科智咨询

宁夏东数西算产融研究院
Ningxia EDWC Industry and Finance Institute

中国通信工业协会数据中心委员会
China Communications Industry Association Data Center Committee

算力底座之争：智算芯片产业链、竞争态势与趋势研判白皮书（2026）

前言

自2023年以来，以大模型为代表的新一轮人工智能浪潮席卷全球，宣告了智能时代的真正到来。这场技术变革的核心驱动力，并非仅仅是算法的精进与数据的积累，而是一个更基础、更硬核的要素——算力。大模型参数量从千亿迈向万亿级别，对底层计算能力的需求呈指数级攀升，“算力饥渴”已成为制约AI产业发展的最大瓶颈。

算力，正在成为数字时代的“石油”与“电力”。它不仅是科技企业训练大模型、迭代算法的刚需，更已成为大国博弈中衡量国家综合竞争力的关键战略资源。在此背景下，智算芯片作为算力的物理载体，其产业链的完整性、技术路线的自主性以及市场竞争的格局，均牵动着从技术研发到产业应用的全局。

近年来，在全球地缘政治深刻演变与AI技术爆发式增长的双重驱动下，中国智算产业经历了前所未有的挑战与机遇。一方面，美国等国家持续升级对华先进制程芯片及制造设备的出口管制，在光刻机、EDA软件、HBM内存等关键环节形成“卡脖子”封锁，供应链安全面临严峻考验。另一方面，巨大的内需市场、坚定的政策支持以及本土厂商在架构创新（如Chiplet）、系统集群（如超节点）和软件生态上的奋力追赶，正推动国产智算芯片从“能用”向“好用”加速跨越，国产替代已进入深水区。

本白皮书由科智咨询联合宁夏东数西算产融研究院、中国通信工业协会数据中心委员会共同编写，旨在系统梳理智算芯片产业链的全球图景与中国态势。报告从宏观驱动因素切入，深入剖析全球算力规模的增长结构与地缘重构，重点聚焦中国市场的竞争格局——既包括华为昇腾、阿里平头哥、百度昆仑芯、寒武纪、海光等主流厂商的排兵布阵，也涵盖供应链风险、市场竞争风险与地缘政治风险的多维研判。

主编单位

北京科智咨询有限公司

宁夏东数西算产融研究院

中国通信工业协会数据中心委员会

参编单位

克拉玛依市云计算产业园区管理委员会

联想科技（苏州）有限公司

科华数据股份有限公司

博浩数据服务有限公司

泓呵智慧科技（上海）有限公司

江苏云工场信息技术有限公司

联通（广东）产业互联网有限公司

中科智道（北京）科技股份有限公司

广州尚航信息科技股份有限公司

北京数字科智技术有限公司

编制组成员

金和平	黄超	弓瑞峰	张晓雪	李宏图
李建军	丁阔	惠丽丽	毕长顺	许志德
兰满桔	余伟雄	夏常义	黄新	

目录

第一章 宏观背景与驱动因素 **04**

1.1 时代背景：AI大模型爆发与算力饥渴 05

1.2 政策与地缘政治的双重驱动 06

1.3 市场需求画像 08

第二章 全球智算产业的地缘重构 **09**

2.1 全球算力规模与需求展望 10

2.2 大国博弈下的算力供应链重构 11

2.3 中国算力产业发展态势 13

第三章 中国智算芯片竞争态势 **14**

3.1 政策驱动下的国产替代加速 15

3.2 主流国产厂商深度剖析 17

3.3 供应链“卡脖子”与突围路径 22

第四章 关键挑战与风险研判 **23**

4.1 供应链风险 24

4.2 市场竞争风险 26

4.3 地缘政治与政策风险 27

第五章 未来发展趋势研判 **28**

5.1 技术趋势 29

5.2 竞争格局演变 30

第六章 中国算力厂商案例介绍 **31**

6.1 泓呵智慧——某省级电信智算中心改造项目 32

6.2 云工场科技——无锡算力调度运营平台 35

第一章

宏观背景与驱动因素

1.1 时代背景：AI大模型爆发与算力饥渴

1.2 政策与地缘政治的双重驱动

1.3 市场需求画像

1.1 时代背景：AI大模型爆发与算力饥渴

时代背景：AI大模型的“算力饕餮”

自2023年以来，以Transformer架构为核心的大模型开启了人工智能的新纪元。Scaling Law（缩放法则）被推至极致，参数量每增加一个量级，对算力的需求则呈指数级跃升。模型训练已从早期的单机多卡演变为数万卡规模的超大规模集群协同运算。

全球每周Token使用量 *22.7* 万亿次

2026年1月初至3月，全球每周Token使用量从6.4万亿次骤升至22.7万亿次，三个月内增幅约250%。

中国大模型周调用量 *7.94* 万亿次

到2026年5月中旬，中国大模型周调用量达到7.94万亿Token，为美国同期的2.11倍，连续数周稳居全国首位。

2026全球AI基础设施支出 *6580* 亿美元

2026年AI基础设施支出达到6580亿美元，同比增长77%。其中近5000亿美元直接投向服务器、GPU、数据中心和网络等。

2026-2031年全球AI基础设施累计支出 *7.6* 万亿美元

预计2026年至2031年全球AI基础设施累计支出达到7.6万亿美元。其中2026年字节计划投入850亿用于智算芯片采购，阿里宣布未来三年投入3800亿元建设AI基础设施。

核心观点：关于“算力饥渴”的四个真相

1. 算力供应从“过剩”转向“极度结构化短缺”

尽管通用算力看似充裕，但针对大模型训练所需的高端算力却极度匮乏。英伟达H系列长期处于供应紧张状态，而国产替代方案正在处于从“能用”到“好用”的关键跨越期。

2026年的市场将不再仅比拼芯片单体算力，而是比拼万卡互联的稳定性与软件栈的生态兼容性。

2. 供应链安全：从“光刻机焦虑”到“HBM与封装困局”

国产芯片厂商正面临复杂的封锁环境。除了先进制程受限，高带宽内存和先进封装已成为制约国产算力性能释放的新瓶颈。

突围路径：Chiplet技术成为国产芯片的战略选择，通过成熟制程的裸片进行3D堆叠，实现等效先进制程的算力表现。

3. 推理侧将成为国产芯片的“蓝海赛道”

虽然训练侧仍是英伟达的坚实领地，但随着大模型进入应用爆发期，推理算力的需求占比将快速提升。

国产芯片在推理侧具备更高的性价比和本地化支持优势，占领边缘计算和特定行业应用市场。

4. 算力即电力：能源效率是最后的物理防线

大模型的尽头是能源。算力饥渴的本质是能源消耗的剧增。未来的竞争将集中在算力能效比上。谁能用更少的电费完成同样的推理或训练，将拥有商业护城河。

信息来源：科智咨询整理

1.2 政策与地缘政治的双重驱动——政策驱动

国家政策

顶层设计与替代

- **《“东数西算”工程》：**通过国家算力网络布局，引导东部算力需求有序向西部清洁能源富集地区转移，同时要求新建数据中心优先采用国产算力设备，提升整体产业链的自主化水平。
- **《算力基础设施高质量发展行动计划》**：由工信部等六部门联合发布，从计算力、运载力、存储力、应用赋能四个维度设定发展目标，明确提出提升自主可控水平，推动国产芯片在算力设施中的规模应用。

资金与税收支持

- **算力券补贴：**对企业使用国产算力服务给予合同金额10-50%的补贴，降低算力成本。北京、深圳、贵州、浙江、天津、河南、安徽、河北等地已推出算力券，深圳最高补贴比例达50%，需求方年度累计额度可达1000万元。
- **专项基金支持：**国家大基金三期注资3440亿元，重点投向AI芯片、光刻机等卡脖子环节，支持国产AI芯片产业的技术研发和产业发展。
- **运营成本补贴：**对使用国产算力的企业给予用电、用网运营费用补贴。例如四川雅安数据中心提供0.35元/度绿电优惠电价，上海对智算中心PUE≤1.25给予能耗指标奖励。

行业应用与强制要求

- **重点行业强制替代：**2023年2月中共中央、国务院联合发布《数字中国建设整体布局规划》，进一步将金融、能源、军工等领域核心系统适配国产技术（例如金融核心系统国产化、军工全栈国产化）。

地方政策

地区	政策名称	关键内容
北京	《北京市算力基础设施建设实施方案（2024-2027年）》	到2025年智算供给规模达45Eflops，推动智算设施自主化建设，集中建设智算大集群；发放算力券：单次算力券申领不超过合同额20%，年度累计不超过200万元。
上海	《上海市智能算力基础设施高质量发展“算力浦江”智算行动实施方案（2024-2025年）》	到2025年，上海市智能算力规模将超过30Eflops，占总算力的50%以上。具备大规模应用赋能价值的智算中心数量达到10家。
广东	《广东省算力基础设施高质量发展行动暨“粤算”行动计划（2024-2025年）》	到2025年算力规模达到38Eflops，智能算力占比达到50%，到2025年底，新增国产化算力占比达到70%。
河南	《河南省重大新型基础设施建设提速行动方案（2023-2025年》	建设智算中心和郑州城市算力网调度中心，资源利用率达到70%。

未来政策趋势

- ✓ **强化国产化比例硬性要求**：各地在智算中心提高国产算力芯片使用比例，并在政府采购、算力券发放中优先支持国产算力方案；
- ✓ **算力券与资金补贴持续加码**：中央及地方将通过超长期特别国债、算力券等定向支持国产智算需求，补贴比例普遍在合同金额10-50%；
- ✓ **关键技术攻关与产业链协同**：国家大基础及地方专项基金将重点投向GPU/ASIC芯片、光模块等“卡脖子”环节，鼓励技术研发与产业化；

信息来源：科智咨询整理

1.2 政策与地缘政治的双重驱动——中美贸易战

美国出台的芯片限制政策对国内算力芯片市场的影响呈现两面性，短期带来供给压力，长期则倒逼国产算力产业加速自主可控。短期来看，英伟达H200进口受限、海外供应链断供风险加剧，直接导致国内互联网大厂算力采购节奏被打乱，部分高端算力需求无法得到及时满足。但长期发展角度来看，限制政策彻底打破了国内企业对英伟达芯片的长期依赖，推动大厂主动转向国产芯片。同时，外部限制也倒逼国内芯片厂商加大核心技术研发、聚焦超节点、芯片协同架构等关键技术突破，推动国产制程、封装、网络芯片等全链条供应链完善，最终实现核心算力供给自主可控。

中美贸易战加速国产化算力发展

技术封锁初始阶段（2019-2022年）

具体事件说明：

- 2019年，美国商务部将华为、中科曙光等企业列入“实体清单”。
- 2022年7月，美国商务部发布《芯片与科学法案》，禁售14nm以下的半导体设备出售给中国大陆企业。
- 2022年8月，美国政府以“防止被用于军事用途”为由禁止英伟达向中国企业出口A100和H100 GPU卡。

本土芯片厂商进入萌芽阶段，国产化算力迎来初步发展。

技术封锁升级阶段（2023-2024年）

具体事件说明：

- 2023年10月美国商务部工业和安全局（BIS）发布针对芯片出口禁令新规，扩大对华出口先进芯片的管制范围：以“性能密度阈值”为标准管制芯片出口，进而将A100/A800/H100/H800/L40/L40S/RTX4090停止供货。进一步加强高算力卡限制。
- 2024年1月，美国商务部针对“基础设施服务商拟议规则”，禁止国内企业通过云服务获取高端算力。
- 2024年12月，美国商务部新规发布遏制中国公司通过第三方国家获取制裁的GPU卡。

AI芯片加速发展，华为、寒武纪、海光等企业逐步扩大市场份额。

全面封锁与精准打击阶段（2025年至今）

具体事件说明：

- 2025年1月，美国商务部发布《人工智能扩散出口管制框架》，对AI芯片实施分级许可制度，中国被列入第三层级的国家被完全禁止出口任何GPU卡。
- 2025年4月，美国政府明确禁止英伟达特供中国市场H20出口，进一步限制高性能AI芯片对华销售。
- 2025年12月，美国允许英伟达H200芯片在收入分成条件下向中国出口。芯片出口前必须在美国进行独立第三方测试。

大陆晶圆厂、封测厂的先进制程&先进封测产能加速扩展阶段。

信息来源：科智咨询整理

1.3 市场需求画像

核心厂商目标客户及关键应用场景

厂商	核心产品	目标客户	关键应用场景
英伟达	H200	头部互联网、AI大模型公司	大模型训练、通用AI训练
华为昇腾	950PR/910C	头部互联网、运营商、政企	逐步渗透互联网智能对话业务、智能代码开发业务场景等，并作为互联网厂商算力租赁服务的算力底座
阿里平头哥	真武M890/810E	阿里自用、智能驾驶、金融	智能体大规模推理、电商推荐
百度昆仑芯	P800/M100	百度自用、运营商、金融	大模型训练（百度文心）、规模化推理
寒武纪	思元590/690	头部互联网、金融、政企等	AI大模型推理、视频分析等
海光	深算系列	政府、头部互联网、金融、运营商等	政府数字化系统、智能风控等

核心客户市场画像

互联网大厂

- 市场画像：互联网头部大厂采购智算芯片一方面是搭建自有AI算力集群，支撑企业自研智能对话机器人、智能代理助手等自研大模型落地，同时部署各类开源主流大模型完成商业化应用。另一方面搭建自研公有云算力服务平台，对外输出算力服务。
- 智算芯片首选：英伟达、华为昇腾、自研芯片。

运营商

- 市场画像：运营商自身自研大模型储备数量少，几乎不自主研发模型，采购芯片主要用来承接部署第三方成熟商用大模型，业务自主性偏弱。通信运营商采购流程严格受财政预算管控，项目审批与采购落地节奏慢，整体推进效率远低于互联网企业。
- 智算芯片首选：华为昇腾、寒武纪、海光等。

垂直行业（政企/金融/医疗等）

- 市场画像：行业用户的算力需求场景更偏向行业智能体建设、专用算力部署等方向，对芯片的兼容性、运行稳定性以及政策合规性要求更高。
- 智算芯片首选：华为昇腾、寒武纪、海光等。

科研院所

- 市场画像：由于制裁，高端训练芯片极度短缺，科研经费开始转向国产。预算有限，需要大量购买用于实验验证，追求单卡性能和性价比。
- 智算芯片首选：华为昇腾、寒武纪、海光等。

第二章

全球智算产业的地缘重构

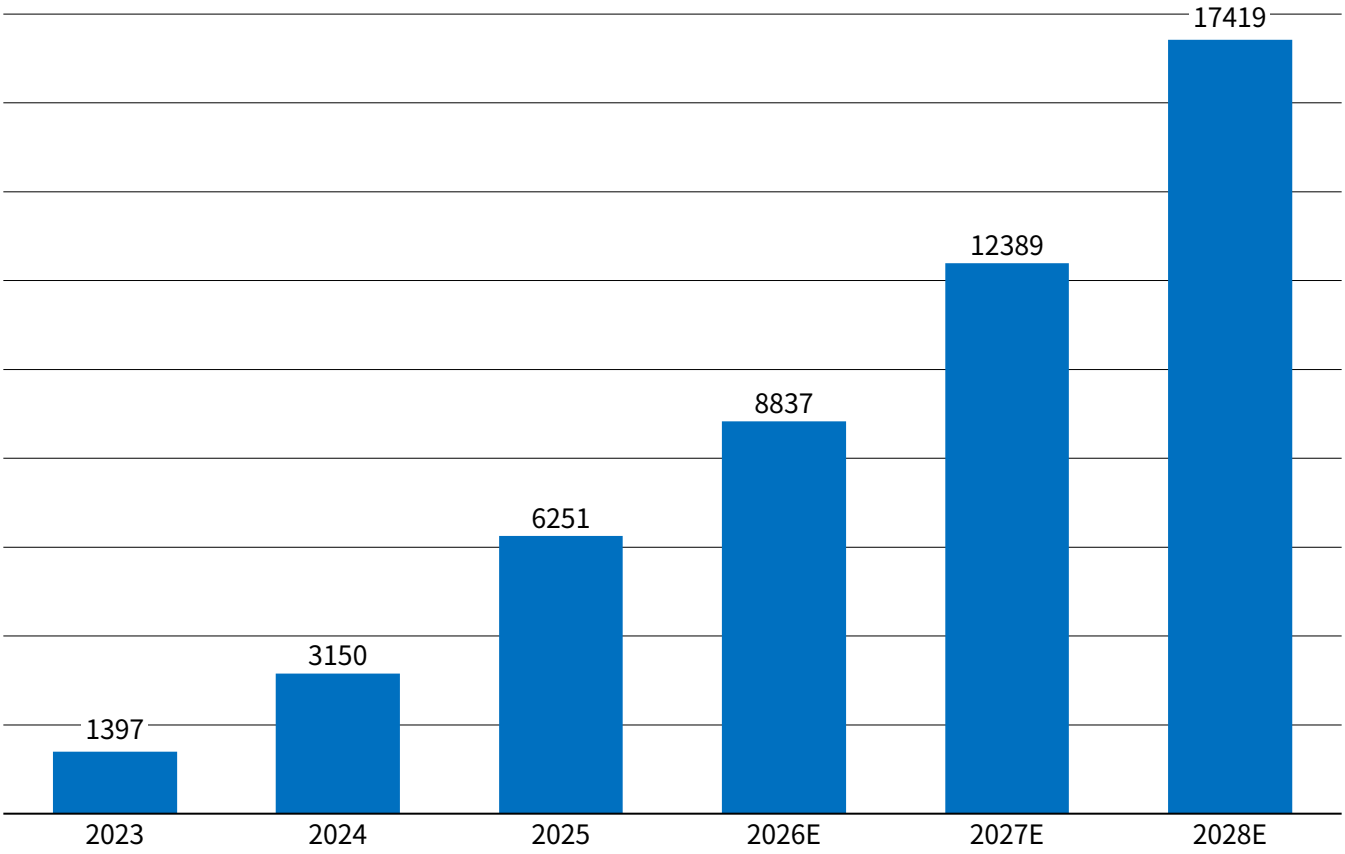
2.1 全球算力规模与需求展望

2.2 大国博弈下的算力供应链重构

2.3 中国算力产业发展态势

2.1 全球算力规模与需求展望

2023-2028E全球算力规模（EFLOPS）



算力需求结构变化：推理成为绝对主导

- 2023-2024年：训练算力需求占比高（约60%），主要受大模型预训练驱动。
- 2025-2026年：随着大模型部署成熟，推理算力需求首次超过训练算力，占比达到55-60%。
- 2027-2028年：推理算力占比将进一步提升至70%以上，主要原因为AI应用从“尝鲜”进入“全场景落地”，智能客服、推荐系统、视频生成、自动驾驶等推理算力需求增长。2028年，全球推理算力需求将是训练算力需求的3-5倍。

2028年展望：算力即“石油”

- 总量目标：2028年全球总算力规模有望突破17000EFLOPS。
- 结构目标：推理算力将彻底超越训练算力，成为市场主流（>70%）。
- 中国角色：中国将边缘计算、垂直行业应用、国产算力生态领域形成全球竞争力，成为全球算力增长的核心引擎。

信息来源：科智咨询整理

2.2 大国博弈下的算力供应链重构——区域格局变化

美国凭借超大规模云服务商的先发优势占据算力基础设施规模和芯片设计龙头两座高地，中国在智算芯片国产替代和系统级创新两个维度上的追赶速度超出预期，而欧洲、亚太、中东、非洲等区域正围绕“主权算力”这一共同主题探索差异化路径。

中国：全球智算的“独立一极”

- ✓ 技术路线：从“单卡性能”转向“集群系统”。通过HCCL互联协议和CANN软件栈，构建独立的“中国版CUDA”。
- ✓ 供应链：依托中芯国际和国内封装厂，解决28nm-7nm成熟支撑的产能问题，通过Chiplet技术突破制程限制。
- ✓ 挑战：高端HBM内存、光刻机仍受制约，软件生态迁移成本高。
- ✓ 机遇：全球最大的AI应用场景，为国产智算芯片提供了海量落地机会。

美国：全球智算的“大脑”与“心脏”

- ✓ 技术优势：掌握3nm/2nm先进制程、HBM存储、Nvlink互联技术。
- ✓ 供应链：本土云巨头自研芯片比例超过50%，减少对外依赖，其他行业英伟达依旧首选。
- ✓ 技术封锁：通过“芯片四方联盟”强化对非美国家的出口管制，构建排他性供应链。

欧洲：“主权”与“依赖”间摇摆

- ✓ 自研尝试：法国Mistral AI、美国Graphcore、德国SAP等尝试推出专用AI芯片，但受限于缺乏先进制程和生态规模，难以撼动英伟达地位。
- ✓ 政策推动：欧盟推出《欧洲芯片法案》投入巨额补贴，试图在2030年前实现20%的芯片自给率。

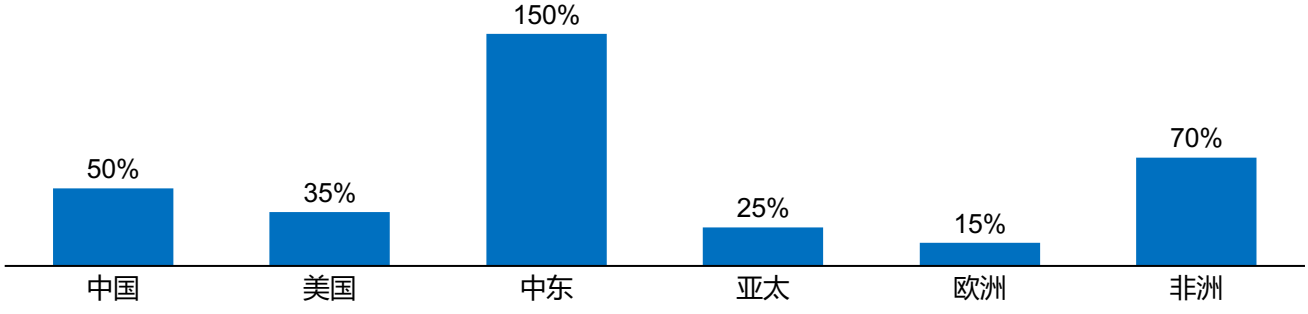
亚太地区：供应链“腰部”核心技术

- ✓ 韩国：全球HBM存储垄断者（海力士、三星），是智算芯片的关键上游。三星自研智算芯片进展缓慢，主要作为封装测试中心。
- ✓ 新加坡/马来西亚：全球先进封装中心，承接台积电、AMD、英伟达的封装订单。
- ✓ 印度：巨大的应用窗口，主要依赖进口芯片和云厂商建设数据中心。

中东与非洲：资本的“新大陆”

- ✓ 中东：大规模采购算力，建设“主权AI”。主要依赖英伟达，通过巨额订单换取技术转移和本地建厂（如沙特与英伟达合资建厂）。
- ✓ 非洲：整体算力基础薄弱，主要集中在南非、埃及、肯尼亚等少数国家。随着5G/6G普及和数字经济发展，非洲成为边缘算力和低功耗芯片的重要市场。

2026年区域算力增长对比



信息来源：科智咨询整理

2.2 大国博弈下的算力供应链重构——供应链变化

在“大国博弈”背景下，智算芯片供应链的重构已不再仅仅是商业竞争，而是上升为国家生存与发展的核心战略。在2026年视角下，供应链重构呈现出“双轨生态成型、国产加速迭代、非美供应链崛起”等特征。

全球供应链分裂为“双极”体系 01

美国对华出口管制已不可逆转，全球智算芯片供应链从“效率优先”转向“安全优先”，形成“美国主导的CUDA生态”与“中国主导的自主生态”两个相对独立，互不兼容的闭环。

- **美国主导圈：**
  - ✓ 核心：以英伟达、AMD、英特尔为核心，依托台积电、ASML光刻机、Synopsys等EDA工具。
  - ✓ 特点：拥有最先进制程（3nm/2nm）和最强生态（CUDA）。
  - ✓ 局限：供应链高度集中，地缘政治风险高，面临中国市场永久性丧失。
- **中国主导圈：**
  - ✓ 核心：以华为、海光、寒武纪等芯片厂商为核心，依托中芯国际、国内EDA、国产光刻设备（北方华创等）。
  - ✓ 特点：追求自主可控，不再追求单点制程的绝对领先，而是通过“架构创新+Chiplet+先进封装”来提升整体算力密度。
  - ✓ 局限：已建立从设计、制造到封装的“内循环”能力，但先进制程（＜7nm）仍受限于光刻机，良率与成本是主要瓶颈。

供应链“去美化”与“国产化替代”进入深水区，成本与生态是最大挑战 02

政策驱动已从“鼓励”转向“强制”，但商业逻辑（成本、生态）仍是最核心制约。

- **政策驱动：**
  - ✓ 信创2.0：党政、金融、能源、电信等关键领域要求100%国产化。
- **商业挑战：**
  - ✓ 生态迁移成本：从CUDA迁移到国产生态，迁移周期长、风险高。
  - ✓ 成本倒挂：在同等算力下，国产芯片的单卡成本可能高于英伟达，且集群运维成本更高。
- **供应链变化趋势：**百度、阿里、腾讯等巨头通过自研芯片压低国产芯片成本，并推动生态成熟。

全球供应链正在寻找“第三极”，非美供应链崛起 03

全球智算产业形成“美国核心+非美外围”的供应链网络，中国则需在“非美供应链”中寻找突破口（如与俄罗斯、中东、东南亚合作）。

- **欧洲：**推动“主权云”和“欧洲芯片法案”，试图建立独立于美国的半导体生态，但在高端AI芯片领域仍依赖美国技术。
- **东南亚（新加坡、马来西亚、越南）：**成为封装测试和晶圆制造的新中心（如台积电、Intel、AMD在马来西亚建厂），以规避地缘风险。
- **日本：**在上游材料（光刻胶、特种气体）和存储芯片（HBM）领域保持垄断地位，成为供应链重构中的关键“卡位点”。
- **韩国：**海力士、三星在HBM领域领先，是智算芯片不可或缺的上游环节。

信息来源：科智咨询整理

2.3 中国算力产业发展态势

基于2026年当前的大国博弈背景，国产智算芯片产业链已不再追求“单点突破”，而是转向“全链条自主可控”与“体系化对抗”。中国算力产业呈现出“上游卡脖子攻坚、中游生态突破、下游应用反哺”的特征。以下是针对国产智算芯片产业链的深度分析：

上游：核心环节的“卡脖子”与“突围”

半导体制造设备（光刻机）

目前上海微电子是国产光刻机的唯一主力整机商，其28nm浸没式光刻机已量产，其他厂商（如科益虹源、华卓精科等）在光学、工作台、涂胶显影等关键子系统上实现突破，共同支撑国产光刻机向14nm进军。

EDA工具

目前国内EDA工具已形成“一超多强、百花齐放”的格局，虽然Synopsys、Cadence、Siemens EDA组成的国际三巨头合计仍占据中国EDA市场超80%的份额，但国产EDA厂商正加速从“点工具”向“全流程”迈进，在各自专注的领域实现了显著突破。

晶圆代工

- 目前国内智算芯片代工格局主要由中芯国际和华虹集团主导。中芯国际是大陆唯一能量产14nm以下制程的代工厂，承担最核心、最先进的AI训练芯片制造任务。不仅是华为独家代工厂，还为海光、寒武纪等头部国产智算芯片公司提供制造服务。
- 三星、台积电受限于出口管制，大陆工厂只能生产16nm及以上成熟工艺产品，但目前互联网大厂通过海外“白手套”公司争取三星、台积电先进制程芯片订单。

中游：芯片设计与生态

智算芯片设计公司

目前部分国产智算芯片已对标英伟达部分产品。但软件和供应链稳定性是最大挑战。英伟达CUDA生态壁垒高，迁移成本巨大。目前国内智算芯片设计公司已推出各自软件栈来适配国产智算芯片。

下游：应用从“政企导入”走向“行业巨头全面拥抱”

- **互联网：**国产智算芯片已深入承载互联网核心业务。除阿里、百度等自研芯片在内部应用外，寒武纪思元系列在字节跳动已累计部署超10万张，用于支持抖音电商、AIGC多模态交互及视频内容审核等业务，海光DCU用于微信、QQ等业务的推理场景。
- **电信运营商：**三大运营商已成为国产智算芯片的“战略买家”。中国联通已建成全国首个全栈国产化智算集群，全部采用华为昇腾芯片。中国电信2025年AI服务器集采中，国产芯片占比已明确要求超过50%。
- **政企与行业：**党政、金融、能源、交通等关键领域强制要求100%国产化，是国产芯片最稳定的基本盘。

信息来源：科智咨询整理

第三章

中国智算芯片竞争态势

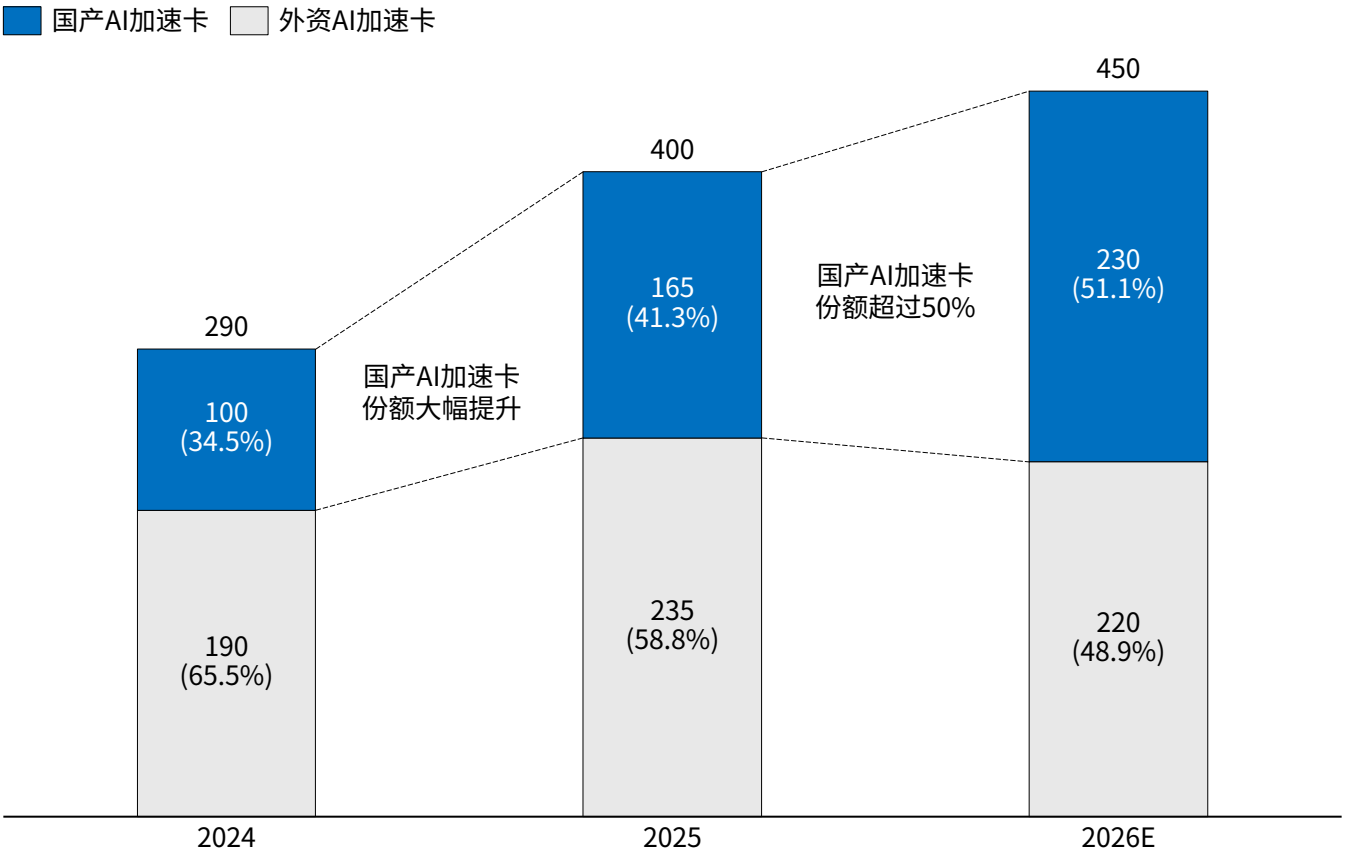
3.1 政策驱动下的国产替代加速

3.2 主流国产厂商深度剖析

3.3 供应链“卡脖子”与突围路径

3.1 政策驱动下的国产替代加速——整体市场

2024-2026E年中国市场AI加速卡出货量（万张）



- **政策驱动：**

  ✓国家战略与顶层设计：面向未来的“十五五”规划明确提出，要从根本上保障供应链安全，将核心技术掌握在自己手中。

  ✓财政与金融支持：“人工智能+”专项资金在2026年预计规模超800亿元，对AI芯片等“卡脖子”项目最高可补贴总投资的60%。

- **外资AI加速卡市场概况：**英伟达在中国AI加速器市场销售份额大幅下滑，但英伟达并未放弃中国市场，2025年12月，美国允许英伟达向中国销售H200处理器，2026年3月，英伟达宣布为中国市场生产H200。据悉，阿里、字节、腾讯、百度共下单200万+张H200卡，其中字节计划在AI硬件上投入预估约140亿美元，目前英伟达持有约70万张H200卡，面对中国互联网大厂需求，正加紧生产。在推理加速卡方面，英伟达去年以170亿美元从人工智能芯片初创公司Groq手中获得了推断加速器技术的授权，已于2026年5月正式发布相关产品，后续计划推出针对中国市场的产品版本。随着国产芯片崛起，国内市场英伟达份额出现下滑，但依旧为各互联网大厂首选供应商。

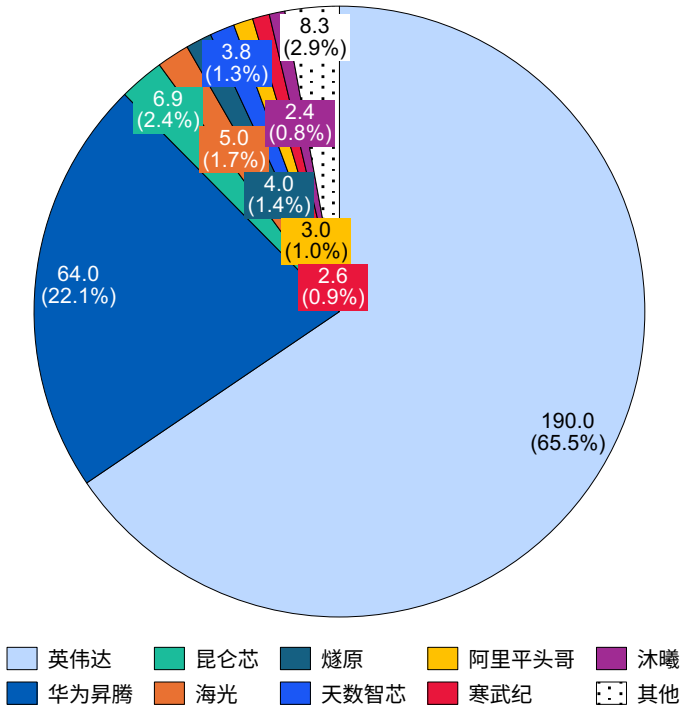
- **国产AI加速卡市场概况：**与英伟达份额急剧萎缩形成鲜明对比，国产AI芯片正在强势崛起。2025年中国AI加速卡总出货量达到约400万张，其中国产厂商出货165万张，市场份额一举跃升至41%。国产AI芯片出货持续增长，增长速度远超英伟达。目前国内至少有九家AI芯片公司出货量超过1万卡，包括华为昇腾、阿里平头哥、百度昆仑芯、寒武纪、海光、沐曦、天数智芯、燧原科技等。其中2025年华为昇腾以81.2万张出货量遥遥领先，占据国产芯片近半壁江山。

信息来源：科智咨询整理

3.1 政策驱动下的国产替代加速——竞争格局

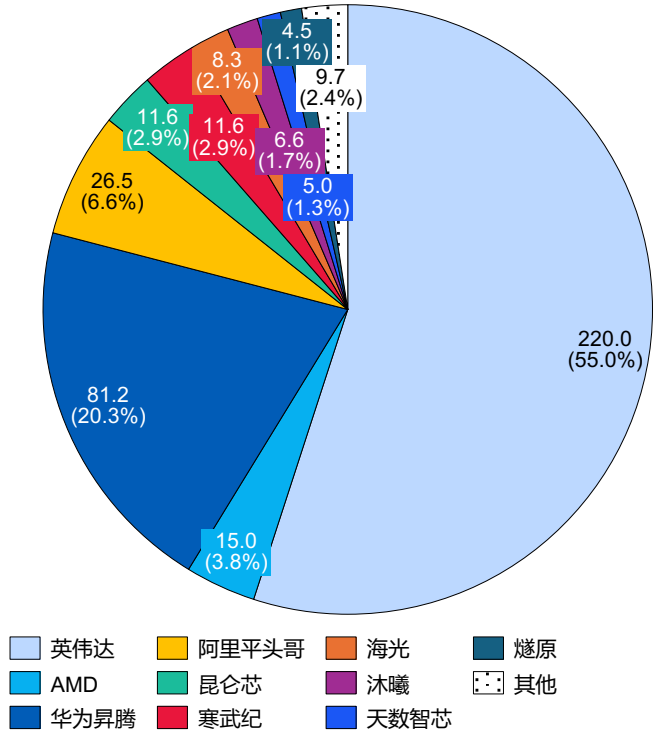
中国市场智算芯片厂商竞争格局（万片）

2024年



- 2024年，英伟达凭借H20等特供型号，以约190万片的出货量占据65%的市场份额，但份额逐步减少，国产智算芯片厂商正开始撼动市场格局。
- **华为形成“超一线”地位：**华为昇腾以64万片出货量稳居国产第一，超过其他国产头部厂商出货量总和。在国产芯片阵营中形成“超一线”的领先地位。同时也被美国列入全面禁运名单，侧面印证已成为中国算力基石的关键角色。
- **第二梯队初具规模：**昆仑芯、海光、燧原、寒武纪等共同构成国产智算芯片的“一线阵营”。不过除华为昇腾外其他国产品牌出货量均未过10万片，市场分层明显。

2025年

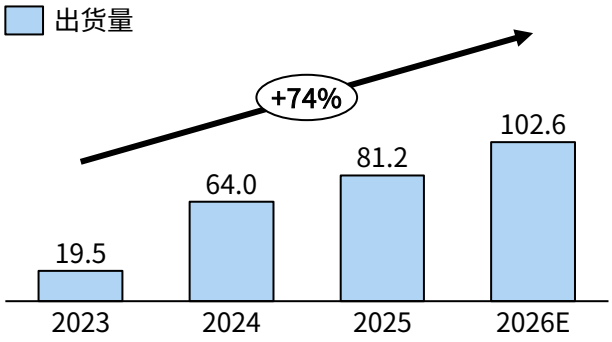


- **出口管制的倒逼效应：**2025年，英伟达仍以约220万张出货量首位，但与制裁前近乎垄断的95%份额相比，已大幅下滑。美国连续多轮管制切断了中国获取英伟达最先进芯片的渠道，迫使市场加速转向国产替代。
- **华为领跑国产阵营：**华为以81.2万张出货量位列国产第一、整体第二，占国产品牌总出货量近一半。
- **政策端的强力推动：**2025年以来，各地智算中心建设加速，项目采购环节普遍向国产芯片倾斜。
- **互联网厂商出货大幅增长：**阿里、百度智算芯片出货量在国产品牌中达到第二、第三位置。字节预计2026年第三季度推出首轮规模化产品。

信息来源：科智咨询整理

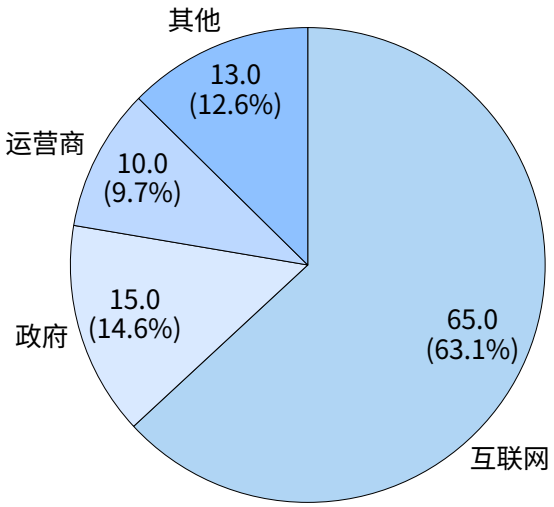
3.2 主流国产厂商深度剖析——华为昇腾

华为昇腾智算芯片出货量（万片）



华为昇腾已完成从910系列到950系列的转换。预计2026年昇腾950PR市场出货量达到约80万片，对于950DT，到2026年底出货量预计约10-20万片，950系列相对上一代产品性能大幅提升，已有互联网大厂预定下单950PR。

华为昇腾智算芯片客户结构（2026年预测、万片）



预计2026年，华为950系列在阿里、腾讯、字节等需求最多，综合来看互联网厂商总需求预计达到65万片，剩余产品将用于满足政府、运营商、企业等客户需求。目前市场货源紧张，一些渠道至今仍未获得配货，据部分渠道反馈，需等待互联网公司的订单情况最终确定后才能分配。

华为的超算节点客户主要包括一些实验室、部分国产车厂的自动驾驶业务单元以及大型运营商。互联网公司通常不会直接采购完整的超算节点，而是采用自建的方式。采购华为的核心部件，但机箱则根据自身尺寸需求定制，外围的网络布线等则自行决定。

华为昇腾智算芯片产品演进

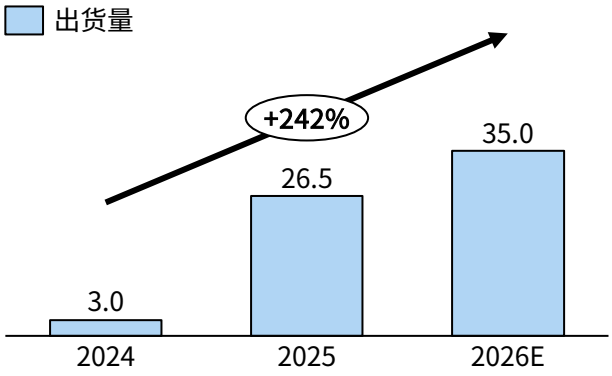
	910系列	950系列	960系列（规划中）
核心参数	**910C**	**950PR**	**960**
典型算力	FP16:0.8PFLPOS	FP8:1PFLOPS FP4:2PFLOPS	FP8:2PFLOPS FP4:4PFLOPS
HBM内存	128GB	128GB	288GB
内存带宽	3.2TB/s	4TB/s	9.6TB/s
互联带宽	784GB/s	2TB/s	2.2TB/s
功耗	280W	600W	-

- **950系列能够实现互联带宽升级至2T**，核心依托先进高端封装技术完成突破，从硬件封装与内部传输架构两大维度协同发力，最终完成2T互联网带宽落地应用。
- 950系列完成从类ASIC专用架构向类GPGPU通用计算架构转型，最核心的作用就是全面提升芯片产品通用性与全场景应用适配能力。**现阶段950系列针对市场主流的CUDA整体兼容覆盖度已突破95%。**

信息来源：科智咨询整理

3.2 主流国产厂商深度剖析——阿里平头哥

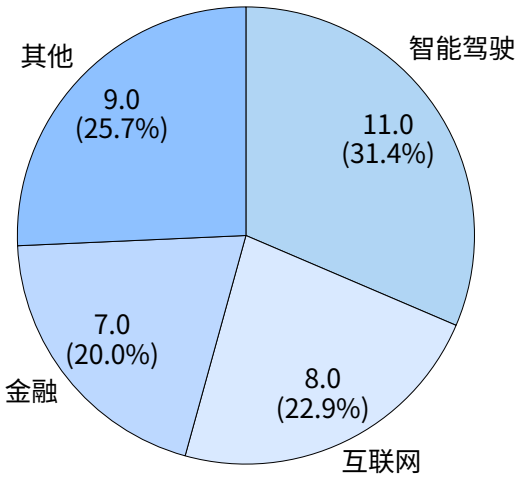
阿里平头哥智算芯片出货量（万片）



2024年，真武810E已开始逐步量产，主要在阿里内部使用。

2025年，真武M890正式发布，一方面美国政府限制H20等特供芯片供应，另一方面，部署在阿里云数据中心同时，针对外部客户拿下400+规模商业订单，实现大幅增长。

阿里平头哥智算芯片客户结构（2026年预测、万片）



在智能驾驶领域，与长安、比亚迪、小鹏、蔚来、理想等超30家头部客户已落地部署。在与车企合作中，通过阿里云完成内部结算，车企相关业务款项先结算至阿里云，再由阿里云与平头哥进行内部账务划转。

在互联网行业中，目前主要部署在阿里云数据中心，用于训练通义千问大模型。

在金融行业中，服务工农中建等超过150家银行、证券及保险机构，覆盖认证比对、风控、智能客服等核心场景。

其余行业覆盖能源、制造等行业，主要面向大型央国企及创新企业布局。

阿里平头哥智算芯片产品演进

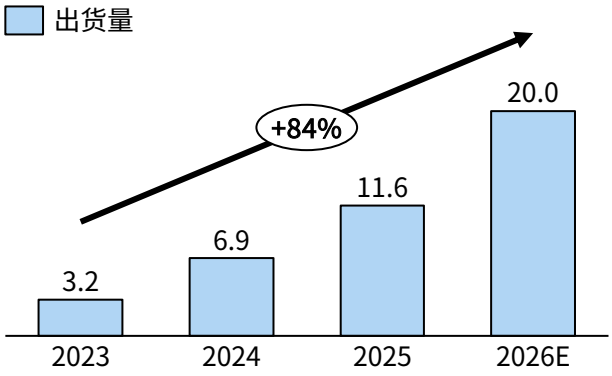
核心参数	真武810E	真武M890	真武V900（规划中）
显存容量	96GB	144GB	216GB
显存带宽	2.7TB/s	-	-
互联带宽	700GB/s	800GB/s	1200GB/s
功耗	400W	-	-

阿里平头哥的真武系列，正以“一年一代”的节奏进行迭代。从驱动多场景的810E，到转为智能体时代设计的M890，再到计划于2027年Q3发布的性能巨兽V900，每一代产品都在突破架构性能上限。

信息来源：科智咨询整理

3.2 主流国产厂商深度剖析——昆仑芯

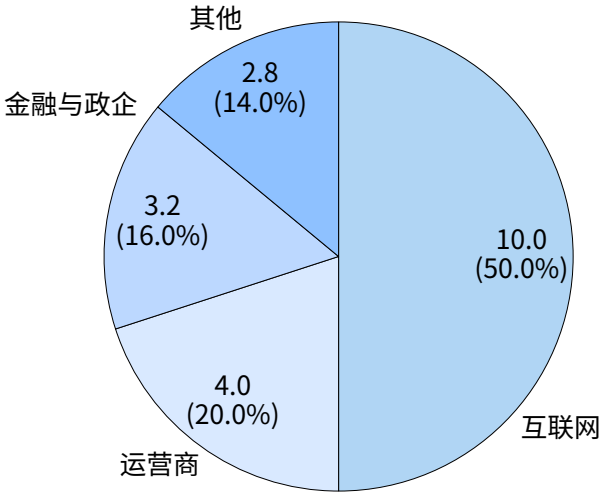
昆仑芯智算芯片出货量（万片）



在客户拓展方面，昆仑芯已成功从内外部两条路径双轮驱动，不仅获得了腾讯和中国移动等巨头的规模订单。在生态构建上，其自研XPU架构支持万卡级集群。

百度在天池256卡超节点上的技术突破，也为其带来了新的增长曲线。

昆仑芯智算芯片客户结构（2026年预测、万片）



在互联网行业中，除百度内部自用外，2026年已有2万片的订单落地，是外部市场的重要突破。并且与字节正在进行灰度测试，预计未来可能达成合作。

运营商庞大的智算中心建设为昆仑芯提供了巨大机遇。以中国移动为例，其2025年AI服务器集采中，斩获数十亿元大单。

金融和政企是国产替代需求最迫切的领域。昆仑芯已成功中标招商银行AI芯片项目，并且其产品已广泛应用于智慧金融、央企、智慧城市等场景。

昆仑芯智算芯片产品演进

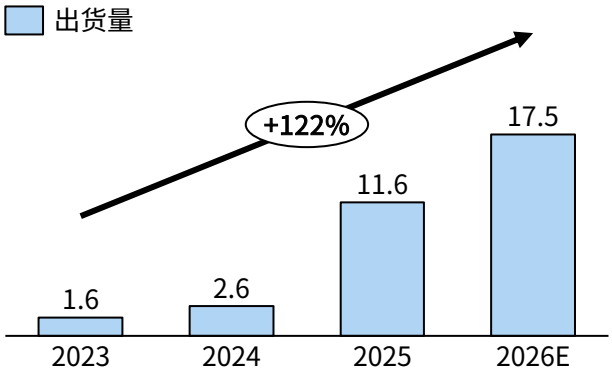
核心参数	K系列	R系列	P系列	M系列（规划中）
制程工艺	14nm	7nm	7nm	
FP16	64 TFLOPS	96-128 TFLOPS	345 TFLOPS	
显存容量	8-16GB	16-32GB	96GB	
显存带宽	512GB/s	512GB/s	3TB/s	
功耗	75-150W	120-150W	400W	

- 2024年，昆仑芯第三代P800系列产品量产，P800是当前主力在售产品，但交付节奏受限于三星产能。昆仑芯当前在国内的流片规模较小，核心流片环节依旧在海外开展，短期内与三星的合作关系保持稳定，不会出现重大合作变动。
- 预计2026年发布M100，2027年发布M300。

信息来源：科智咨询整理

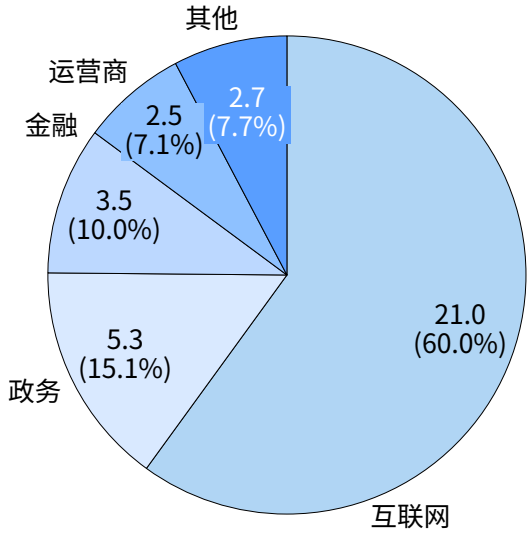
3.2 主流国产厂商深度剖析——寒武纪

寒武纪智算芯片出货量（万片）



2025年寒武纪在政务、金融等领域实现大幅增长。目前寒武纪MLU590、MLU690等核心产品已实现流片。但在产品良率仍存在明显短板，整体产能处于紧缺状态，无法满足大规模出货要求，直接影响寒武纪2026年规模化出货目标的实现，这也是其当前面临的核心供应链难题。

寒武纪智算芯片客户结构（2026年预测、万片）



预计2026年寒武纪在互联网行业迎来大幅增长，其中：在字节跳动主要交付型号为MLU580、MLU590、MLU690，预计交付量在10万片以上。在阿里，寒武纪MLU590签订了框架协议，目前已有1万片进行灰度测试。在腾讯，寒武纪MLU590签订了近10万片的框架协议，预计2026年实际交付量约几万片。

政务与金融为寒武纪第二大重点布局领域，主要受惠于政策驱动的国产化替代。针对高安全需求场景，寒武纪强化了芯片的数据加密模块和全栈国产化适配。推出思元590的行业定制版，以满足特定需求。

寒武纪智算芯片产品演进

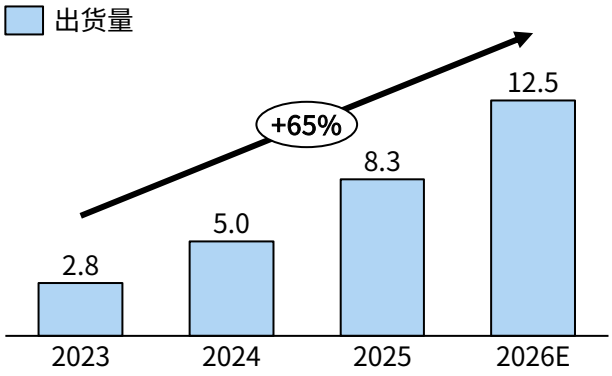
核心参数	思元590	思元690	思元790（规划中）
FP16	256 TFLOPS	700 TFLOPS	2 PFLOPS
显存容量	96GB	196GB	-
互联带宽	372GB/s	512GB/s	-

- 2025年寒武纪交付的主力产品为590芯片，2026年将向客户交付采用全合规国产供应链的590及690芯片，年度主力产品为基于690芯片打造的超节点方案。690超节点则直接对标华为910C 384 超节点，争夺国内高端算力市场份额。2026年下半年至2027年初，寒武纪预计推出下一代产品790，性能相较690实现翻倍提升，同时大幅优化网络性能以及HBM显存与带宽配置。
- 寒武纪690的产能约束远大于昇腾950系列，产能瓶颈集中在三大核心环节：一是芯片流片环节受限，二是采用CoWoS封装路线带来的封装产能瓶颈，三是HBM高宽带内存供给不足。

信息来源：科智咨询整理

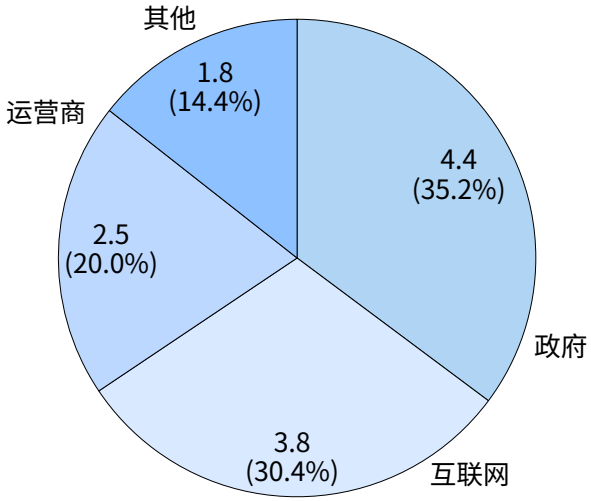
3.2 主流国产厂商深度剖析——海光

海光智算芯片出货量（万片）



在字节、阿里、腾讯等客户中，海光DCU（深度计算单元）正处于渗透率快速提升的关键阶段。2026年第一季度，海光营收同比增长68%，达到40.34亿元，在海光DCU带动下，业绩实现飞速增长。

海光智算芯片客户结构（2026年预测、万片）



海光在2025年底与腾讯、阿里、字节逐步完成合作。腾讯大规模采购深算三号主要是因为当前H20芯片供应受阻，阿里和字节对深算三号的采购量较为保守。海光虽具备较强的架构与研发实力，但项目经验主要集中在国央企与政府领域，缺少大型互联网客户的服务经验，在规模化落地过程中面临的阻碍更多，整体完成出货目标的不确定性更高。

政府和运营商行业当前为海光最大的营收来源，尤其在政府主导的智算中心项目中，中标率高。

海光智算芯片产品演进

核心参数	深算二号	深算三号	深算四号（规划中）
FP16	46/100 TFLOPS	500 TFLOPS	960 TFLOPS
FP64	21 TFLOPS	30 TFLOPS	80 TFLOPS
显存容量	64GB	64GB	96GB

- 海光在2025年开始销售兼具超算与AI算力能力的深算3号（KW1000）芯片，2026年下半年预计提前推出深算四号芯片，性能相较深算三号实现翻倍增长。2026年海光出货主力为深算三号与深算四号，产品将搭载海光自研CPU，结合沉没式液冷技术，推出640张液冷超节点及全新架构超节点方案。
- 2027年深算四号将成为市场主流产品，深算五号预计在2027年下半年或年底正式推出。海光的核心竞争优势在于CPU和AI芯片的协同算力能力，产品布局重点聚焦超算与AI融合应用场景，主要服务对架构兼容性要求高的行业客户。

信息来源：科智咨询整理

3.3 供应链“卡脖子”与突围路径

四大核心“卡脖子”节点

尽管国产化进程加速，但在高端智算领域，供应链的完整性仍面临严峻考验：

1 制造之困：先进制程与光刻设备

制造层面的短板尤为突出。中芯国际等代工厂的先进工艺产能（如7nm、5nm）非常有限，且晶圆良率仍需提升。更关键的是，制造环节高度依赖从海外进口的先进设备（尤其是ASML和EUV光刻机），这构成了制造环节“卡脖子”的物理瓶颈。

2 设计之困：EDA工具依赖

芯片设计、仿真和验证高度依赖EDA工具、其长期被Synopsys、Cadence、西门子等海外巨头垄断。相比之下，国产EDA工具链在先进工艺等系统级设计验证环节仍存在关键短板，直接制约了国产芯片向更高性能迭代的进程。

3 性能之困：HBM严重短缺

随着AI大模型的发展，高带宽内存（HBM）已成为制约智算芯片性能的关键瓶颈。目前，全球HBM几乎被三星、海力士和美光三家巨头垄断，国产化几乎为零，国内产业链在此领域的积累与国际领先水平仍有差距。

4 生态之困：软件栈生态（CUDA壁垒）

英伟达CUDA生态经过十余年积累，锁定了全球数百万开发者。国产芯片不仅要“算的快”，更要解决大模型“移不动、配不上”的生态隔阂。

破局突围的四大路径

技术路线创新：通过计算架构提升算力

国产芯片已不再单纯追求缩小单个晶体管尺寸，而是通过重构计算架构（如存算一体、异构计算）提高整体算力。如壁仞科技的BR100芯片通过Chiplet技术，将大芯片拆分为小芯粒制造再封装互联，在同等工艺下实现了更高的集成度和良率。

系统级突围：以“超节点”发挥综合优势

当单科芯片性能受限时，通过系统架构的创新实现“堆量换质”。华为的CloudMatrix 384和昆仑芯的超节点技术通过将数百乃至数千颗芯片高效互联，构建了强大的计算集群，在整体算力和能效上实现了对单卡性能的超越。

生态突围：软硬协同构建护城河

国产厂商正在通过多种方式构建自己的生态护城河。沐曦的MXMACA软件栈、海光基于“类CUDA”的DTK工具链等兼容层、寒武纪NeuWare、华为CANN等自研生态，都在努力降低用户的迁移和使用门槛。

供应链协同：在多环节寻求突破

制造端方面，国产28nm DUV光刻机已进入产线试用，标志着设备自主化迈出第一步。设计端方面，以华大九天为代表的国产DEA企业通过战略并购（如入股思尔芯）补齐数字EDA流程。同时，产业界通过组建“芯片-模型”联盟等方式，共同攻克从芯片设计到应用落地的难题。

信息来源：科智咨询整理

第四章

关键挑战与风险研判

4.1 供应链风险
4.2 市场竞争风险
4.3 地缘政治与政策风险

4.1 供应链风险——上游设备与材料断供风险

高端制造设备：目前国产化率严重不足

根据中国电子专用设备工业协会的数据，2025年中国大陆半导体设备市场规模约495亿美元，但国产设备在关键环节的渗透率严重不足。

设备类别	国产化率（2025年）	主要依赖
光刻机（浸没式DUV/EUV）	＜5%	ASML
薄膜沉积（PVD/CVD/ALD）	约15%	AMAT、Lam、TEL
检测与测量设备	约10%	KLA、Onto、Nova
刻蚀设备	约25%	Lam、TEL、中微
清洗设备	约35%	Screen、盛美
涂胶显影设备	＜10%	TEL
离子注入机	约10%	Applied Materials、Axcelis

光刻机是“卡脖子”最严重的环节。尽管上海微电子28nm DUV已量产交付。但其产能远不能满足全行业需求，且更先进的浸没式光刻机尚未进入大规模商用。2025年10月，美国进一步收紧对华DUV光刻机出口许可审批，导致国内晶圆厂已订购的多台设备延迟交付。

关键材料：多维度收紧，高端材料已断供

01 光刻胶

- **高端实质性断供：**自2025年末起，日企已停止接受ArF及EUV光刻胶的新增订单，审批通过率趋近于零。2026年，中国海关自日本进口的ArF/EUV光刻胶报关量为零。
- **外资厂商占主导地位：**光刻胶是半导体制造的核心材料，直接决定了芯片制造的精度与良率。日本企业信越化学、JSR、东京应化在全球高端市场占据绝对主导地位。

02 大尺寸硅片

- 12英寸硅片国产化率约20%，主要依赖日本信越、SUMCO和中国台湾环球晶圆。沪硅产业、中环股份虽已实现部分供货，但正片（用于先进制程）良率和稳定性仍有差距。

03 特种气体

- 高纯电子特气（如氖气、六氟化钨）部分品类仍依赖乌克兰、美国供应。2025年地缘冲突导致氖气价格波动超300%，暴露出供应链的地缘脆弱性。

04 CMP抛光液

- Cabot、Dow、Fujimi占据全球70%以上份额，国产安集科技、鼎龙股份在成熟制程已有突破，但14nm以下仍依赖进口。

信息来源：科智咨询整理

4.1 供应链风险——晶圆代工风险

海外代工通道全面封堵

2025年1月生效的BIS出口管制新规，对海外代工厂向中国大陆客户提供先进制程代工服务设置了严格限制。具体执行情况如下：

代工厂	中国大陆客户先进制程接单状态（截至2026年6月）
台积电	全面停止新订单，现有订单需在2025年Q3前交付完毕
三星	全面停止接受中国大陆AI芯片公司新设计定案，三星在华工厂仅限制成熟制程
联电	不接受16nm以下新单，16nm及以上需申请许可，许可审批通过率极低
格罗方德	不接受任何中国大陆AI芯片代工订单

- 国产智算芯片公司已无法通过台积电等海外代工厂获得先进制程产能。历史上依赖台积电流片的海光、寒武纪等公司，被迫紧急将订单转移至中芯国际。
- **海外子公司“绕道”路径被封堵**：2026年5月31日，BIS发布新规：所有总部位于中国（包括香港）的企业，无论其子公司位于何地，出口先进芯片均需申请许可证。彻底堵住了以往中国企业通过马来西亚、新加坡等地子公司向台积电下单的路线。

单一代工依赖的风险

（一）中芯国际：唯一选择下的超额承压

当前，能够为国产AI训练芯片（7nm级）提供代工服务的，大陆仅有中芯国际一家。这种高度集中的格局带来风险：

- **产能排挤风险：**华为凭借战略合作关系和技术积累，在产能紧张时优先获得排产。2025年下半年，部分中小AI芯片公司的流片排期被推迟了4-6个月，直接导致产品上市延迟和订单流失。
- **议价权失衡：**代工价格由卖方主导。2025年中芯国际已对先进制程进行两次调价，累计涨幅约15-20%，直接拉高了国产AI芯片的成本。

（二）华虹作为第二选择供应不足

华虹集团旗下的华力微电子正在推进7nm工艺研发，原计划2025年底量产，但目前进度已滞后。7nm产线仍处于小批量试产阶段，尚未通过主流AI芯片公司的可靠性认证。即便2026年顺利量产，产能也仅约中芯国际的1/5-1/4，短期内难以形成有效备份。

（三）代工产能对需求的供应不足

2026年国产先进AI芯片总需求约450万颗，而中芯国际+华虹合计供给约270-280万颗，缺口接近200万颗，这一缺口需要由海外代工填补的，但目前已不可行。或应用先进NV产能，或只能被迫采用成熟制程生产（性能受损），或延迟交付（失去市场窗口）。

4.2 市场竞争风险

市场竞争风险	核心风险子维度	具体风险内容
国产厂商之间的同质化竞争风险	1.数量繁荣下的质量堪忧	• **技术路线趋同：**绝大多数国产智算芯片在硬件参数上，不同厂商的产品往往高度相似，技术参数与NV存在一定差距. • **目标市场重叠：**超过70%的国产智算芯片厂商将“互联网大厂推理场景”作为首要攻坚目标，导致大家在同一片红海中竞争。 • **差异化不足：**除华为昇腾在集群方案、海光在“类CUDA”兼容性上建立了相对清晰的差异化外，多数中小厂商停留在对标NV某款产品，价格优势成为唯一卖点。
	2.价格战初见端倪，利润空间承压	• 2025年下半年，部分国产AI加速卡市场报价逐步下滑，多个中小厂商跟进降价，试图以低价换取市场份额。2025年国产智算芯片公司的平均毛利率约为45%，较2023年下降约10个百分点。若价格战持续，部分中小厂商将在1-2年内陷入亏损。
	3.头部集中加速，长尾厂商生存空间收窄	• 国产厂商竞争出现“马太效应”加剧。2025年TOP5厂商芯片出货量占比达到约85%。未来2-3年内预计国产智算芯片市场将经历一轮“洗牌期”，大量缺乏核心技术和资金储备的创业公司将出局。
英伟达的反扑风险	1.政策窗口期的精准打击	• 美国已对H200有条件放行，未来若政策进一步松动，英伟达可迅速向中国投放大量产能。未来可能会针对中国市场的特殊需求开发“定制版”芯片，直接瞄准国产厂商当前的优势赛道。
	2.生态壁垒的“锁定效应”仍在强化	• 英伟达最大的护城河并非硬件，而是CUDA生态。截至2026年初，CUDA全球开发者数量已超过400万，支持超过3000款AI应用框架。国产芯片即使在政策保护下也难以真正撼动其根基。
互联网大厂自研芯片的“外销”冲击	1.从“自用”到“外销”	• **阿里平头哥：**真武系列芯片已通过阿里云对外提供算力服务。并开始向长安、比亚迪等外部客户直接销售芯片。 • **百度昆仑芯：**已在中国移动、招商银行等客户中实现批量部署，并在运营商集采中与华为、寒武纪等直接竞争。 • **字节跳动：**虽然自研Seedchip目前主要用于内部推理，但一旦产能规模上升，字节同样会考虑将其芯片产品化并推向市场。

市场竞争风险的综合研判

市场竞争风险	当前严重程度	演进趋势（2026-2028）	关键观察指标
国产厂商之间的同质化竞争风险	中危	加剧分化，头部集中，长尾淘汰	中小厂商融资情况、毛利率变化
英伟达的反扑风险	高危	取决于美政策松绑程度	H200对华实际出货量、CUDA中国开发者增长
互联网大厂自研芯片的“外销”冲击	中危	逐步升温，2-3年后形成实质冲击	平头哥/昆仑芯外部客户收入占比

信息来源：科智咨询整理

4.3 地缘政治与政策风险

地缘政治与政策风险	核心风险子维度	具体风险内容
美国制裁逐步升级	1.实体清单的持续扩容	• 从2023年154家、2024年264家、2025年美方新增约133个中国实体进入实体清单。2026年这一趋势仍在延续，浪潮信息在内地及港澳台的6家子公司被集体列入，显示出美方对高性能计算产业链的全面封堵。
	2.从“芯片禁运”到“设备、软件、服务、人才”的全方位封锁	• 美国的制裁已不再局限于禁止某种型号的智算芯片出口，而是构建起一张覆盖芯片、设备、EDA软件、技术合作、人才流动的全方位网络。
	3.“外围直接产品规则”的杀伤半径持续扩大	• “外国直接产品规则”（FDPR）是美国出口管制中最具威慑力的条件。任何外国生产的、只要使用了美国技术（含EDA、设备等）的产品，都受到美国出口管制法的管辖。 • 2026年1月，只要任何被列入实体清单的中国股东直接或间接持有某非美国公司50%及以上股东，该公司受到同等限制，直接击碎试图通过海外子公司获取技术的幻想。
	4.H200“有条件放行”的陷阱与博弈	• 2026年5月，美国正式批准英伟达H200出口中国，标志着出口管制政策从“一刀切禁止”调整为“附条件放行”。但美国政府要求抽取高达25%的销售收入分成，实则等同于加征高额关税。
盟友协同管制的“合围”效应	1.日本：坚决跟随的“政策急先锋”	• 日本是美国盟友中对华半导体出口管制最为坚决的国家。自2023年起，日本将23种关键半导体设备纳入出口管制，2025年进一步将21个物项纳入出口管理，范围涵盖光刻、蚀刻、薄膜沉积等。
	2.荷兰：光刻机的“水龙头”正在拧紧	• 在美方持续施压下，荷兰已将限制范围从EUV光刻机扩大至中高端DUV光刻机，管制节点从7nm收紧到14nm级别以下。
国内政策的“双刃剑”效应	1.信创市场的“围墙”效应	• 过度依赖政府采购和信创市场，可能在客观上限制国产芯片在更广阔的国际市场竞争中自身产品能力，形成“温室效应”。

地缘政治与政治风险的综合研判

市场竞争风险	当前严重程度	演进趋势（2026-2028）	关键观察指标
美国制裁逐步升级	高危	持续强化，将更多第三国企业纳入管辖	制裁是否持续升级，新增中国实体清单企业数量
盟友协同管制的“合围”效应	高危	美日荷协同加强	MATCH法案立法进程，ASML在华服务政策
国内政策的“双刃剑”效应	中危	需持续优化	地方智算中心利用率、补贴效率

信息来源：科智咨询整理

第五章

未来发展趋势研判

5.1 技术趋势

5.2 竞争格局演变

5.1 技术趋势

智能算力结构：从“训练为主”到“训推一体、推理主导”

- ✓ **AI计算中心开始转移：**AI计算分为训练和推理两个层面，训练是一次性投入，对单卡峰值算力和集群规模敏感，推理则是持续性、碎片化的运营支出，对单位Token成本更为敏感。随着AI大模型从实验室走向大规模商业化部署，算力需求的结构正在改变。预计2026年推理需求将达到训练需求的4-5倍。
- ✓ **推理算力爆发驱动芯片架构重新定义：**智能体在AI应用中的加速渗透，从根本上改变算力消耗模式。与传统AI一问一答的调用模式不同，智能体在执行任务时往往需要多轮推理，单个任务消耗的Token数量是传统对话的数十倍。推理场景的差异化驱动芯片架构多元化：硬件厂商必须在吞吐、时延、成本之间做出取舍，单一规格的通用推理芯片难以同时覆盖全部负载。
- ✓ **芯片设计从“峰值算力”转向“能效比与单位成本算力：**推理算力所需算力通常比GPU提供的要少，但需要更多内存，如果内存不足，芯片无法快速获取数据。在此背景下，智算芯片厂商正在从追求”峰值算力“的单一指标，转向同时优化能效比、单位成本算力、推理延迟的吞吐量等多维指标，以实际业务场景的性价比作为核心竞争力。

超节点与集群：系统级创新的持续深化

- ✓ **从“单芯性能”到“系统效率”的范式迁移：**传统算力集群因通信延迟导致大量计算资源处于空闲状态。超节点通过全互联架构将算力利用率提升50%以上。
- ✓ **超节点规模部署引领产业新需求：**当超节点算力密度大幅提升，AIDC功率迎来新挑战。由于包括华为CloudMatrix 384、英伟达GB200 NVL72在内的超节点单机柜功耗普遍突破100KW，在算力密度指数级增加的情况下，超节点机柜的温控和电源系统面临严峻挑战。当Atlas 950超节点采用全液冷模式时，其互联带宽速率和算力速率有望迎来大幅提升。2026年有望成为液冷规模部署元年。

制程工艺：先进制程加速发展

- ✓ **先进制程——供需矛盾的现实与突破：**先进制程产能与国内需求之间存在显著缺口。2026年昇腾910系列出货目标从80万片大幅下调至60万片。凸显了国产先进制程产能的现实约束。中芯国际正积极扩产，目标是到2027-2028年将7nm及类5nm级工艺的月产量从目前不足2万片提升至10万片。
- ✓ **成熟制程——规模化优势与协同配套：**除先进制程外，成熟制程（28nm及以上）路线仍在持续发挥规模化优势。中芯国际和华虹半导体的28nm等成熟制程已具备全球竞争力。对于智算芯片设计而言，成熟制程与先进制程并非替代关系，而是以“Chiplet混搭”策略互为补充，将计算核心与I/O、存储控制等周边模块解耦，实现性能、良率和成本的三重均衡。

5.2 竞争格局演变

2026年中国智算芯片竞争格局演变（预测，万片）

排名	厂商	2026年出货量（预测）	2025年出货量	2026年市场份额（预测）	2025年市场份额
1	英伟达	200.0	220.0	44.4%	55.0%
2	华为昇腾	102.6	81.2	22.8%	20.3%
3	阿里平头哥	35.0	26.5	7.8%	6.6%
4	昆仑芯	20.0	11.6	4.4%	2.9%
5	AMD	20.0	15.0	4.4%	3.8%
6	寒武纪	17.5	11.6	3.9%	2.9%
7	海光	12.5	8.3	2.8%	2.1%
8	沐曦	8.0	6.6	1.8%	1.7%
9	天数智芯	6.0	5.0	1.3%	1.3%
10	燧原	5.5	4.5	1.2%	1.1%
-	其他	22.9	9.7	5.1%	2.4%
-	总计	450.0	400.0	100.0%	100.0%

- **国产芯片市场“一超多强”格局持续固化**

华为凭借全栈自研能力，生态成熟度，在国产智算芯片市场中处于“超一线”地位，市场份额进一步提升，短期难以撼动。国内智算芯片厂商长期趋势为“华为一家独大，其他各家各守赛道”。

- **互联网自研芯片对外销售趋势加速**

阿里平头哥、百度昆仑芯已明确将“向外部客户销售”为未来核心层战略，字节自研SeedChip未来也可能对外输出。这意味着国产智算芯片厂商不仅要与传统海外巨头竞争，还要与自己的客户竞争。竞争格局正从“华为+第二梯队厂商”演变为“华为+互联网大厂自研芯片+独立芯片厂商”的多层嵌套模式。

- **英伟达重返中国市场的战略选择**

英伟达的高端训练芯片（B200/Rubin系列）在中国全面禁售是长期确定性事件，H200的回归取决于政策松紧节奏，而在推理芯片市场，英伟达通过定制化产品和生态优势持续施加竞争压力。预计2026年，英伟达出货量及市场份额持续下滑。

信息来源：科智咨询整理

第六章

中国算力厂商案例介绍

6.1 案例1：某省级电信智算中心改造项目

6.2 案例2：无锡算力调度运营平台

6.1 案例1：某省级电信智算中心改造项目

● 项目背景

该省级电信已建成专业智算中心，面向市场开展裸金属算力租赁、大模型训推等业务。项目落地后暴露出两大核心问题：一是前期 GPU 选型不符市场主流需求，部署的英伟达 A100 40GB 显存机型难以满足绝大多数裸金属租赁客户的使用标准，设备长期闲置，空置时长接近一年，硬件资源严重浪费；二是原有管理平台能力不足，无法完成算力 Token 转化输出，分布式推理集群的算力无法正常对外提供服务，集群业务价值无法释放。客户亟需通过技术改造优化集群架构、盘活闲置算力、打通推理算力输出能力，实现智算中心商业化运营。

● 项目介绍

本次改造从网络架构、算力管理、存储支撑三大维度整体升级，搭建一体化高性能智算集群架构。首先开展网络改造，新增 IB 网络并完成集群网络标准化规整，全面提升集群内网传输性能。其次上线天邃精益智算管理平台，对现有 GPU 集群进行统一管控与算力转化，启用分布式推理能力，结合 PD 分离、GDS、LMcache 等核心技术，并配套部署天邃 AI 高性能存储系统，协同优化集群 Token 输出效率。同时依托平台原生分布式训练能力，对接 deepspeed、magtron 等主流训练框架，配置 checkpoint 断点续训、热备节点等功能，构建完整的大模型训练、推理一体化集群，全方位完成智算中心能力升级。

● 解决方案核心亮点

- 全栈技术融合，适配场景全面：整合网络优化、算力调度、加速技术、分布式框架与高性能存储，同时覆盖大模型训练、分布式推理两大核心场景，一站式解决集群功能缺失、性能不足问题。
- 成熟技术组合，性能优化突出：融合 PD 分离、GDS、LMcache、断点续训、节点热备等多项前沿技术，大幅提升 Token 输出效率与集群运行稳定性，强化集群承载能力。
- 生态兼容性强：原生兼容 deepspeed、magtron 等主流 AI 框架，对接主流智算生态，无需大规模重构现有业务体系，改造成本低、落地速度快。
- 资源盘活能力强：依托精细化算力管理能力，充分挖掘存量 GPU 硬件潜力，将长期闲置设备转化为可商用的合规算力资源。

● 用户价值

- 改造完成后，原本空置近一年的 GPU 算力资源全部实现对外租赁，上线 1 个月内算力资源 100% 出租，彻底解决硬件闲置问题；
- 分布式推理集群成功实现 Token 正常输出，智算中心完整具备裸金属租赁、大模型训练、在线推理全业务能力，集群商用价值充分释放，大幅提升项目营收与资产利用率。



6.1 厂商简介及产品服务——沨呵智慧

● 厂商简介

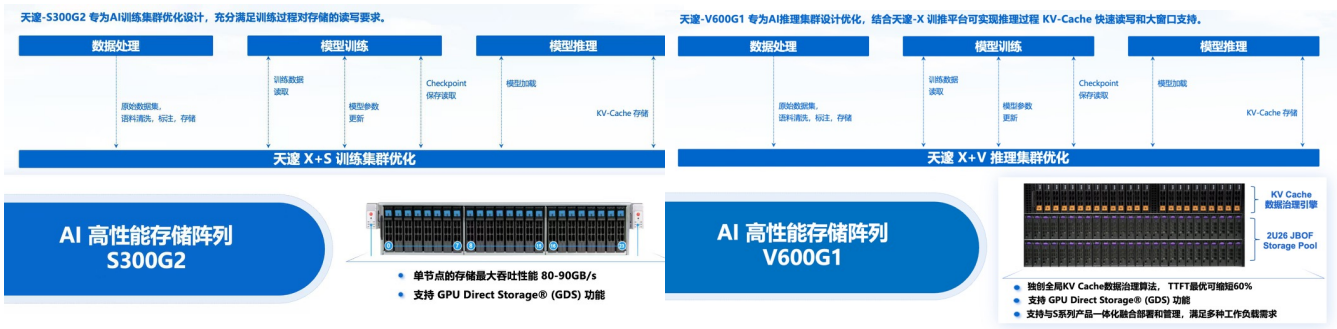
沨呵智慧科技（上海）有限公司专注打造全链条智算综合解决方案。业务覆盖智算中心规划设计咨询、算力调优、AI大模型训练推理等全流程，同时深耕高性能存储、边缘计算、异构算力调度等前沿领域。公司核心团队拥有国际化北京，核心成员均来自HP、IBM、Oracle等全球科技巨头，平均具备 20 年以上大型数据中心设计、建设与落地实施经验，擅长攻克智算集群运行稳定性、算力利用率等行业痛点，持续优化客户整体 TCO 成本，并依托 AGI 技术为各类客户赋能。

● 产品服务

- **天邃精益智算管理平台**：该平台是面向智算集群打造的综合性管理工具，聚焦算力低效、运维繁琐、大模型应用门槛高等行业难题，集成算力调度、全链路智能运维、大模型全生命周期管理三大核心能力，实现智算运营精益化、自动化、低成本化。



- **天邃 AI 高性能存储系统**：天邃 S/V 系列存储是专为大模型训练，推理场景量身打造的专用分布式高性能存储产品，作为AI算力体系的核心数据引擎，凭借优异的硬件架构、极致的读写性能、高可靠性与强兼容性，全面满足智算场景海量数据存取、高速流转、安全管控的核心需求，其中 S300G2 专为训练 Checkpoint优化，实现训练过程对ckpt文件的高速读写需求；V600G1 专为推理 KV-Cache 优化，结合天邃 X 训推管理平台实现推理Tokens的最优输出实现。



6.2 案例2：无锡算力调度运营平台

● 项目背景

随着人工智能、大模型和行业智能化应用加速落地，无锡区域算力呈现两侧矛盾。

从供给侧看，区域内智算中心、边缘节点、企业自持算力和社会化算力资源分散在不同主体、不同地点和不同技术架构之下，芯片类型、软件栈、调度方式和计量标准不一致，存在资源状态不透明、供需匹配低效、调用方式不统一、服务交付链路长等问题，容易形成"有资源但难调用、有算力但难交易"的局面。

从需求侧看，中小企业、AI创业团队、高校科研用户和行业客户希望以更便捷、灵活、普惠的方式获取算力，但传统算力采购前期投入高、部署周期长、技术门槛高、弹性不足，难以匹配AI应用快速验证、快速部署和持续迭代的节奏。

因此，无锡需要一个能够连接区域算力资源、模型能力和产业需求的城市级调度运营平台，将算力从基础设施资源转化为可被普遍调用、灵活获取的公共服务能力和产业服务能力。

● 用户价值

- 调度规模与运行效率：平台目前支持并发调度任务 300 个，资源利用率约 45%，调度成功率约 90%；后续目标提升至并发 3000、利用率 80%、调度成功率 99%，逐步实现多架构资源的规模化统一调度。
- 多元客户群体覆盖：通过算力超市与模型超市，平台同时服务企业客户（按需采购训练与推理算力）、高校与科研机构（科研项目算力补给）、政府与行业用户（产业智能化应用），单一平台覆盖多类用户需求。
- 降低用算门槛：客户无需自行承担硬件采购、环境部署、集群运维和底层适配工作，从传统的"长周期自建采购"转向"线上下单、按需弹性"，显著缩短AI应用的研发、验证与上线周期。
- 支撑国产算力规模化应用：平台推动 AMD、沐曦、昇腾等多元算力进入真实AI训练、推理和行业应用场景，提升区域AI基础设施的自主可控能力和产业韧性。
- 典型应用场景落地：平台已为大模型训练推理、代码生成、数据标注、图形渲染、仿真计算、行业智能体等场景提供算力支撑，帮助企业、高校和科研机构完成模型开发、应用验证和场景部署。

● 项目介绍

云工场科技携手无锡数据集团共建无锡算力调度运营平台，按"一体两翼"架构推进。

一体为城市级算力调度平台，承担区域算力资源的统一纳管、智能调度、交易撮合与市场化运营职能；

两翼为算力超市与模型超市——算力超市面向企业、高校、科研机构和开发者提供多元化算力服务，模型超市提供模型能力调用、行业模型服务和AI应用支撑能力。

平台坚持市场化运营与公共服务支撑并重，通过"统一纳管、统一调度、统一交易、统一服务"机制，推动政府资源统筹、企业按需用算、社会算力释放。

● 解决方案核心亮点

- 跨架构异构纳管能力：平台已实现对 AMD、沐曦、英伟达、昇腾四类主流芯片路线的统一接入与调度，自研异构算力适配层封装 ROCm/CUDA/CANN 等不同软件栈，向上提供标准化接口，上层调度系统无需感知底层芯片差异。
- CPU 与 GPU 混合调度机制：区别于单纯 GPU 智算平台，平台将 X86 CPU 纳入混合异构算力体系，使 CPU 不仅是辅助资源，更可参与轻量推理、数据处理、任务编排和分布式计算，与多路 GPU 协同构建混合算力池，提升整体资源利用率。
- 万卡级集群运营能力：依托中国首个 AMD 万卡智算集群和无锡沐曦集群（千卡已点亮、万卡建设中）的实战经验，平台具备大规模国际主流芯片与国产芯片集群的运维、调度和稳定性保障能力，支撑7×24小时连续训练与推理任务。
- 算力超市与模型超市双引擎：算力超市解决"算力怎么找、怎么选、怎么用"，模型超市解决"模型怎么调用、怎么适配、怎么落地"，形成从算力供给—模型服务—场景应用的完整闭环。
- “四统一”市场化运营机制：通过统一纳管、统一调度、统一交易、统一服务，将技术调度系统延伸为城市级算力运营平台，连接政府统筹、社会算力释放与企业用算需求。

6.2 厂商简介及产品服务——云工场科技

● 厂商简介

江苏云工场信息技术有限公司（以下简称"云工场科技"）是国内领先的边缘计算与智算服务商，长期深耕边缘计算、异构算力调度与AI算力服务运营领域。

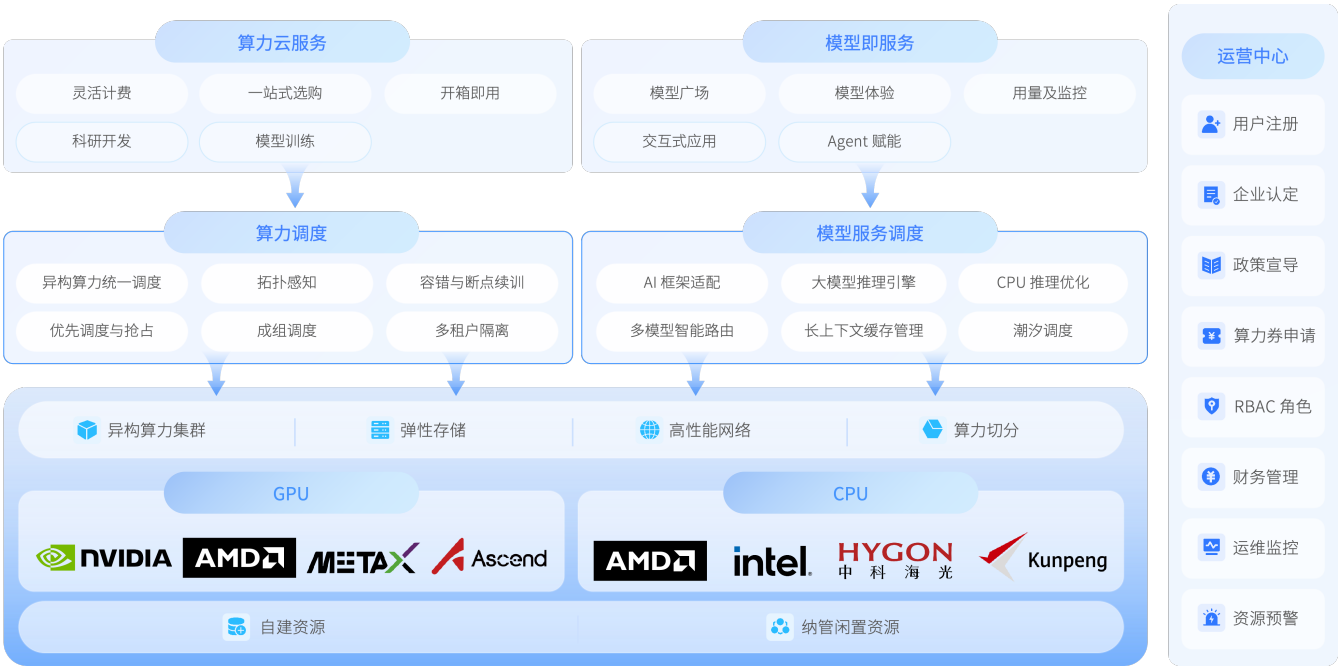
在算力基础设施层面，云工场科技已建成 2000+ 边缘节点和 9 个智算中心。2025年，公司主导建成中国首个 AMD 万卡智算集群；2026年，在无锡推进沐曦万卡集群建设，其中一期千卡项目已点亮并投入运营，形成覆盖国内外主流芯片路线的异构算力资源底盘。

依托自研异构算力调度平台，云工场科技构建了以算力超市、Token服务为核心的算力服务体系，将底层算力资源转化为可交易、可调用、可计量、可运营的标准化服务，帮助客户降低算力获取门槛、提升AI应用落地效率。

● 产品服务

异构算力调度平台

- 资源接入层：支持 AMD、沐曦、英伟达、摩尔线程、寒武纪等国内外主流芯片路线的算力资源统一纳管，解决不同芯片架构、不同集群环境、不同地域资源之间管理分散、调用复杂的问题。
- 调度管理层：基于任务类型、资源状态、芯片适配、性能需求和成本目标，对训练、推理、模型服务等任务进行智能匹配与动态调度，提升算力资源利用效率，降低资源闲置和重复建设成本。
- 服务输出层：依托算力超市与 Token 服务，将底层算力资源转化为标准化的算力商品和模型调用服务，客户可按需查询、订购、调用和计量。



科智咨询

科智咨询是国内领先的数字领域专业咨询机构，致力于为政府、企业和社会组织提供全方位的数字科技战略规划、数字行业与市场洞察、投资决策与管理等服务，助力企业把握数字经济发展机遇、构建数字化核心能力，在新一轮产业和技术革命浪潮中取得竞争优势。

如欲了解更多信息，请联系电话：13161291179

宁夏东数西算产融研究院

宁夏东数西算产融研究院以“整合资源、推动创新、促进产业”为宗旨，积极发挥平台作用，协同宁夏东数西算工程相关产业发展，形成了课题研究与产业规划咨询、创新研发与技术平台建设服务、招商引资与创新企业孵化、人才培训及行业交流四大业务体系，全力打造宁夏算力产业的智库和桥梁，助力宁夏数字化产业发展。

中国通信工业协会数据中心委员会

中国通信工业协会数据中心委员会（简称“CIDC”）在相关部门指导下成立，致力于数据中心生态环境建设，推动国内数字化产业发展。协会将协同数据中心上下游企业展开诸多探索与实践工作，推动数字经济和实体经济深度融合，为实现“碳中和、碳达峰”目标及数字产业高质量发展目标做出应有的贡献。

如欲了解更多信息，请联系电话：18600716466

免责声明

本研究报告由科智咨询联合宁夏东数西算产融研究院、中国通信工业协会数据中心委员会共同撰写和制作。报告仅作为研究内容介绍之用。未经许可，不得对文中内容进行任何形式的复制。尽管我们对所依据的信息和资料保持高度谨慎，但无法对其中的准确性和完整性做出绝对保证，请勿绝对化地加以利用。文中所述观点有可能在未经知会的情况下进行调整。报告内容亦非根据任何公司所处独特环境而提供的具体咨询建议。如果您需要就文中提及内容获得具体建议及进一步的详细资料，请与我们进行联系。