国算力结构占比预计进一步提升



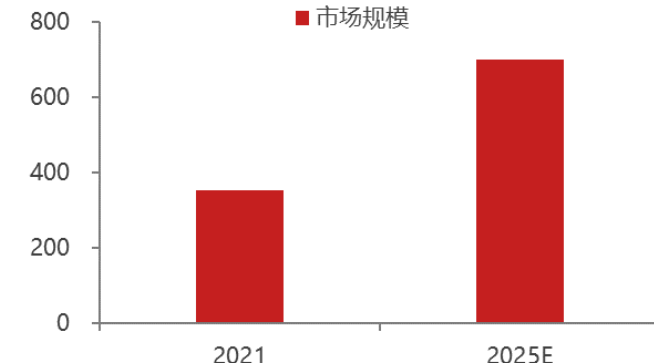
资料来源：中国通服数字基建产业研究院，方正证券研究所

图表24:中国智能算力规模预测



资料来源：IDC, cnii, 方正证券研究所

图表25:2021-2025 中国 AI 服务器市场规模预测 (亿元)



资料来源：IDC, 量子位智库, 方正证券研究所

东部算力供给增长斜率预计将更加低于需求扩张斜率，东西协同发展算力大有可为。正如上一部分提及，东部地区由于自然资源、能源荷载、土地规划、以及运维成本等方面的限制，数据中心建设的局限性较大。同时，特别是在行业早期，数字经济及人工智能发展的东西部差异性更加突出，东部地区快速增长的算力需求与供给的差距将进一步扩大，东部对西部算力的需求也预计只增不减。

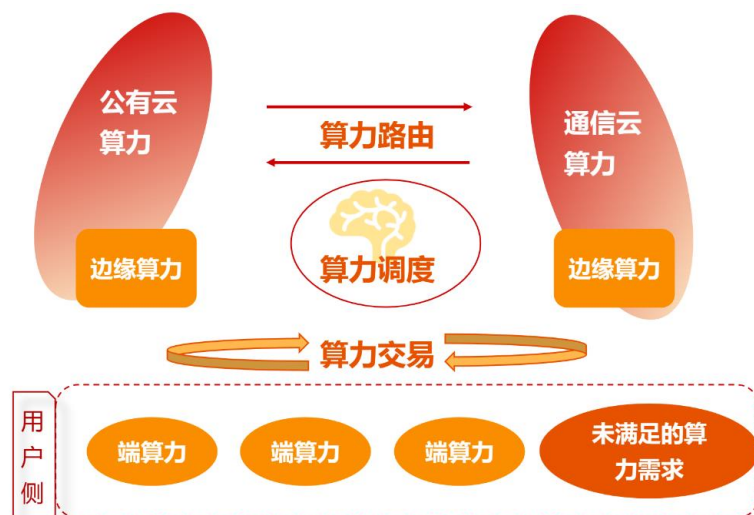
传统上对数据在西部存储的认知多集中在灾备，后续预计更多数据将在西部存储和计算，主要原因为：1) 数据交换及通信技术进步，算力一体化调度平台进一步完善，有效提升东西部算力及数据的实时互通性。因此，当前存储在东部

的“冷数据”将更大规模，更大比例的转移到西部，同时，部分时效及调用频度介于“冷数据”与“热数据”之间的“温数据”，转移的经济性及必要性也将上升。2) 数字经济不断发展，数据量及数据种类将快速膨胀，数据也将作为未来的重要资产，将数据合理、经济的存储成为政府及企业需要思考的课题，西部地区在经济性上更胜一筹。3) 人工智能技术快速发展，模型及应用提供方更倾向于将数据存储与训练集中在一个节点，进而减小传输带来的数据丢失、延迟等问题，提升训练效率。总结看，算网成熟、数据量增大、模型训练提效的需求将共同推动更多数据在西部存储和计算。

2.2 算力租赁释放中小实体算力使用需求，算力报价可作供需刻画的晴雨表

在各行各业数字化发展需求的催生下，适用于更多、更广用户需求的算力产品和服务也将加速落地。算力调度和租赁进一步降低算力的使用门槛和使用成本，抬升算力需求的天花板，是当前众多产业数字化参与主体所期盼的发展方向。

图表26: 算力调度、算力路由、算力交易



资料来源：华为官网，方正证券研究所

2.2.1 政策利好算力服务增长，行业供给及需求抬升共振

需求侧：新技术革命使得算力需求群体，由小众扩展为千行百业。主要基于两点判断：1) 数智化转型不仅针对大型企业，中小型公司在降本增效、提升运营管理水平、开发新的业务增长方面亦有旺盛的需求。传统企业有大量的数据资源由于缺少采集及分析而被浪费，因此，数字化转型需要深挖数据价值，这需要更大算力的支撑。2) 随着业务智能化水平的提高，适用于中小厂商业务的智能化模型的部署需要相应算力的支撑。因此，算力的主要需求方将由原来的科研及国防机构扩展为希望实现数智化转型的千行百业。

供给侧：

1) 存量角度看：与水电资源类似，算力的使用亦存在峰谷。各行各业对算力的使用需求并不是所有时间恒定不变的。以一天的维度看，推理的算力需求较大集中于白天的工作时间，或者晚间休闲娱乐时间，在休息时间需求较少。从中期维度看，一些企业在项目前期搭载的算力资源在项目后期，或者随着企业战

略转型可能面临闲置。将闲置资源变现将为相关企业带来额外收益，减少算力闲置浪费的问题。

2) 增量角度看：由具有资源优势、渠道优势、资金充裕的大型企业建设，通过租赁的形式提供给其他中小型企业及团体，预计将产生较大的市场。资源优势主要体现为数据中心建设用地、政府支持、融资能力等方面，这类公司多为上市公司，拥有较多资源的大型企业。在高端计算芯片短缺的背景下，与主要芯片厂商关系紧密可以视为渠道优势。在算力需求快速增长的背景下，算力租赁市场预计较快增长。

算力普惠普及是政策所向，多地算力券政策相继推出。推动算力建设的同时，使得算力更多服务经济发展，使其成为容易获得的社会资源是各级政府思考发力的重点。年初以来，在中央推动数字经济发展的定调下，成都、上海、北京等各地方政府发放算力券可以视为明显的政策支持风向。最终实现如中国移动负责人在 2021 年世界互联网大会上所表述的，实现“网络无所不达、算力无处不在、智能无所不及”的愿景。

图表27: 各地方政府纷纷发放算力券支持算力普惠

城市	时点	文件&政策
成都	2023/1/12	《成都市围绕超算智算加快算力产业发展的政策措施》，每年 1000 万元的算力券，支持算力中介、科技型中小微企业和创客、科研机构、高校等使用成都智算中心算力资源，按算力合同费用的 50% 支持
上海	2023/5/17	助力中小企业数字化转型，发放“AI 算力券”，重点支持租用本市智能算力且用于核心算法创新、模型研发的企业，最高按合同费用 20% 进行支持
北京	2023/7/14	正在筹划通过算力券等形式支持模型伙伴和模型观察员，首期预计支持不低于 4000 万元算力券，补贴到模型伙伴企业。
杭州	2023/7/13	全市每年设立总额不超过 5000 万元的“算力券”，重点支持中小企业购买算力服务。鼓励头部企业、高校院所开展多模态通用大模型研发并向中小企业开放模型应用
河南	2023/8/5	《河南省支持重大新型基础设施建设若干政策》，每年发放总规模不超过 5000 万元的“算力券”，支持企业、科研机构、高校等使用国家超算郑州中心，按照算力资源使用费用的 20% 予以奖励

资料来源：中国政府网，新华网，中国证券报，河南政府官网，方正证券研究所

2.2.2 算力服务报价可作为跟踪行业算力供需情况的重要指标

模型训练需求扩张推动短期算力景气度攀升，中期看推理算力需求释放的节奏。基于对产业发展特点的研究，由于各家厂商的模型仍处在不断提升推出迭代的过程中，当前算力需求主要集中在训练侧，到明年年初的算力需求增长都有比较好的预期。向后期发展，随着基础模型的成熟，预计大模型的格局将演化为“基础模型+专业模型”的模式，新增通用大模型的算力需求逐步走向收敛。随着训练侧算力增长逐渐平滑，需要推理侧的算力需求接力。在此过程中，亦可能会有由于训练与推理两端需求起量节奏异步，导致总体算力需求短期下滑的风险。

当前 AI 算力核心上游硬件处于供不应求的状态，下游订单量较为乐观。以 AI 服务器最核心的 PGU 芯片的供需情况看，从年初以来，全球各地厂商向英伟达不断追加订单。根据英国《金融时报》8 月 10 日报道，中国大型互联网公司已

经向英伟达订购价值达 50 亿美元的智能芯片。同时，英伟达高端芯片现货价格年初以来由于供应紧张也在不断攀升。据媒体报导，7 月中旬英伟达芯片国内售价达 10 万元/片，单周涨幅超 20%。新华三运营总经理在 MWC 上海展向媒体表示，搭载英伟达高端芯片的服务器供货周期在 30 周左右。

图表28:智算芯片订单相关新闻枚举

时间	新闻
5 月 14 日	英伟达 A100 价格从去年 12 月开始上涨，截至今年 4 月上半月，其 5 个月价格累计涨幅达到 37.5%；同期 A800 价格累计涨幅达 20.0%。同时，英伟达 GPU 交货周期也被拉长，之前拿货周期大约为一个半月，现在基本都需要三个月或更长。
6 月 29 日	甲骨文今年将斥资数十亿美元购买英伟达芯片，以扩展针对新一波人工智能浪潮的云计算服务。此外，埃里森还透露，除了英伟达，他们还将从 Ampere Computing 和 AMD 购买 CPU。
7 月 3 日	英伟达、博通、AMD 陆续上调对台积电订单，整体下单规模较 2023 年至少再增 2 成以上
8 月 7 日	英伟达 H100 涨到 4.5 万美元 海外已经有创业企业开始利用 GPU 进行抵押融资
8 月 11 日	百度、字节跳动、腾讯和阿里已向英伟达订购了价值 10 亿美元的共约 10 万张 A800 芯片，将于年内交付，另外价值 40 亿美元的芯片将于 2024 年交付。
8 月 15 日	沙特和阿联酋正购买数千颗高性能英伟达芯片。沙特通过阿卜杜拉国王科技大学购买了至少 3000 块英伟达的 H100 芯片；阿联酋在研发自己大语言模型 Falcon 同时采购数千块英伟达芯片。

资料来源：金融时报，U. S. News，财联社，澎湃新闻，方正证券研究所

单维度的算力下游订单数据可能会有虚报的成分。2019 年-2020 年的比特币等虚拟货币的挖矿热，以及 2020 年-2021 年的智能电动车缺芯潮中，在关键硬件供给紧张的趋势下，会出现下游需求方报高意向订单量，以保证上游关键紧缺零部件拿货的优先等级，后续如果下游需求放缓，可能会发生砍单的情况。因此，订单数据的跟踪相对低频和滞后一些，很多情况下亦缺乏透明度，需要更多维度的算力需求预警指标。

算力服务价格可作为判断行业算力需求节奏的依据。随着算网建设深入，算力交易市场的完善，算力服务报价将实现实时更新，是可以直观刻画算力供给及需求情况的高频数据。对于市场投资者而言，算力服务报价实时度高、透明度高，可以作为判断算力市场供需热度的重要指标。

2.3 智算中心由“资源服务”向“应用服务”转变，产业进一步丰富

算力“新基建”，新型算力中心绝不是简单理解为提供硬件的概念，为应对数字化、智能化需求，未来数据中心预计通过沉淀的基础通信、人工智能、大数据、区块链、安全认证、精准定位等优质能力，提供统一封装、灵活调用的“能力即服务”，从而促进经济社会效率和效益的提升。

2.3.1 连接+算力+能力相辅相成，大厂积极布局

“能力即服务”是指在算力中心硬件的基础上，通过平台能力，为客户使用算力进行数智化转型提供应用服务。人工智能推理及训练是系统性的工作，需要连接、算力、模型开发形成一个全方位的解决方案。

各大 IaaS 厂商致力于提供整套的算力解决方案以提升竞争力。包括华为、百度、阿里、商汤等大型科技厂商，也包括三大运营商，在提供整套解决方面均有较多的布局。除了提供硬件的“资源服务”之外，在“应用服务”方面的发

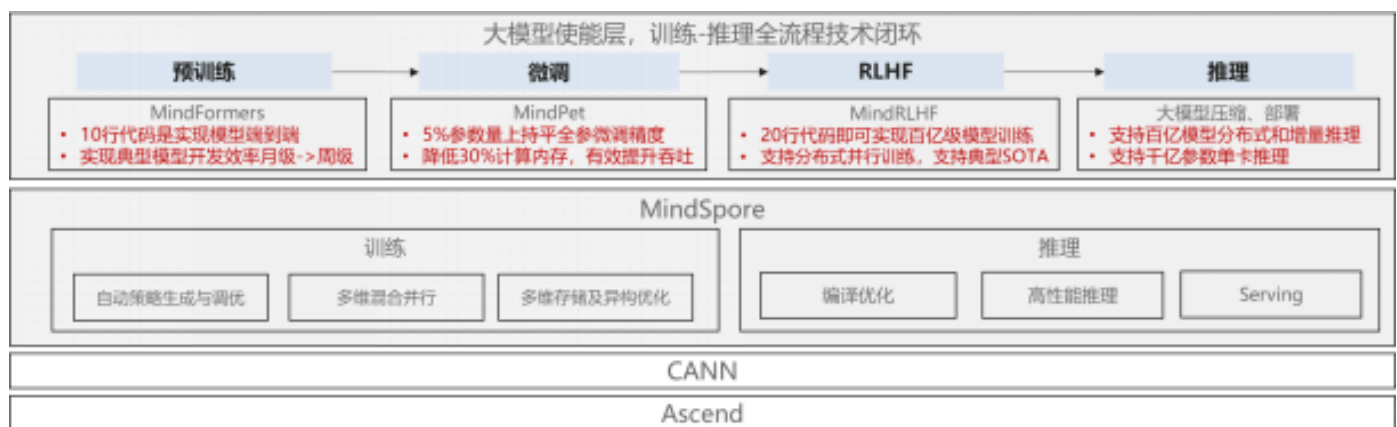
力可能包括算法服务、应用软件优化、算法与软件研发等。以商汤为例，2023年1月AIDC正式启动，致力于实现不同场景的算法模型的底层抽象，以模块化平台套件打造通用型服务平台。主要愿景之一为“扶持一批芯片企业，培育一个软件生态，繁荣一群应用开发”。华为昇思基于昇腾硬件算力，提供工具组件、领域套件和拓展包、科学计算套件，涵盖从模型训练到推理的一整套技术支持。

图表29:商汤 AIDC 意在提供更高性能、更多拓展的 AI 技术系统



资料来源：新浪新闻，方正证券研究所

图表30:华为昇思基于昇腾硬件提供从训练到推理的技术支持

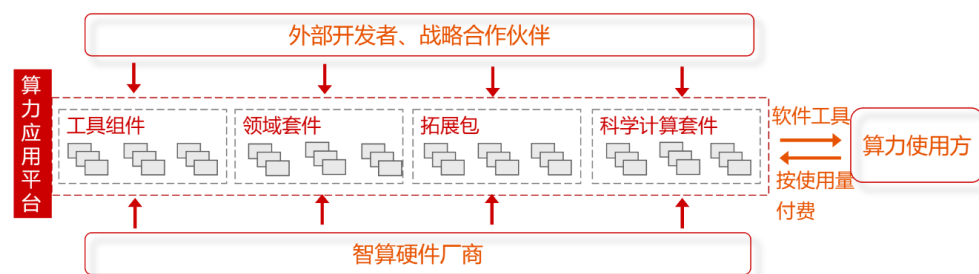


资料来源：昇思 MindSpore 官网，方正证券研究所

预计开辟按使用量收费的新模式，为算力平台和第三方软件开发公司带来新的增长点。各厂商通过在算力平台上建立与运营应用，平台软件按需使用。在为广大算力使用方提供便利，降低使用门槛的同时，算力用户对平台软件的付费意愿将逐步提升。在按软件使用量付费的新模式下，算力平台和软件开发方都将分享这块增量的“蛋糕”。

算力服务生态将进一步丰富，逐步形成正向经济循环。数字经济及人工智能行业仍在快速演进，新增需求不断，需要有更多开发者进入智算超算领域开发软件。人工智能算力国产化不仅需要硬件，更需要软件应用生态配合。随着软件收费模式的跑通，预计将很快形成开发者投入产出的良性循环。

图表31:算力服务生态将进一步丰富

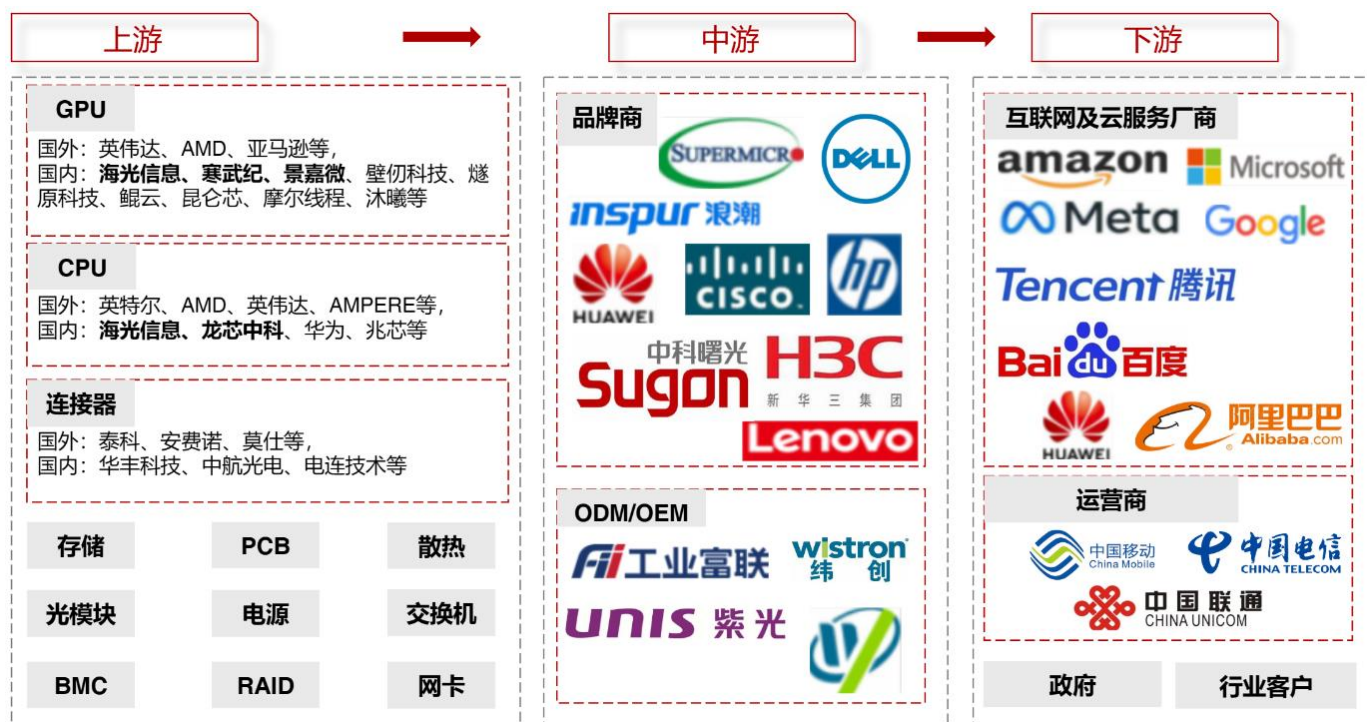


资料来源：方正证券研究所

3 产业链受益范围更广，国产化重要性不断提升

算力产业链上中下游涉及环节较多，AIDC 时代产业链容量预计将进一步扩展。AIDC 下游主要为较为熟知的互联网及云服务厂商、运营商、政府客户、行业客户。中游按照服务器整机提供商类型可以分为品牌商和 ODM/OEM 两种。上游服务器硬件涉及繁多，硬件来看，最核心的为计算推理芯片，包括 GPU、CPU，其余包括存储、连接器、PCB、散热、电源等。软件来看，核心涉及操作系统，数据库。

图表32:AI 算力产业链



资料来源：中国信通院，方正证券研究所

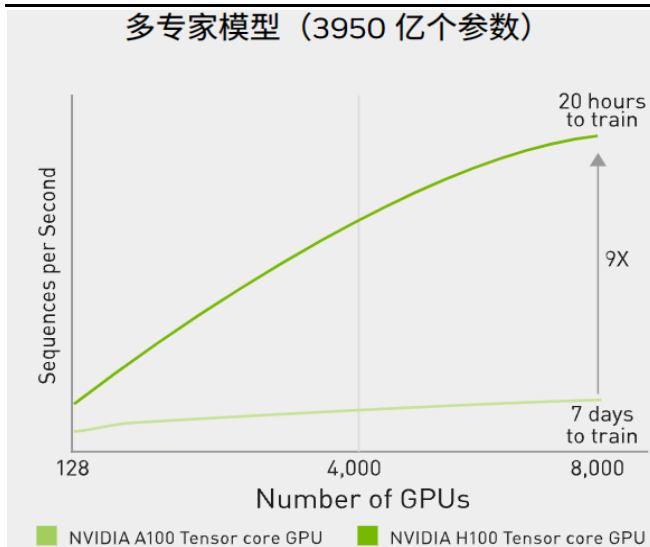
巨大的新增算力需求催生行业巨头的产生。每一轮技术浪潮带来新的机遇都会催生掌握关键技术的巨头产生，例如，个人电脑普及造就了英特尔、微软等巨头，移动互联网时期，国内腾讯、阿里、百度，国外谷歌、苹果等站上风口，当下 AIGC 的发展也将在智能芯片，服务器，算力网络等环节，使得抓住新一轮关键技术的厂商获得快速扩张。

参与方对核心资源的占有能力仍为考量因素之一，但重要性权重相对降低。与上一轮东数西算建设略有不同，当前建设更加考虑产业落地，与各行各业的融合，因此，参与方预计更加多元，更注重产业链竞争力和市场化能力。预计将是地方政府及大型央国企牵头，能够支撑具体数智化场景，更高效运营的厂商共同参与的格局。更好匹配行业发展需要，具备优秀运营管理能力的厂商将不断走出来。

3.1 英伟达生态占据主要地位，落地速度较快

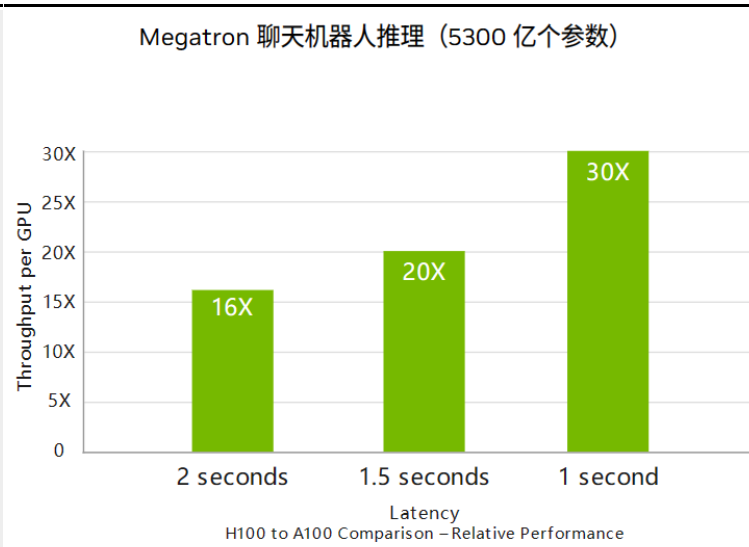
英伟达高端芯片性能领先，产品解决方案丰富。英伟达新一代 GPU 芯片 H100 较上一代 A100 在训练性能提升约 9 倍，推理性能提升约达 30 倍，使用 NVIDIA® NVLink® Switch 系统可连接 256 个 H100 来加速百亿亿级工作负载，另外可通过专用的 Transformer 引擎来处理万亿参数语言模型。第三季度将出货用于高性能 AI 计算的 GH200、用于通用数据中心的 L40s、服务器参考设计方案 MGX、加速网络平台 Spectrum-X、RTX 工作站等，拥有较强的技术领先性和解决方案丰富度。

图表33: 英伟达 H100 训练性能提升约 9 倍



资料来源：英伟达官网，方正证券研究所

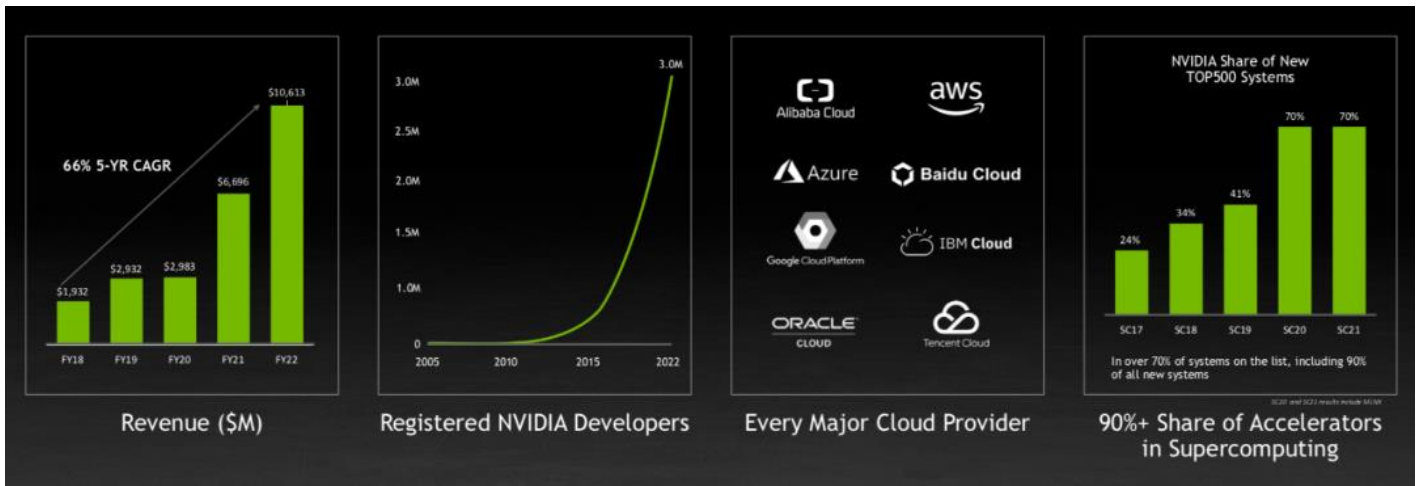
图表34: 英伟达 H100 可将推理性能推高约 30 倍



资料来源：英伟达官网，方正证券研究所

业绩持续高增验证算力需求确定性。根据英伟达 2024 财年第二季度财报，收入同比上升 101% 至 135.1 亿美元，超市场预期，第三季度指引收入 160 亿美元，预计持续高增。收入贡献最大的为数据中心业务，2024 财年第二季度收入 103.2 亿美元，同增 171%。

图表35: 英伟达数据中心业务保持高速增长



资料来源: 英伟达业绩说明会, 方正证券研究所

3.2 算力侧国产化率预计将进一步提升

国内发展智能算力实则是对软硬件产品的供应商提出更高的要求, 包括供应链安全可靠, 快速响应能力及国内生态的落地能力。今年, 中美两国在高科技领域的摩擦不断升级, 美国限制涉及高性能计算的产品及技术对华出口, 确保中国不落后于智能化发展, 智能算力国产化迫在眉睫。在这一过程中, 国产优质厂商预计将凭借政策支持, 产业扩容, 后发优势等因素, 获得更多产业链联合研发机会, 项目定点落地, 并最终加速研发成果在报表端的体现。

图表36: 美国对华科技封锁不断升级

时间	事件概述
2022 年 7 月 27 日 - 28 日	美国参议院和众议院通过“芯片与科学法案”, 强制规定获得美国政府补贴企业不得在中国投资或扩建先进制程的半导体工厂。
2022 年 8 月 31 日	美国芯片厂商英伟达和 AMD 收到通知要求其停止向中国出口 2 种尖端人工智能芯片, 分别为英伟达 A100 与 H100, 以及 AMD 的数据中心级 GPU MI100 与 MI200。
2022 年 10 月 7 日	美国商务部公布了一系列全面出口管制措施, 包括将某些先进、高性能的计算机芯片和含有此类芯片的计算机产品加入管制清单, 对最终用途在中国的超级计算机或半导体开发及生产应用项目增加新的许可证要求等。
2022 年 12 月 15 日	美国商务部宣布将 35 家中国企业和研究机构以及位于日本的 1 家中资企业加入“实体清单”, 实体主要涉及人工智能芯片、半导体装备、航空航天等行业。
2023 年 7 月 25 日	美国参议院通过“2024 财年国防授权法案”(NDAA) 修正案——“对外投资透明度法”, 要求美国企业向联邦机构通报对中国技术领域的投资情况, 包括人工智能、先进半导体、卫星通信和量子计算等。
2023 年 8 月 10 日	拜登签署行政命令, 严格禁止美国对中国敏感技术的某些投资, 包括半导体和微电子、量子信息技术以及某些人工智能系统这三个领域。

资料来源: 清华五道口 NIFR, 方正证券研究所

政府项目及国内大厂积极推动国产算力落地。根据工信部印发的《新型数据中心发展三年行动计划(2021-2023)》, 提出要推动 CPU、GPU 等异构算力提升, 逐步提高自主研发算力的部署比例, 推进新型数据中心算力供应多元化。根据寒武纪 6 月 30 日公告, 中标沈阳市大东区城市建设局的沈阳市汽车城新型基础

设施建设项目-智能计算中心项目，中标金额为 1.55 亿元。7 月 26 日寒武纪中标“浙东南数字经济产业园数字基础设施提升工程（一期）”项目，联合体拟中标金额为 7.5 亿元，预计寒武纪所占金额约为该项目总价格的 70%。在 2023 年人工智能大会上，国内大模型领先厂商科大讯飞公布了与华为的合作，表示讯飞星火与昇腾 AI 将联合打造中国通用智能新底座。在政策鼓励下，国产算力落地预计将进一步加速。

图表37:寒武纪中标浙东南数字基础设施项目

近日，公司（牵头方）与中国移动通信集团浙江有限公司台州分公司（以下简称“中国移动”）、浙江省公众信息产业有限公司（以下简称“浙江信息”）组成联合体，参与了“浙东南数字经济产业园数字基础设施提升工程（一期）”的投标。根据浙江政府采购服务平台于2023年7月25日发布的《关于浙东南数字经济产业园数字基础设施提升工程（一期）的结果公告》，公司与中国移动、浙江信息组成的联合体为上述公开招标的中标单位。

一、中标项目的基本情况

1. 项目名称：浙东南数字经济产业园数字基础设施提升工程（一期）

2. 采购人：台州市黄岩置成物业管理有限公司

3.

中标金额：联合体拟中标金额为75,308.1806万元，根据联合体成员单位分工情况，公司作为联合体牵头方，负责智算硬件部分的供货、安装和后续服务，预计公司所占金额约为该项目总价格的70%，实际金额以最终签订的正式合同为准。

资料来源：寒武纪公告，方正证券研究所

图表38:华为联合科大讯飞发布推训一体解决方案

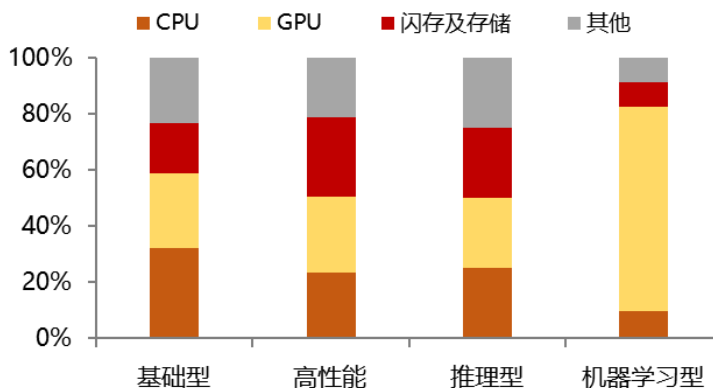


资料来源：澎湃新闻，方正证券研究所

3.3 国产芯片厂商快速崛起，生态加速丰富

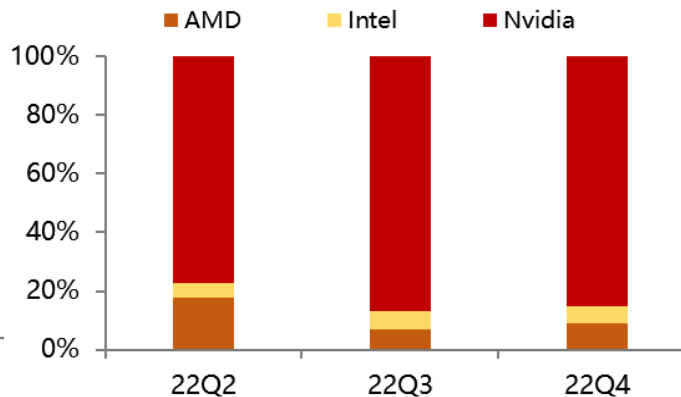
智能芯片占 AI 服务器大部分成本，当前国外厂商仍占绝大部分市场份额。智能芯片在以往高性能服务器中的价值量占比约 50%，在训练型服务器中，CPU 与 GPU 芯片将占据服务器超 70% 的成本，智能芯片厂商占据 AIDC 产业链核心的地位。当前全球独显 GPU 市场主要被英伟达、AMD、英特尔三家垄断。

图表39:服务器成本构成



资料来源：IDC，智研咨询，方正证券研究所

图表40:2022Q2-Q4 全球独显 GPU 市场份额



资料来源：JPR，方正证券研究所

3.3.1 国产智能芯片硬件性能不断提升，海思研发实力优秀

国产智能芯片厂商可以大致分为两类，一类是华为、海光代表的，具有一定从芯片、服务器，到软件生态优势的大型厂商。另一类是以寒武纪、壁仞科技、

摩尔线程、燧原科技等代表的国产智能芯片独角兽企业。两类公司各有优势和特点。

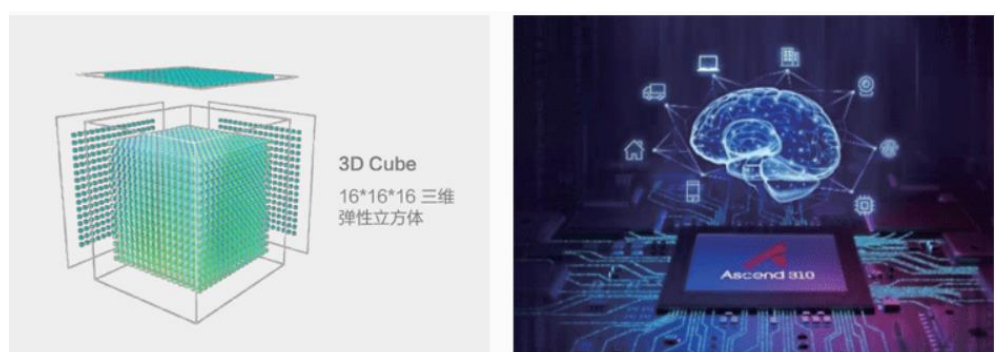
华为海思有超 20 年技术积累，芯片超 200 项自主知识产权，技术实力领先。公司 1991 年启动研发，前身为华为集成电路设计中心，经过 20 余年的发展，成为领先的 Febless 半导体与器件设计公司。目前有智能终端和处理器两大产品线，处理器产品包括麒麟、昇腾、鲲鹏、巴龙、凌霄。目前在全球设有 12 个办事处和研发中心，产品和服务遍布全球 100 多个国家和地区，有超 8000 项专利，研发实力突出。其中昇腾作为新一代智算芯片，有昇腾 310 与昇腾 910 两款产品。昇腾 310 是华为首款全栈全场景人工智能芯片，昇腾 910 是华为目前推出的算力最强的 AI 芯片，昇腾系列后续发展也被各界给予较高期待。昇腾芯片搭载的华为独创达芬奇架构神经网络处理单元 NPU (Neural Network Processing Unit)，该架构具有极具创新的高性能 3D Cube 计算引擎等关键技术，大幅提升 AI 计算的效率和灵活性，能够在多场景，云、边、端，提供最优算力支持。

图表41: 华为海思有较优秀的研发实力



资料来源：海思官网，方正证券研究所

图表42: 华为达芬奇 3D CUBE 计算引擎提升 AI 计算效率和灵活性



资料来源：华为官网，方正证券研究所

图表43:昇腾可提供云、边、端全面的硬件支持

				
AI模块 在端侧实现目标识别、图像分类等，广泛用于智能摄像机、机器人、无人机等端侧AI场景。	AI加速卡 提供AI推理、视频分析等功能，支持检索聚类、OCR识别、语音分析、视频分析等场景。	智能边缘 具有超强计算性能、体积小、环境适应性强、易于维护等特点，可以在边缘场景广泛部署。	AI服务器 具有超强计算性能，可广泛应用于中心侧AI推理、深度学习模型开发和训练场景。	AI集群 具有超强AI算力、更优AI能效、极佳AI拓展等特点，可广泛应用于深度学习模型开发和训练。
Atlas 200 DK 开发者套件 (型号: 3000) > Atlas 200 AI加速模块 (型号: 3000) >	Atlas 300I 推理卡 (型号: 3000/3010) > Atlas 300T 训练卡 (型号: 9000) >	Atlas 500 智能小站 > Atlas 500 Pro 智能边缘服务器 >	Atlas 800 推理服务器 (型号: 3000) > Atlas 800 推理服务器 (型号: 3010) > Atlas 800 训练服务器 (型号: 9000) > Atlas 800 训练服务器 (型号: 9010) >	Atlas 900 PoD (型号: 9000) > Atlas 900 AI集群 (型号: 9000) >

资料来源：华为官网，方正证券研究所

海光信息 CPU 及 DCU 双布局，背靠中科系，具有研发及生态优势。海光信息完全掌握 X86 CPU 完整架构，已具备自主迭代能力。中科曙光作为国产服务器龙头，是公司的第一大股东，有望充分受益中科系软硬件的赋能。海光 7000 系列 CPU 最多集成 32 个处理器核心，最大支持 8 个内存通道和 128 个 PCIe 接口，主要应用于高端服务器，面向数据中心、云计算等复杂应用领域。8000 系列 DCU 深算一号已与国内多家开发大模型的企业和研究所达成认证。海光国产处理器已得到众多 OEM 支持，支持国内外主流操作系统，应用于各类国产服务器及工作站。公司持续保持较强的研发投入，CPU 及 DCU 产品未来预计将进一步受益信创及算力需求爆发。

图表44:海光信息系统集成商客户



资料来源：公司官网，方正证券研究所

图表45:海光信息产品支持国内外主流操作系统



资料来源：公司官网，方正证券研究所

智算芯片独角兽凭借优秀的创始研发团队，及精尖技术的积累异军突起。AI 芯片独角兽企业创始团队大多来自顶尖科研机构及国外科技公司，为公司提供了优秀的研发及技术基因。上市公司主要包括寒武纪，以及众多优秀的待上市企业。

华为及部分 AI 独角兽企业智能芯片硬件性能正逐渐接近国外龙头厂商。通过对国内外智算芯片厂商当前最领先产品关键硬件性能指标的梳理，可以看出华为昇腾 910 芯片处于领先的地位，与英伟达 A100 性能较为接近。此外，寒武纪、壁仞科技、燧原科技、天数智芯等公司的智算芯片性能也处于领先的地位。随着政策支持及产业发展，看好以华为为首的 AI 芯片国产化替代前景。

图表46: 国内外厂商 AI 芯片主要产品性能对比

厂商	型号	芯片类型	显存	显存带宽	INT8 算力	FP16 算力	FP32 算力	总线接口	TDP (功耗)
英伟达	A100	GPU	80GB HBM2e	1935 GB/s	624 TOPS	312 TFLOPS	19.5 TFLOPS	PCIe 4.0 x 16	300W
	H100								
AMD	MI250X	GPU	128 GB HBM2e	3276.8 GB/s	362.1 TOPs	362.1 TFLOPs	45.3 TFLOPs	PCIe 4.0 x16	500W
	MI250X								
华为	昇腾 310	ASIC			16 TOPS	8 TOPS			8W
	昇腾 910								
海光信息	深算一号	DPU	32GB HBM2	1024 GB/s				PCIe Gen4 x 16	350W
	深算一号								
百度	昆仑二代	ASIC	32 GB GDDR6	512 GB/s	256 TOPS	128 TFLOPS			120W
	昆仑二代								
寒武纪	MLU290-M5	ASIC	32GB HBM2	1228 GB/s	512 TOPS			PCIe 4.0 x16	350W
	MLU370-X8								
摩尔线程	MTT S3000	GPU	32GB GDDR6	448GB/s			15.2 TFLOPS	PCIe Gen5 x 16	250W
	MTT S3000								
壁仞科技	壁仞 100P	GPU	64GB HBM2e	128GB/s	1920 TOPS		240 TFLOPS (峰值)	PCIe 5.0 X16	450-550W
	壁仞 100P								
燧原科技	云燧 T20	ASIC	32GB HBM2E	1.6TB/s	256TOPS	128TFLOPS	32TFLOPS	PCIe 4.0 x 16	300w
	云燧 i20								
景嘉微	JM9	GPU	8GB	128GB/s			1.5TFlops	PCIe 4.0 x8	30W
	JM9								
沐曦	MXC500	GPG PU					15 TFLOPS		
	曦思 N100								
天数智芯	天玑 100	GPG PU	32GB DRAM HBM2		295 TOPS	147 TFLOPS	37 TFLOPS	PCIe Gen4.0 x 16	250W
	天玑 100								

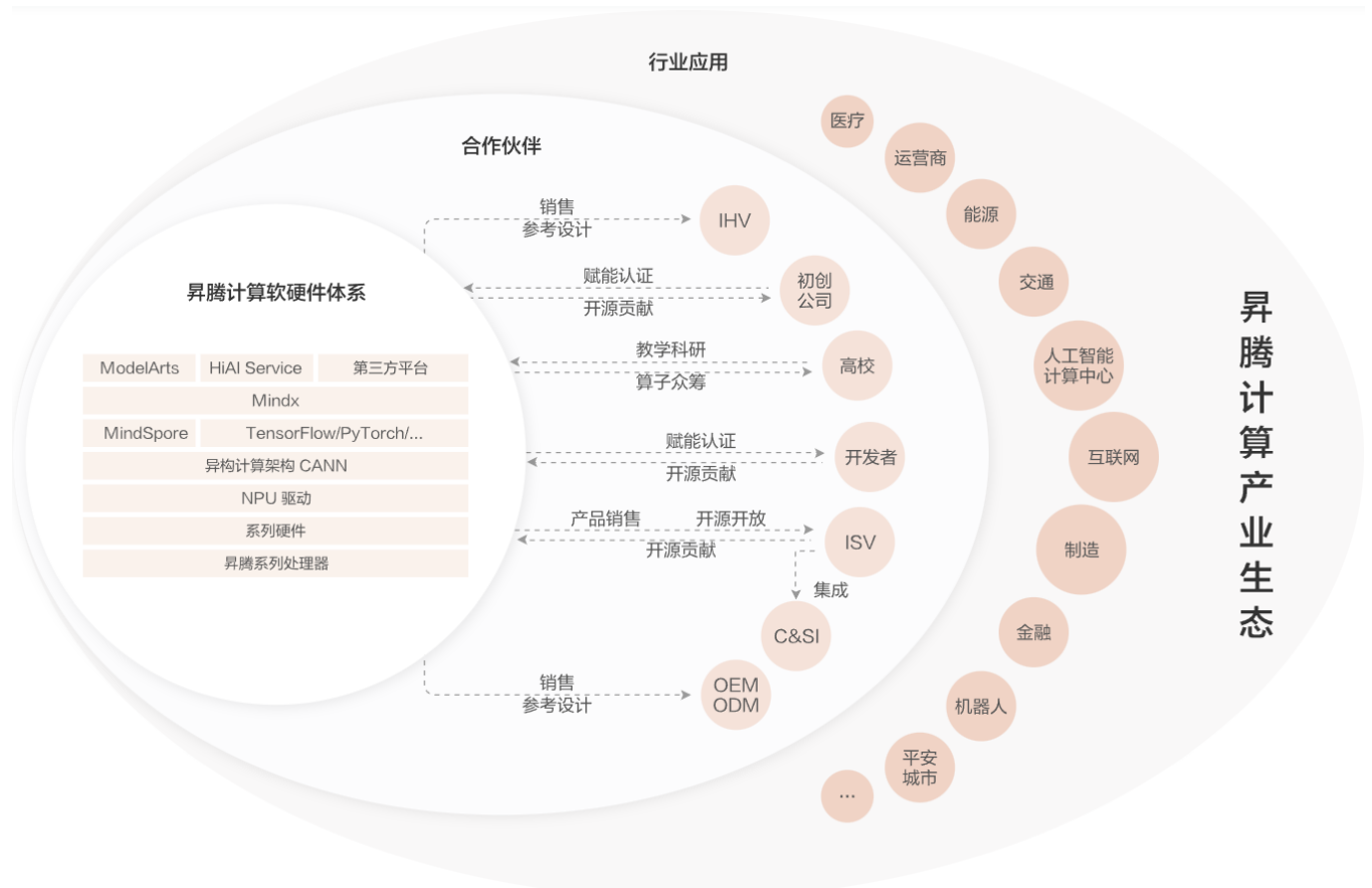
资料来源: 各公司官网, 招股说明书, 方正证券研究所

3.3.2 国产芯片厂商软件生态仍在不断丰富，华为较为领先

CUDA 生态是英伟达关键壁垒之一，国产厂商加速建立自身开发者生态。国外厂商英伟达除芯片硬件性能之外，软件 CUDA 生态亦是其重要壁垒。借助 CUDA，开发者能够利用 GPU 的强大性能显著加速计算应用。CUDA 工具包中包含多个 GPU 加速库、一个编译器、多种开发工具以及 CUDA 运行环境。国内厂商由于起步较晚，在开发者生态上与国外成熟产品仍有差距。在政府推动及下游产业支持下，国产芯片厂商开发者数量逐步上升，生态渐渐形成。

以华为昇腾计算生态为代表的国产开放式生态快速发展。华为在生态建设持开放的态度，在华为建立的全场景开源 AI 框架昇思 MindSpore 外，支持与各种计算框架对接。除自研的应用使能工具 MindX 外，支持第三方平台应用。实现了从模型研发到落地千行百业的全链条赋能。在今年 5 月的鲲鹏昇腾开发者峰会上，华为公布目前鲲鹏和昇腾 AI 开发者已经超过 350 万，合作伙伴超过 5600 家，解决方案认证超过 15500 个。昇腾 AI 基础软硬件平台已孵化和适配了 30 多个主流大模型，覆盖中国一半的原创 AI 大模型，包括鹏城盘古、紫东太初。有 25 个城市基于昇腾构建人工智能计算中心，其中 14 个已经上线并饱和运行。

图表47:昇腾具有较为完整的软硬件生态



资料来源：华为昇腾官网，方正证券研究所

沐曦等 AI 独角兽企业则采取兼容 CUDA 生态的方式，以及与华为等国内大厂共建国产生态的路线。摩尔线程携手浪潮共建元脑生态。在 2023 年人工智能框架

生态峰会上，沐曦科技、壁仞科技作为华为昇思开源社区理事的身份，深度参与昇思生态的构建。

图表48: 华为昇思开源社区成就生态共赢



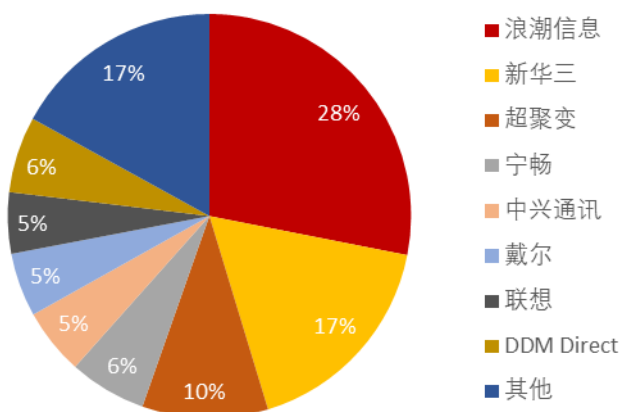
资料来源：新华网，方正证券研究所

3.4 国产 AI 服务器从生态完整性和供应链能力两维度考量

IDC 提供商更加要求一体化方案及生态型运营。服务器厂商负责算力的最终集成和落地，从技术壁垒来看相对上游关键零件较小，因此，AI 服务器厂商的主要优势体现为整合供应链上游硬件资源，保证关键零部件的供应稳定，同时更好匹配下游市场需求变化，为算力客户提供尽可能多的解决方案。当前国内智算中心较为核心的参与厂商包括：1) 华为系，神州数码、拓维信息、东华软件、四川长虹；2) 中科系，中科曙光；3) 海外系，浪潮信息、云赛智联等。

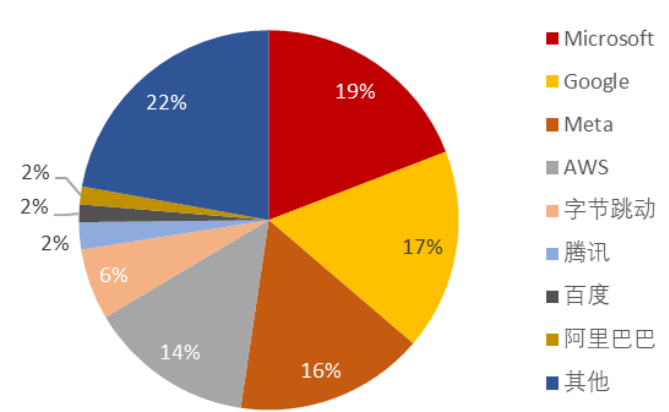
浪潮信息、新华三、超聚变占据中国总体服务器市场超一半份额，其中浪潮占比 28%，新华三占比 17%，紧接着是超聚变，占比 10%。出于统计口径考量，我们认为 IDC 的统计应主要以商用通用服务器为主，应还有较多超算服务器并未在其中统计。从 TrendForce 的统计看，全球 2022 年 AI 服务器的下游主要需求方为北美及国内云和科技厂商。

图表49: 2022 年中国服务器市场份额



资料来源：IDC，量子位智库，方正证券研究所

图表50: 全球 AI 服务器下游出货占比

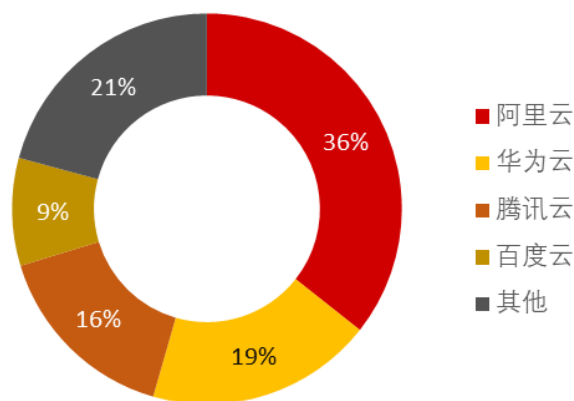
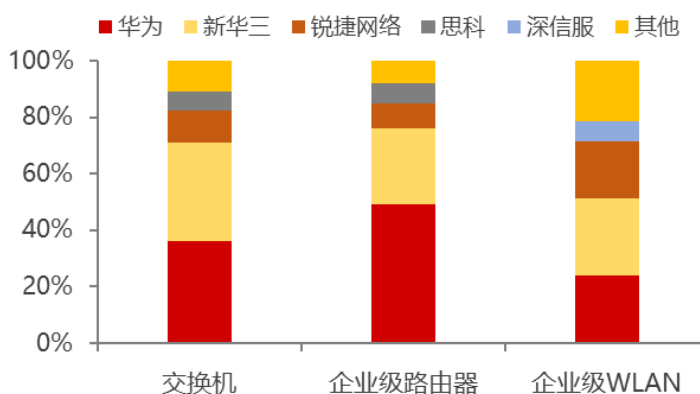


资料来源：TrendForce，方正证券研究所

华为服务器战略合作伙伴有望更多获得华为较全面的软硬件生态赋能，受益华为在云、计算、存储、网络、能源方面的综合优势。华为出身通讯设备厂商，在先进网络通信方面持续引领创新，稳稳处于国内计算机网络设备的第一梯队。根据前瞻产业研究院8月2日发布的报告，中国计算机网络设备上市公司市场格局看，华为在交换机及企业级路由器行业国内市占率处于第一位，企业级路由器市占率50%，企业级WLAN市占率与新华三接近。同时，华为在云基础设施投入上也处于领先地位，2022年占据国内云基础设施投入的19%。华为在算网方面全面的布局，及技术领先性保障了华为总体解决方案的竞争力。在2023年人工智能大会上，华为发布业内首个万卡AI计算集群解决方案，我们认为华为将在后续中国算力网络建设方面扮演重要角色，产业链合作伙伴将更多受益。

图表51: 中国计算机网络设备上市企业竞争格局

图表52: 2022年中国云服务投入厂商分布



资料来源: 前瞻产业研究, 方正证券研究所

资料来源: CanaIys, 方正证券研究所

图表53: 华为在网络、计算、存储等方面的优势赋能中国算力集群建设



资料来源: 搜狐新闻, 方正证券研究所

华为服务器硬件交付合作伙伴主要包括:

1) 神州数码-IT 分销基本盘稳固, 受益信创与智算建设加快。公司与华为同行十余载, 2023年华为中国合作伙伴大会上, 神州数码子公司神州鲲泰获得华为“总经销商杰出贡献奖”、“服务销售突出贡献奖”、“优秀联合营销奖”、“计算优秀整机合作伙伴奖”。作为首批加入鲲鹏计算生态的企业, 拥有自主品牌“神州鲲泰”

系列产品及解决方案，助力政企、金融、运营商、能源等众多行业数字化转型。以神州鲲泰 KunTai R822 服务器为基础打造的金融高性能服务器解决方案入选“2022 年度 Powered by Kunpeng 精选解决方案鲲鹏最佳整机硬件奖”。

图表54: 神州鲲泰获华为计算优秀整机合作伙伴奖



资料来源：新浪新闻，方正证券研究所

图表55: 神州鲲泰服务器入选鲲鹏 2022 年最佳整机奖



资料来源：新浪新闻，方正证券研究所

神州数码是华为首批 AI 服务器合作伙伴之一，是华为整机硬件优先级合作伙伴。未来预计也将明显受益华为智算方案的推广。公司在云管理（MSP）领域可为公司提供全生命周期服务，根据 IDC 发布的《中国云运维管理服务市场（2022 上半年）跟踪》报告，神州数码市场份额位于国内第二，约占 5%。同时，公司 IT 分销基本盘稳固，随着公司逐渐向高毛利产品倾斜，预计总体毛利率将进一步提升。

图表56: 华为昇腾整机硬件伙伴



资料来源：昇腾官网，方正证券研究所

2) 拓维信息-智能计算+开源鸿蒙操作系统战略合作伙伴，绑定华为软硬生态。公司作为华为首批同舟共济伙伴，硬件上负责华为服务器的交付实施，软件上致力于华为 OS 在工业、教育等领域的行业拓展，拥有嵌入式的“霄瀚”全场景操作系统，提供软硬一体的解决方案。公司基于鲲鹏和昇腾处理器开发了兆瀚 RA5900 系列服务器，可以用于智算场景。拓维信息控股湘江鲲鹏 70% 股权，湘江鲲鹏系华为整机硬件优先级伙伴。

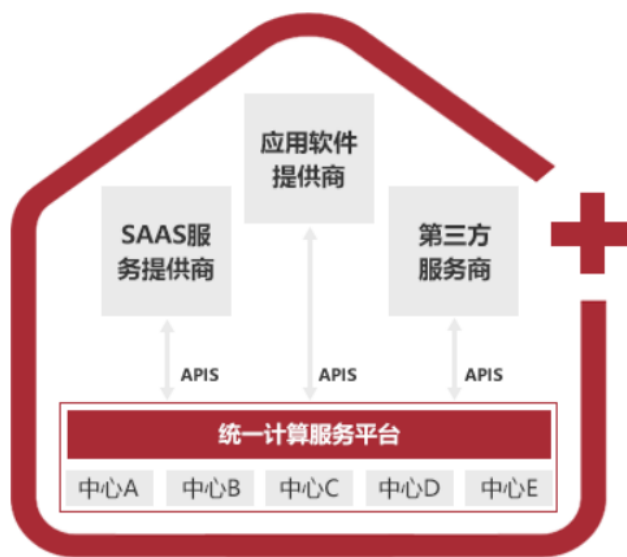
3) **东华软件-大型软硬件代理集成商，参与人工智能中心建设。**软件及信息系统解决方案提供商，主要领域涉及医疗、金融、政务等。AI 服务器方面，公司东华鹏霄一体机可定制化部署各种软件，以及可搭配华为昇腾芯片，适合为大数据、分布式存储、原生应用、高性能计算和数据库等应用高效加速，满足数据中心多样性计算、绿色计算的需求，结合自有数据安全软件能力，持续在智慧城市、医疗、能源、金融等诸多业务领域进行推广应用。东华软件子公司神州新桥于 21-22 年累计算力中标项目超 40 亿元，已中标的智算中心多使用华为芯片，华为昇腾和鲲鹏服务器占 90%以上。

4) **同方股份-昇腾优选级整机硬件伙伴，算力+数据双驱动。**公司与华为鲲鹏、昇腾、存储产品线全面合作，获得华为昇腾产品线优秀整机合作伙伴称号，整机硬件伙伴优选级认证。同方股份和清华大学合作建设中国华北最大的内蒙古超算中心“青城之光”、宁波智算中心、北京昇腾智算中心等项目。公司 AI 服务器具备英伟达和昇腾两条技术路线，2022 年销售的主要为昇腾技术，具备批量出货能力。公司对科研数据库同方知网 100%控股，具有数据资源优势，同时积极探索知识增强型大模型在教育科研、法律医疗等垂直领域的产业化应用。

5) **四川长虹-与华鲲振宇关系为最终控制方的联营企业。**公司与华鲲振宇同为长虹控股集团子公司，四川长虹董事潘晓勇先生担任该公司董事长、法定代表人，具四川长虹公告，两公司关系为最终控制方的联营企业。华鲲振宇作为昇腾领先级硬件合作伙伴，鲲鹏和昇腾整机生态中的出货量位列第一，22 年业绩实现三倍增长。2022 年四川长虹向华鲲振宇采购商品 4.16 亿元，销售商品 0.53 亿元。

中科曙光作为中科体系内唯一高性能服务器交付平台，在资源及生态方面充分受益。1) **在生态方面：完整的 AI 计算服务体系，建设“全国一体化算力服务平台”。**公司具有从 AI 核心组件、人工智能服务器、人工智能管理平台、到软件的多项创新，构建了完整的 AI 计算服务体系。在生态建设方面，曙光亦秉承“共建繁荣、开放、共赢的智算新生态”的理念，支持第三方平台集成、应用集成、接入第三方定制平台，与百度飞桨建立合作伙伴关系。公司布局建设“全国一体化算力服务平台”，将实现多样融合算力的弹性供给，大幅降低算力应用门槛。2) **在技术领先性方面：高端服务器产品全栈自研，存储及液冷方案业内领先。**公司分布式存储方案单套存储容量已能达到 300PB 容量，ParaStor 存储在业内权威存储性能测试 IO 500 榜单中位列 10 节点榜单第一名。公司研发的硅立方浸没液冷技术、Torus 网络架构技术实现了计算技术的绿色节能、集约高效和安全可靠。曙光提出 5A 级智算中心整体解决方案，建设全国多个城市的先进数据中心，覆盖科研、金融、通信、互联网等多个关键行业。

图表57:曙光致力于打造开放共赢的智算生态



资料来源:曙光官网,方正证券研究所

图表58:中科曙光算力产品体系较完善



资料来源:曙光官网,方正证券研究所

图表59:中科院投资上市企业枚举



资料来源:国科控股官网,方正证券研究所

浪潮信息-创造 JDM (联合设计开发) 模式, 互联网客户合作基础较好。1) 浪潮信息采取客户需求驱动的定制化商业模式, 具有较领先的技术平台规划和产品定义能力, 对个性化订单承接能力较强, 对复杂供应链的管理是公司的优势之一。公司 JDM 实现 96%的生产订单自动生成, 98%的物料计划自动决策和执行, 产品交付效率提升 40%。2) 在智算建设方面, 公司致力构建元脑生态。截至 2022 年底, 浪潮信息生态伙伴数量达 2 万余家, 元脑 AIStore 已经入驻了超过 1200 家伙伴, 上线产品和解决方案超过 150 个。基于银行多业务场景常规并发需求, 通过算力资源的细粒度切分, 实现一卡多用。最大化提高算力利用率, 在几乎零性能损失情况下 ($\approx 1\%$), 将算力利用率提升至 3-5 倍。AIStation 是专为人工智能开发和部署打造的端到端平台, 可以实现从模型开发、训练、部署、测试、发布、服务的全流程一站式高效交付, 获得 2023 全球人工智能产品应用博览会「产品金奖」。3) 公司当前主要为基于海外 AI 芯片的服务器方案, 出货一定程度受到海外芯片供应节奏影响, 预计随着供应恢复, 或更多元芯片的服务器方案的推出而改善。此前合作头部互联网客户, 客户基础较好。

图表60:浪潮信息致力于构建元脑生态



资料来源：美通社，浪潮信息，方正证券研究所

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格，保证报告所采用的数据和信息均来自公开合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法存在局限性。特此声明。

免责声明

本研究报告由方正证券制作及在中国（香港和澳门特别行政区、台湾省除外）发布。根据《证券期货投资者适当性管理办法》，本报告内容仅供我公司适当性评级为 C3 及以上等级的投资者使用，本公司不会因接收人收到本报告而视其为本公司的当然客户。若您并非前述等级的投资者，为保证服务质量、控制风险，请勿订阅本报告中的信息，本资料难以设置访问权限，若给您造成不便，敬请谅解。

在任何情况下，本报告的内容不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求，方正证券不对任何人因使用本报告所载任何内容所引致的任何损失负任何责任，投资者需自行承担风险。

本报告版权仅为方正证券所有，本公司对本报告保留一切法律权利。未经本公司事先书面授权，任何机构或个人不得以任何形式复制、转发或公开传播本报告的全部或部分内容，不得将报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其它用途。如需引用、刊发或转载本报告，需注明出处且不得进行任何有悖原意的引用、删节和修改。

评级说明：

类别	评级	说明
公司评级	强烈推荐	分析师预测未来12个月内相对同期基准指数有20%以上的涨幅。
	推荐	分析师预测未来12个月内相对同期基准指数有10%以上的涨幅。
	中性	分析师预测未来12个月内相对同期基准指数在-10%和10%之间波动。
	减持	分析师预测未来12个月内相对同期基准指数有10%以上的跌幅。
行业评级	推荐	分析师预测未来12个月内行业表现强于同期基准指数。
	中性	分析师预测未来12个月内行业表现与同期基准指数持平。
	减持	分析师预测未来12个月内行业表现弱于同期基准指数。
基准指数说明		A股市场以沪深300 指数为基准；香港市场以恒生指数为基准，美股市场以标普500指数为基准。

方正证券研究所联系方式：

北京：西城区展览馆路 48 号新联写字楼 6 层

上海：静安区延平路71号延平大厦2楼

深圳：福田区竹子林紫竹七道光银行大厦31层

广州：天河区兴盛路12号楼隼峰苑2期3层方正证券

长沙：天心区湘江中路二段36号华远国际中心37层

网址：<https://www.foundersc.com>

E-mail：yjzx@foundersc.com