智算产业发展研究报告

(2024 年)

中国电信研究院（天翼智库）

2024 年 9 月

编制说明

主编单位：中国电信研究院

参编单位：深圳海兰云数据中心科技有限公司

中国电信股份有限公司上海分公司

中国电信股份有限公司湖北分公司

顾问专家：

中国电信研究院战略发展研究所 所长：饶少阳

编委成员：

孙雪媛、陈元谋、李思思、赵静、熊小明、谢林翰、马腾滕、王田媛、魏玥

王勇、傅蓉蓉

廖志鹏

张能

智算产业发展研究报告

（2024 年）

目 录

一、全球智算产业新动向

1、智能算力将成为 AI 发展的关键支撑与引擎

（一）AI 推动算力需求超线性发展，智能算力需求井喷。根据 EPOCH AI 数据，全球人工智能经历从传统模型进入深度学习阶段，模型所需算力规模年增长 1.5 倍突破至 4.1 倍/年，算力规模实现 7 个数量级的增长（2010-至今）[1]，其主要在于：

大参数模型、高质量数据集和大算力（智算）成为 AI 大模型发展的关键要素，Scaling 法则带动大模型不断突破新的瓶颈。Transformer 架构成为人工智能的事实标准，模型架构趋于稳定，数据规模、质量双提升，GPU 单卡性能、集群有效算力的持续迭代升级，为大模型的性能、规模爆发奠定了基础，加剧模型训练算力需求指数级增长，目前全球千亿级参数模型所需算力至少为 10^23 数量级，所需算力至少为万卡级规模以上（参考 A100），持续训练周期为周级时长。

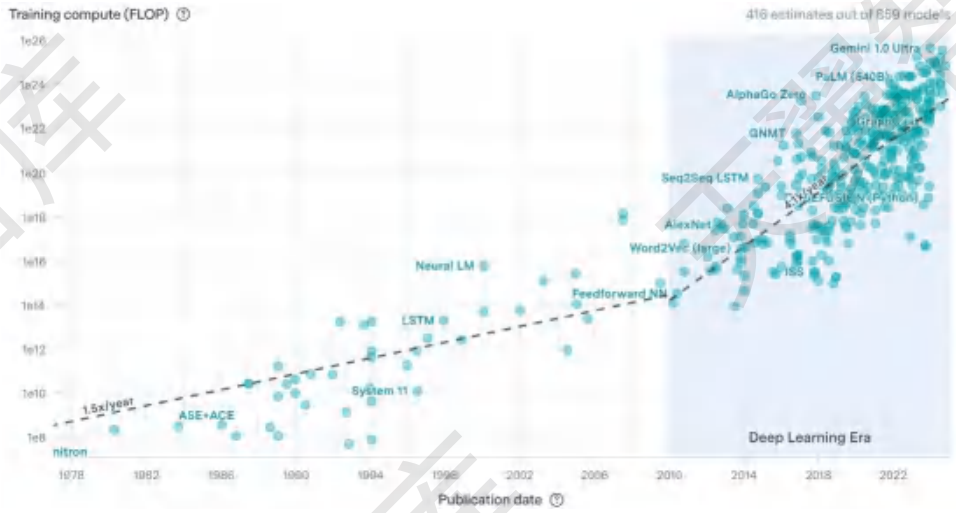

图 1 主流大模型训练算力需求

大模型加速行业渗透，叠加多模态大模型发展，催生算力需求持续增长。随着 AIGC 技术发展，IDC 预测 2024 年全球将涌现出 5 亿个智能化应用，相当于过去 40 年间的应用数总和[2]，在巨大的应用蓝海市场面前，算力缺口显著。Sora、GPT-4o、Gemini 等多模态大模型持续迭代，多模态的海量数据、高清晰度的多轮去噪将带来算力百倍以上的增长。根据测算，与语言大模型（GPT-3）相比，Sora 训练阶段的算力需求是大语言模型 170+倍；推理阶段，即完成一项常规任务，算力需求是大语言模型 600+倍。

（二）智算正在成为我国算力主赛道。2023 年我国算力总规模达到 230EFLOPS，近五年年均增速 30%；智能算力规模达 70EFLOPS，增速超 70%，占算力总规模 30%。2023 年 10 月工信部等六部门联合发布《算力基础设施高质量发展行动计划》，明确了到 2025 年算力规模超过 300EFlops，智能算力占比达到 35%。多省跟进总体目标，上海、广东等省提出到 2025 年智能算力在总算力中占比达到 50%及以上，截至 2024 年 6 月，中央企业智能算力规模同比实现翻倍增长。2024 年 7 月国务院“推动高质量发展”新闻发布会上，国资委 AI 赋能产业焕新工作计划“有序推进智算中心和算力调度运营平台建设，做强智能算力供给”。

2、AI 投资热潮推动智算产业进入快速增长期

多国政府将 AI 基础设施上升至国家战略，持续加大投资及政策支持。面对新一轮人工智能引发的科技竞赛，各国纷纷加码支持，针

对 AI 基础技术、算力基础设施等出台一揽子投资举措。**美国**瞄准细分领域，2024 财年 AI 投资预算增长至 31 亿美元，创历史新高；对通用人工智能基础技术保持高强度投入、算力基础设施等资助力度逐年增强，并于 2024 年 9 月美国政府宣布成立 AI（人工智能）数据中心基础设施工作组，该工作组将由国家经济委员会、国家安全委员会和白宫副幕僚长办公室领导，以协调政府各部门的政策，进一步打造 AI 基础设施的全球领先优势。**加拿大**政府 2024 年 4 月宣布 24 亿加元 AI 投资举措，通过 AI 开发和使用来提高生产力。其中 20 亿加元为 AI 研究人员、初创企业和规模化企业提供算力和技术基础设施。**欧盟** 2024 年 8 月正在推进设立“人工智能工厂”，鼓励成员国建设人工智能基础设施建设；欧委会将向数字欧洲计划拨款 8 亿欧元，用于购买新的 AI 专用计算资源或升级现有基础设施。**沙特**致力于成为全球人工智能中心。2024 年 3 月沙特阿拉伯称将启动成立一只高达 400 亿美元的国家基金，专注于人工智能领域投资，特别是对初创企业、芯片制造商和数据中心扩张。沙特阿美旗下风投公司 Wa’ed Ventures（实际上是沙特的投资工具）近期围绕 AI 芯片、AI 平台持续投资，加快沙特在全球 AI 竞赛中布局。

国家加大力度支持算力产业发展，重点聚焦国产算力。政府通过“国家大基金”、“算力券”“模型券”等方式支持 AI 产业发展。国家大基金目前已进行到第三期，目标推动国内集成电路产业及半导体产业的持续发展、促进半导体以及芯片技术的创新。2024 年 2 月中央企业人工智能推进会上，国资委强调“中央企业要把发展人工智

能放在全局工作中，统筹谋划加快建设一批智能算力中心，国产算力进一步加速”。3 月两会上，政府工作报告首提“加快形成全国一体化算力体系”。

资本市场加大 AI 投资力度，基础设施及模型成为资本关注重点领域。从全球看，2024 年上半年 AIGC 行业投融资总额达 1384 亿元，较去年同期增长 23.3%；累计发生投资事件 363 次，同比猛增 307.9%。国内方面，AIGC 行业融资金额约为 221 亿人民币，同比增长 192.4%；累计发生投资事件 170 次，融资次数快速增长了 178.7%。到目前为止，大部分资金都投向 AI 基础设施和基础模型（Foundational layer）；七家科技巨头[1]（下图所示 M7）已将 AI 作为竞争焦点，尤其是 AI 硬件层和 AI 模型层的竞争正在升温[3]。摩根士丹利认为，算力周期的发展进入第二阶段，红利逐渐从芯片转向基础设施，具体包括服务器、网络设备、冷却系统、数据存储等。

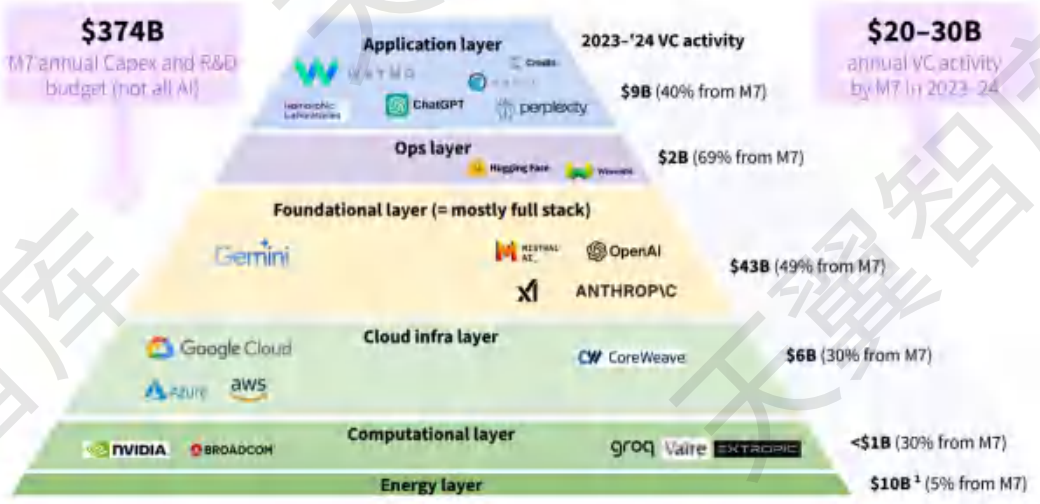

图 2 AI 投资分布情况

[1] M7 指的是 Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia & Tesla

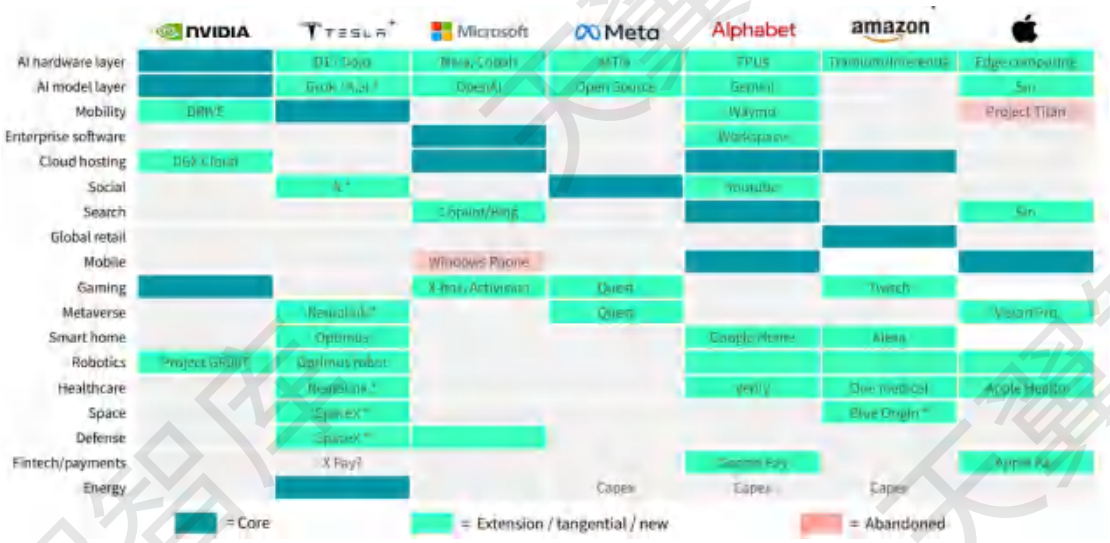

图 3 M7 人工智能布局情况

头部科技公司持续加强全球化 AI 基础设施布局。OpenAI 正计划在美国投资数百亿美元用于建设 AI 基础设施，并寻求美国政府支持，组建全球投资者联盟；微软计划投资超 100 亿美元在英国、日本、印尼、法国等国家地区建设人工智能中心、扩大 AI 基础设施、培训 AI 劳动力技能等；三星电子表示将在美国、韩国成立半导体 AGI（通用人工智能）计算实验室；字节投资 21.3 亿美元在马来西亚建立人工智能中心。

3、智算产业开启“速度”与“质量”并行

在人工智能领域，各国持续加大监管行动的开展，全球加快推进“人工智能治理”。根据世界经济论坛（WEF）发布《生成式人工智能与国际贸易分析》数据，自 2023 年 1 月以来，有超 600 项针对人工智能供应商的监管行动[4]。**美国开启“全面立法”治理人工智能。**2023 年 1 月，美国商务部国家标准与技术研究院发布《人工智能风险管理框架》，旨在对人工智能系统全生命周期实行有效的风险管理。

2023 年 10 月，发布总统行政令《安全、可靠和值得信赖的人工智能开发和使用》，明确了美国政府治理人工智能的政策法律框架，具有里程碑意义。2024 年 5 月，美国参议院人工智能工作组发布《推动美国在 AI 领域的创新：参议院 AI 政策路线图》。**欧盟展现领先全球的立法速度。**2024 年 3 月，其发布的《人工智能法案》被认为是全球首部综合性人工智能治理立法，对人工智能系统进行了分类，并根据风险等级制定了相应的监管要求，设定了普适、全面的人工智能监管方法，包括数据管理、透明度、可解释性、人类监督等方面。同时新设了人工智能办公室，推动实施《人工智能法案》。

“算力治理”成为人工智能治理抓手，倒逼智算产业高质量发展。2024 年世界人工智能大会提出“选取人工智能技术发展中最直观量化的算力作为治理对象，通过算力治理间接实现人工智能治理”。算力治理的提出，反映了对于人工智能发展的全面考量。通过算力治理，可以更好地协调和优化计算资源的分配，提高计算效率，从而提升人工智能系统的整体性能。此外，算力治理还有助于监控和管理人工智能技术的潜在风险，确保其在法律和伦理框架内运行，保护个人信息安全和国家安全。算力治理不仅关乎底层计算基础设施，更是对智算全产业链的完整治理。生态伙伴按治理规则，全面提升智算基建、智算运营、智算安全等一体化智算服务能力。

我国明确“算力一体化”战略，生态联合实现算力产业高质量发展。2024 年政府工作报告明确提出，“加快形成全国一体化算力体系，培育算力产业生态”。多部委牵头联合正在加紧构建“全国一体

化算力网”，推动建设中国式现代化数字基座；建设“全国统一算力服务大市场”，降低中小企业用算成本。三大运营商支撑了全国一体化算力网原型试验场的构建，在建设算网基础设施，深化算网资源布局、创新算力网络服务模式方面发挥了主力作用；阿里云、华为云、天翼云等现已加入全国一体化算力调度平台。

4、智算产业发展的几点认识

（一）算力正在成为塑造未来社会经济发展格局的核心驱动力，其与数据、算法交互融合，正在催生数字化、智能化、生态化的新质生产力形态。

当前人工智能已进入大模型时代，算力、数据、模型快速增长，相比之前的人工智能，大模型具有更好的通用性、更广的应用范围，具备赋能各行各业的潜力，颠覆传统的生产流程、创新模式，引领产业加快向智能化升级，是形成新质生产力的关键力量之一。大模型推动算力爆发式增长，由“算力-数据-算法”三位一体构成的智能化生产力形态，是数智时代最具颠覆性的“新质生产力”，在各行各业、应用场景中落地生根，释放产业新动能。根据 IDC 预测，到 2030 年人工智能（AI）将为全球经济贡献 19.9 万亿美元，推动全球 GDP 增长 3.5%[5]。

（二）智算规模稳定增长，年均增长 70%左右；能效水平不断提升，每瓦特消耗的算力规模每 2 年翻一倍。

多模态大模型快速发展，推动智能算力需求规模大幅增长。以文

生视频大模型 Sora 为例，Sora 生成 60 秒视频对比 GPT-3 生成 3000 字文本，对应推理计算负荷增加超 600 倍。预计未来 1-2 年，我国智算规模能将维持在 70%以上增长率，到 2025 年有望超过 200EFLOPS。同时，智能算力的高能耗特征日益显著，但整体能效水平将不断优化提升。参照英伟达/AMD 等主流 AI 芯片系列产品，AI 芯片每 W 消耗算力约每 2.2 年翻倍[6]，同时结合数据中心高效制冷模式普及所带来 PUE 优化，预计未来 1-2 年每瓦特所承载算力每 2 年左右可翻一倍。

（三）底层算力具有高价格、高利润特点，智算产业活力难以释放，影响对数字经济的带动作用。

受需求激增、供应链紧张等因素影响，高性能 GPU 呈供不应求发展态势，带动价格持续飙升，如英伟达 H100 首发价为 3.5 万美元，二手市场一度炒到 5 万美元甚至更高。面对商业模式不清晰、收入无法覆盖成本的困境，GPU 的高价格导致智算产业中上游成本压力巨大，不利于智算产业的长期、可持续、高效发展。随着智算集群由千卡向万卡/十万卡发展，AI 加速卡及系统搭建成本占据模型训练比例高（60%~）[7]，未来将逐步向平台及应用转移。该现象将进一步加剧。此外，GPU 售价远远高于成本，如英伟达 H100 的售价约是成本的 10+倍，英伟达 2024Q1 毛利水平达 79%。同时，GPU 的高利润促使其上游封装企业台积电持续提价，如先进封装 2025 年报价涨幅约 10%-20%。随着英伟达 GPU 持续供不应求，溢价趋势将持续向上游转嫁，制约智算产业释放活力。

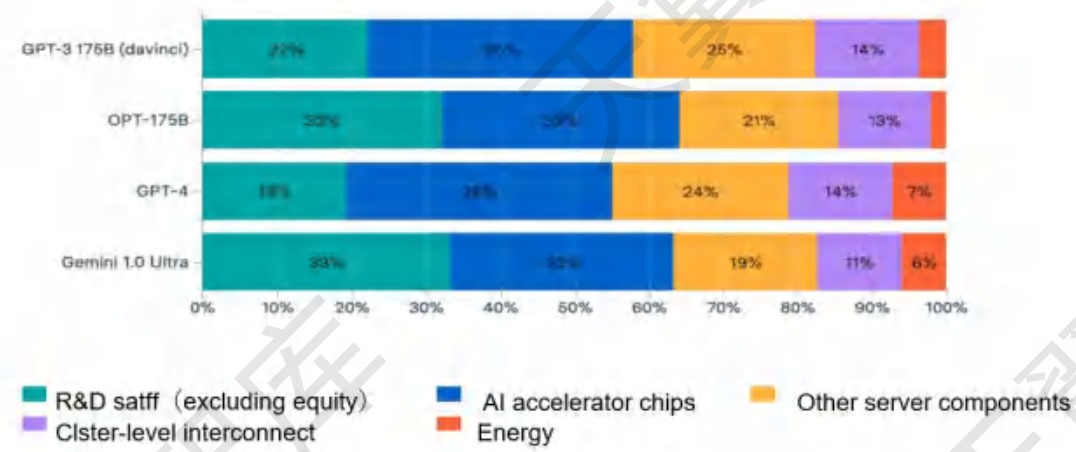

图 4 典型模型训练成本对比明细

二、智算产业图谱

国内人工智能技术加速向行业渗透，叠加西方对我国的算力封锁持续加码，国产智算产业链正在加速构建。目前已形成纵向包含智算基础设施、大模型平台服务、行业应用；横向提供算力资源集成服务的“三纵一横”庞大产业链。如图所示：

图 5 智算产业图谱

1、智算基础设施：全球建设如火如荼，国产化进程加速

智算基础设施是产业链上游，包括 AI 芯片、AI 服务器、智算中心等；上游供应商主要以租赁形式为客户提供一站式智算服务，包括面向生成式 AI 和非生成式 AI 两个细分市场。据 IDC 报告，2023 年下半年智算基础设施市场规模 78.1 亿元，同比增长 86%；从全年情况看，2023 年较上一年全年智算服务市场增长 81.6 亿元，生成式 AI 市场增量贡献 59%，是主要驱动力；非生成式 AI 市场仅贡献 3%的增量[8]。

智算中心建设提速，呈现集群化、灵活调度特征。根据赛迪顾问数据显示，2024 年上半年，全国已经建设和正在建设的智算中心超过 250 个，招投标事件 791 起，同比增长高达 407.1%；已有超 20 个城市建设了智算中心，其中既有北京、上海等一线城市，也有郑州、武汉等中心地区城市，还有内蒙、宁夏等西部城市，目前正在加速向县城下沉。从建设主体来看，头部大模型厂商与云商选择适度超前建设超大规模智算集群，如 Meta 为训练 LLaMA3 构建 2 个 2.4 万卡集群，并计划 2024 年底搭建 35 万卡集群；字节跳动、百度等已建立万卡级智算集群，主要面向自有大模型训练。截至 5 月底，全国规划具有超万张 GPU 集群的智算中心已达十余个。其中，三大运营商加大国产 GPU 的应用，预计 2024 年将建成 6 个万卡集群智算中心，已投产使用近 3 万张国产 GPU。

云实现智能算力资源的灵活调度。信通院联合中国电信打造“全国一体化算力算网调度平台”，是我国首个实现多元异构算

力资源调度的全国性平台，目前天翼云、华为云、阿里云等云商均已接入。除此之外，云商也纷纷基于自身云底座实现算力调度，如青云打造“AI 算力调度平台+AI 算力云服务”，前者打破算力边界，延申算力从中心到边缘到端侧，对算力资源进行统一调度、管理、运营；后者基于青云的云服务，联合生态共同对外提供 AI 产业链上的服务。

AI 芯片供不应求，国产化万卡级集群量质同步提升。据 Gartner 预测，2024 年 AI 芯片市场规模将较上一年增长 25.6%，达到 671 亿美元；预计到 2027 年，AI 芯片市场规模将达到 1194 亿美元。英伟达凭借成熟的 CUDA 软件生态，垄断高性能人工智能芯片市场，2023 年在数据中心 GPU 占据了 98%的市场份额，总收入达 362 亿美元（约 2626.9 亿元）；自身芯片产业加速形成。根据 IDC 数据，截至 2023 年底，我国自主 AI 芯片市场占比已达 14%，年提升 4 个百分点，国产化能力稳步提升。国产 AI 芯片厂商面向 AI 推理和 AI 训练需求持续发力，芯片架构呈现出多元化，国内多个万卡级智算集群投入运营，自主能力不断提升。运营商依托华为 910B 构建多个万卡级集群，燧原科技国产万卡集群在庆阳点亮。

2、大模型平台服务：国内外云厂商模式创新，差异化布局

大模型平台服务是产业链中游，供应商提供 API 等工具供用户灵活调用其基础大模型，并针对不同业务场景，开发、训练和

部署专属大模型。据 NTCysd 预测，到 2026 年我国 MaaS（模型即服务）市场规模有望突破 130 亿元；2023-2026 年复合增长率为 117.9%[9]。云平台提供从数据、模型到应用服务的全周期管理和工具。

国外云厂商的 MaaS 服务专注于构建通用能力。国外云厂商利用自身基础设施优势，开发全流程工具和套件，满足用户对模型预训练、模型精调、模型部署、智能应用开发等多样化需求。**微软云的 Azure OpenAI 服务**，支持开发者调用 OpenAI GPT-4、GPT-3、Codex 和 DALL-E 等模型的 API 来构建、微调模型，Azure 主要提供一些通用型功能，如安全性、合规性和区域可用性等。**亚马逊 AWS 的 SageMaker 服务**，为大型语言模型提供了全生命周期管理工具，研发者可用它进行大模型的训练、微调和部署，并且与 AWS 的其他服务无缝集成。**MaaS 服务极大带动了云商收入增长。**Microsoft、Google、Amazon 云收入增速自 23Q3 逐步提升，24Q1 三家公司云收入同比增速分别为 31%、28%、17%。

国内云厂商的 MaaS 服务强调搭建生态。除了通用能力外，国内云厂商还会参与开发垂直行业大模型、集成软硬件服务。**腾讯云**依托 TI 平台打造一站式行业大模型精选商店，其中包含了腾讯企点、腾讯会议、腾讯云 AI 代码助手等多款头部 SaaS 产品，并启动与金融、文旅等数十个行业客户共建行业大模型。**华为云**在其 MaaS 服务平台“ModelArts”上推出了“昇思大模型服务”，

支持跨平台的模型部署与推理，用户可一键式远程调用昇思 NPU 芯片的海量算力，大幅缩短推理等待时间，避免在本地部署 NPU 芯片的繁重操作。**百度云**基于其 MaaS 服务平台“千帆”，推出了千帆 AI 原生应用商店，成为大模型商业机会的汇集地，为商家提供品牌曝光、流量支持和销售资源等支持。

3、行业应用：全球进入应用元年，智算能力全面升级

我国大模型数量及规模快速增长，能力逼近 GPT-4。根据信通院《全球数字经济白皮书（2024）》数据，全球人工智能大模型 1328 个（包含同一企业、同一模型的不同参数版本），美国数量占比 44%位居第一，我国数量占比 36%位居第二[10]；据大模型之家预测，预计到 2028 年全球大模型市场规模将达到 1095 亿美元；我国大模型市场将达 1179 亿元。百度文心一言、讯飞星火、清华智谱 ChatGLM4、商汤“日日新 SenseNova5.0”整体表现逼近 GPT-4，零一万物的千亿参数闭源大模型 Yi-Large 在最新总榜中排名世界第七，中国大模型中第一，超过 Llama-3-70B、Claude 3 Sonnet；其中文分榜更是与 GPT-4o 并列第一。

科技巨头加大 AI 产品赋能，ToB 市场初见商业化成效。微软旗下的 GitHub Copilot for Business 已有超过 2.7 万家组织用于提高开发人员生产力，超过 6.3 万家组织在 Power Platform 中使用了 AI 驱动的功能且环比增长 75%；微软将旗下的 Microsoft 365 Copilot 定价提升至 30 美金/人/月。Salesforce 从推出首个

生成式 AI 客户关系管理应用 Einstein GPT 开始，全面拥抱生成式 AI，推出 AI Cloud 系列应用；并随着 Einstein 模块用量增大，宣布提价。

我国大模型正加速迈向行业纵深，赋能应用场景。从行业看，据赛迪《2024 年中国人工智能行业典型大模型 100 强企业》披露，金融、工业、政务三大行业因信息化基础好、数据结构化程度高、应用场景相对成熟，100 强企业在这三个行业有落地应用的企业数量最多；医疗、交通两大行业数据丰富、应用场景众多，较多企业开展探索，处于第二梯队[11]。从场景看，腾讯研究院总结大模型落地的快慢呈现“微笑曲线”的特征，即产业链高附加价值的两端（研发/设计和营销/服务），大模型应用落地较快；而在低附加价值的中部（生产、组装等），大模型应用进程较慢[12]。

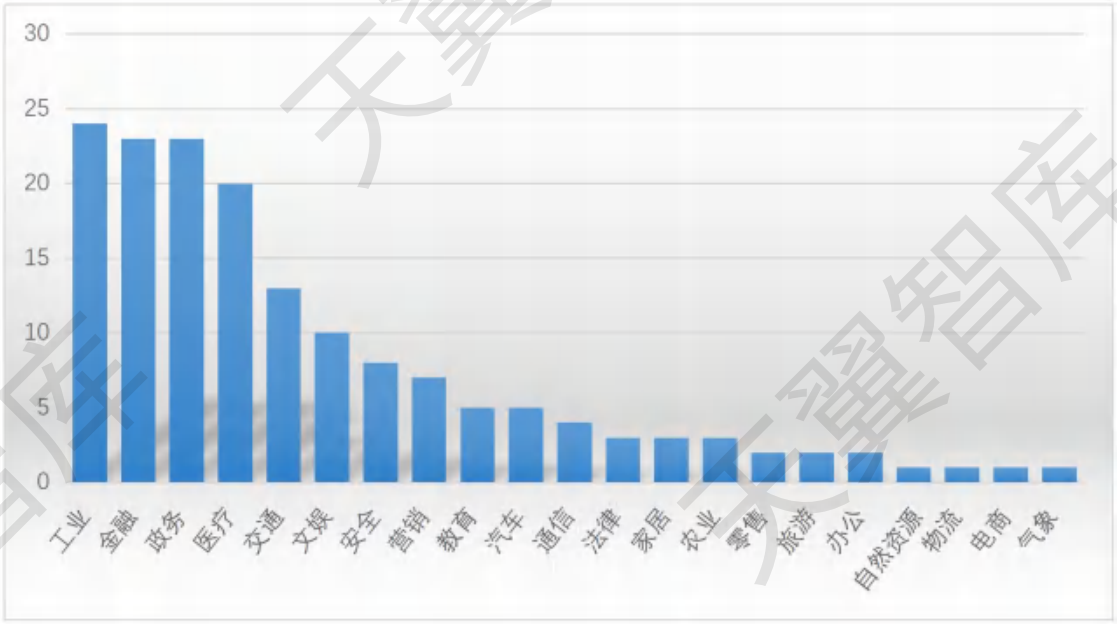

图 6 百强企业大模型分布

金融行业是 AI 渗透率领先行业，大模型驱动金融业务场景革新。据前瞻产业研究院预测，2025 年我国智慧金融市场规模将

达 3638 亿元[13]。金融行业由于数字化基础好（数字化转型超 90%以上）、具备支付能力（行业本身盈利能力强）等特征，成为目前 AI 大模型覆盖最多的行业。AI+金融应用目前已覆盖产品设计、市场营销、风险控制、客户服务、支持性活动，形成覆盖全生命周期、一系列配套的解决方案，推动金融行业高质量发展。如中国人寿联手第四范式基于 AI 大模型打造智能核保系统，实现了自动、精准、个性的核保决策；打造智能理赔系统，以最快速度自动、公正处理理赔，有效降低理赔成本和风险、提升理赔效率。建设银行启动“方舟计划”，打造具备大模型、大算力、大数据的金融大模型基座与能力体系，推进 AI 技术在智能客服、智能运营、智能风控等场景的应用，渠道综合化运营，旗舰类、综合类网点占比提升 2.86 个百分点，基层网点减负赋能成效显现。

医疗场景融合 AI 将释放丰富价值，目前 AI 医疗项目建设方式多样化。据前瞻产业研究院预测，2028 年我国智慧医疗市场规模将达 2332 亿元[13]。AI 大模型可赋能医疗行业“医、教、研、管”等场景的各个环节。医疗机构：可提升管理效率；医护群体：可辅助科学诊疗；制药企业：可降研发成本提生产效率；病患群体：可缩小医疗资源不均矛盾，因此医疗行业可借助 AI 实现商业、人文等丰富价值。友谊医院联合信通院成立“算力+医疗健康工作组”，共同探索算力在医疗健康领域的应用场景与发展趋势；医渡科技与华为合作启动医疗大模型联合创新，基于昇腾 AI 硬件，推出面向 B 端的训推一体机解决方案，医院可以根据自身需求来

购置相应设备；国内 AI 制药头部企业晶泰科技选择自主建设 AI 药物研发所需的算法平台与高性能计算算力平台。

工业算力分布在云-边-端，但应用普及率仍有较大提升空间。据 Capgemini 统计数据显示，欧洲顶级制造企业的 AI 应用普及率超 30%，日本和美国制造企业的 AI 应用率分别达到了 30%和 28%。相较于这些发达国家，我国制造企业 AI 普及率尚不足 11%。目前我国工业算力有四种典型应用场景，（1）边缘算力：多个计算能力较弱的工业终端，将计算任务或数据迁移到邻近的边缘计算设备，实现数采、分析、检测、控制等功能。（2）云化服务：将云资源池以容器或虚机的形式划分出来，远程为工业产线提供应用服务，具有灵活重新配置、成本较低和软件故障恢复快等优势。（3）群智算力：是指在缺乏边缘计算和云计算资源时，利用存在计算/数据依赖的多个生产设备之间，调整任务分配，使得整个设备集群的计算任务具有实时性。（4）算力协同：该模式充分利用了边缘计算的实时性和云计算的大量资源，可以逐级部署计算任务，在计算能力和实时性之间取得折中。

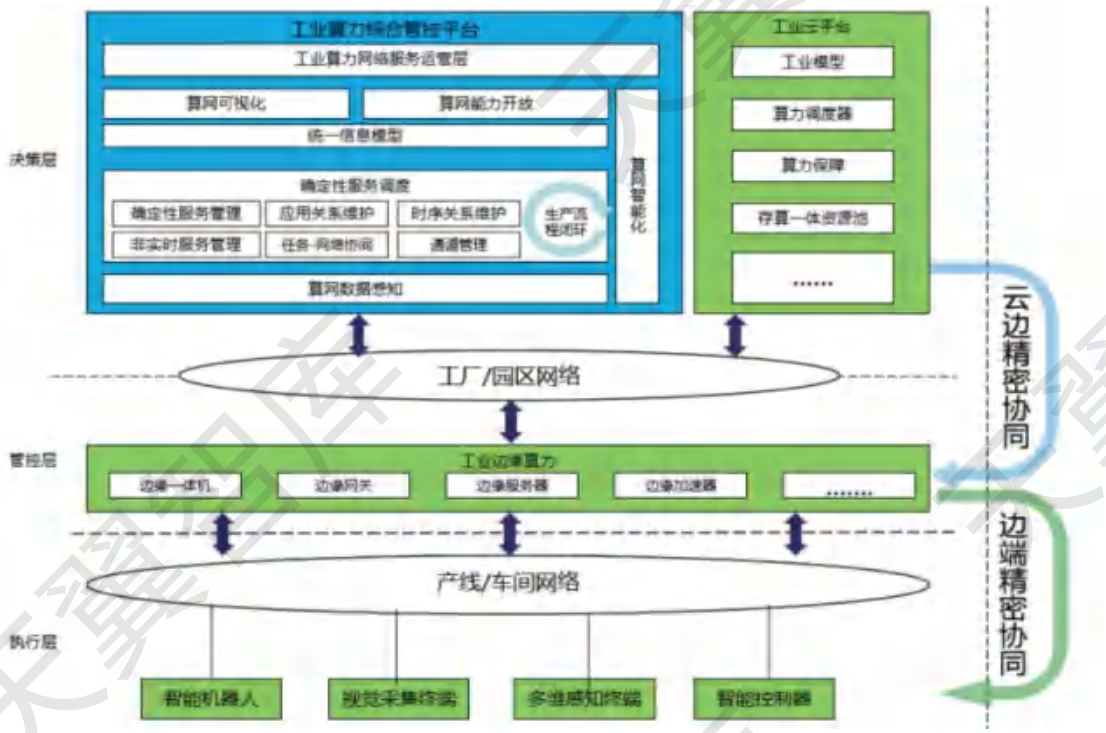

图 7 新型工业算力基础设施

4、智算集成服务：ToB 市场火热，智算集成释放巨大潜力

智算集成服务主要是指厂商在帮助客户建设私有智算基础设施过程中提供的咨询、集成、开发、运维等专业和管理服务。据 IDC 报告，2023 年下半年智算集成服务市场规模 36 亿元，同比增速 129.4%[8]。

智算集成服务市场呈现出“一超多强”的特征。华为依托其领先的芯片能力及全栈服务能力，市场份额领先，同时，华为可为客户提供完整的从咨询规划、平台建设、模型开发、集成实施到辅助运营等全生命周期服务，通过 3+1 算力产业体系—“算力、存力、运力”基础设施以及智算服务，助力客户打造多样性算力中心；新华三通过图灵小镇（产业链式发展）模式、百度依托建、管、运的服务式思维不断取得各地政府的认可；寒武纪同样依托

其领先的推理芯片及全栈服务参与多地台州、沈阳等多地算力基础设施建设项目；中国电子云依托“CECSTACK V5 一体化算力平台”为客户提供智算和高性能计算基础设施，相关智算中心项目目前已在北京、石家庄、武汉等地正式落地。

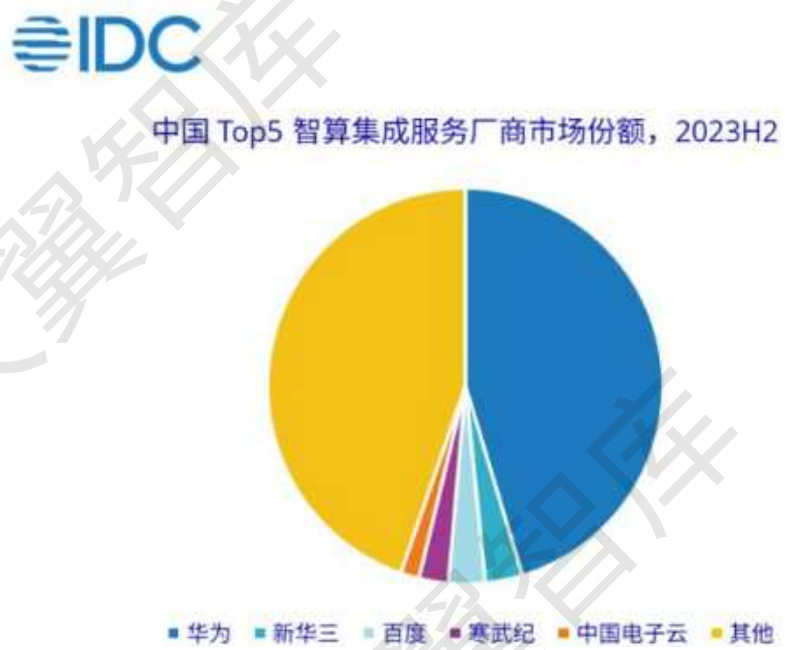

图 8 2023H2 中国 Top 5 智算集成服务厂商市场份额

大模型降价抢占 ToB 市场，推动大模型在各行各业应用落地。麦肯锡的研究报告显示，应用生成式 AI 大模型每年为企业端（即 2B）带来的经济价值为 2.6-4.4 万亿美元[14]，大模型对 ToB 用户吸引力旺盛。在掌握“企业一旦使用一家大模型，替换成本极高”的普遍规律后，大模型厂商通过降价提前卡位，推动自身大模型产品被更多 B 端企业应用，建立数据飞轮，强化用户粘性，进而加速构建 AI 开发生态。5 月字节豆包大模型降至 0.0008 元/千 Tokens；紧接着阿里宣布其主力模型全面降至 0.0005 元/千 Tokens，其通义旗下的 12 款模型已开源，全部免费下载；百度紧跟上宣布最新两款主力模型“免费，立即生效”；而后参与者又

多了科大讯飞和腾讯。面向开发者打造社区生态。百度推出的一众包括飞桨在内的开发社区；阿里的魔搭社区致力于以开源力量助力中国类 Sora 模型的探索和创新。

三、智算发展七大趋势预判

趋势一：软硬协同优化助力大模型降本增效

如何平衡性能和成本成为大模型发展面临主要难题。根据斯坦福大学发布的《2024 年人工智能指数报告》，训练成本随着模型规模增加而急剧上升，如 2017 年的 Transformer 模型训练成本约为 900 美元，而 2023 年的 GPT-4 和 Google 的 Gemini Ultra 的训练成本分别约为 7800 万美元和 1.91 亿美元[15]。

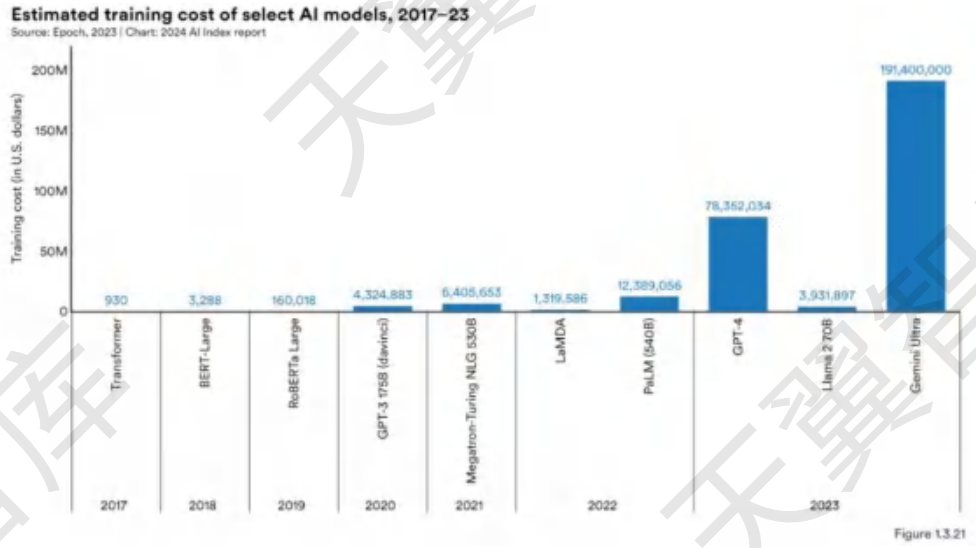

图 9　2017-2023 年部分 AI 大模型的训练成本估算

追求长期降本和供应安全，大型企业发力 AI 芯片自研。随着 AI 算力需求和英伟达芯片价格持续高涨，国外头部云商及科技企业加快自研 AI 芯片，如谷歌 TPU v5、微软 Maia 100、亚马逊

Trainium、Meta 的 MITA V1 等。考虑到供应链的稳定性，国内芯片厂商及大型企业加快推进 AI 芯片国产化进程，如华为昇腾 910B、寒武纪 MLU370、摩尔线程 MTT S4000、海光 DCU、百度昆仑芯等。

采用“分时复用”策略和系统级优化手段，提升算力利用率和训练效率。一方面，借助需求预测和任务调度等方法，在高峰期给新兴业务分配更多算力资源，而在低峰期将多余的算力资源重新分配给其他业务或用户，以提高整体的算力使用效率。另一方面，通过数据并行、张量并行、流水线并行、分组参数切片并行等多种并行加速策略，结合通信、容灾及监控工具，搭建大模型训练/生产系统，加快大模型训练和调优速度，并降低人员操作门槛，如字节跳动 MegaScale、微软 DeepSpeed、NVIDIA 的 Megatron-LM、清华大学 BMTrain、百度飞桨 PaddePaddle 等。

从模型技术创新角度切入，主流企业加快研发 MoE 大模型以平衡模型升级效果及成本。基于 Transformer 的 MoE 已成大模型主流架构，谷歌、OpenAI、阿里、华为、腾讯、昆仑万维、MiniMax 等国内外主流企业加快推进 MoE 大模型布局和落地。2024 年，全球 MoE 大模型数量呈爆发增长态势。据公开统计，2024 年 1-5 月全球发布 MoE 通用大模型数量约 20 个，远超 2021-2023 三年总量（约 10 个），且以多模态大模型为主（占比约 90%）。

趋势二：高质量数据集是大模型能力跃迁的关键

AI 发展正在从“以模型为中心”加速转向“以数据为中心”。

OpenAI 强调，增加大模型的参数量不再是提升大模型能力的最有效手段，大规模、高质量数据和数据高效处理工程化才是关键。传统“以模型为中心”AI 范式主要围绕模型进行迭代、优化设计，数据集相对固定，“以数据为中心”范式更侧重于提升数据集的数量、质量，关注数据集本身，模型相对固定。增加百科、书籍、期刊等高质量、大规模、多样性的数据集占比对于提高模型精度、可解释性和减少训练时长效果显著。如 GPT-4 相比 GPT3 训练数据规模提升约 40 倍（达 13 万亿个 token），Llama 2 相比 Llama1 相比，训练数据规模增加 40%(达 2 万亿个 token)。

目前数据集的市场需求以定制化服务为主。相关数据显示，2021 年我国数据标注及审核市场中定制化服务占比 85.41%，而标准化的数据集产品仅占 13.33%。大模型时代下，“基础模型+微调”成为 AI 开发新范式，微调是让 AI 获取特定领域知识，并赋予其组织、应用知识的能力，可以预见，贴合垂直场景的高精准定制化数据标注服务在未来将是市场需求主流。

合成数据是模型能力跃迁的关键。根据 Epoch AI Research 团队《Will we run out of data? Limits of LLM scaling based on human-generated data》，当前的存量数据中，高质量数据将在 2026 年耗尽，低质量数据将最晚在 2050 年耗尽，图像数据将最晚在 2060 年耗尽[16]。为了解决高质量数据不足的问题，OpenAI 主要采用合成数据的方法，即借助生成对抗网络（GAN）来生成数据。Gartner 预测，2024 年用于训练大模型的数据中有 60%将是合成数据，到 2030

年大模型使用的绝大部分数据将由人工智能合成。合成数据因其高质量、高垂直的特性，将有可能最先在在自动驾驶、金融欺诈、医疗等场景率先应用，并将在 2030 年超过真实数据。目前，英伟达、微软、Meta 以及国内云商等均已在合成数据领域开展布局。

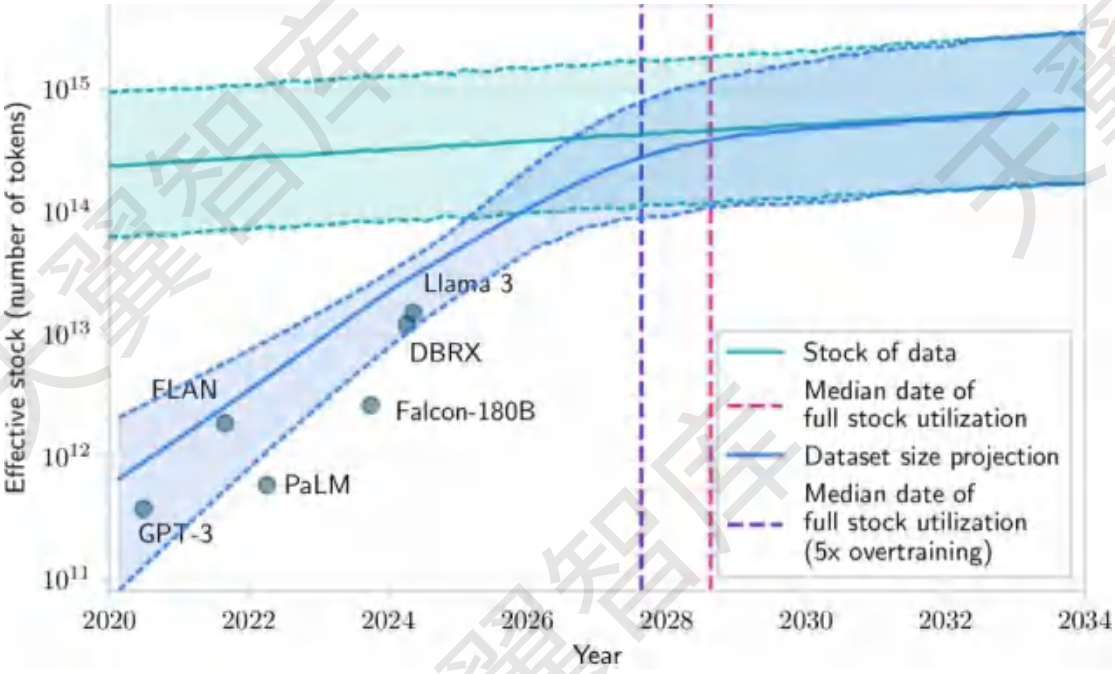

图 10 训练 LLM 的人工生成的公共文本的有效存量和数据集大小的预测

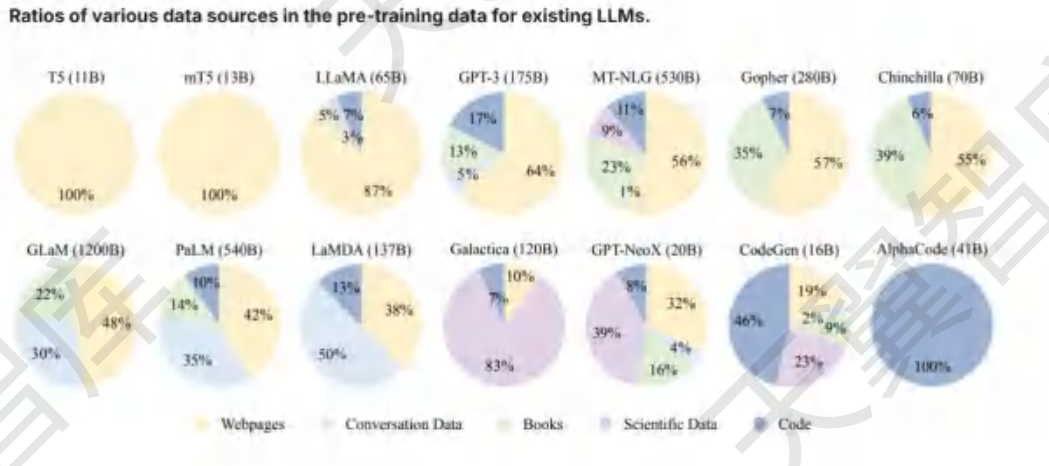

图 11 大模型参数量及其训练所需高质量数据源构成[17]

趋势三：超大规模智算集群成为人工智能发展基石

万卡集群成为大模型军备赛的标配。生成式 AI 的演进推动底层

基础设施不断升级，万卡及以上超大规模智算集群成为人工智能发展的新要求。万卡集群目前代表业界前沿技术能力，主要集中在头部科技公司，指单一集群拥有一万张及以上的计算加速卡（如 GPU、TPU 或其他专用 AI 加速芯片），充分整合高性能 GPU 计算、高性能存储以及网络、智算平台等关键技术，将各类底层基础设施整合成为一台“超级计算机”，主要用来训练参数和数据量越来越庞大的大模型。

智算集群从万卡向十万卡演进。随着模型参数量从千亿迈向万亿，模型能力更加泛化，大模型对底层算力的诉求进一步升级，万卡及以上集群成为大模型基建军备竞赛的标配，有助于压缩大模型训练时间，实现模型能力的快速迭代。如 OpenAI 训练 GPT4 在大约 2 万个 A100 上训练了 90 到 100 天，峰值吞吐量约为 6.28EFlops，若在 10 万 H100 集群的理论峰值吞吐量是 2 万 A100 集群的 31.5 倍，训练时长仅需 4 天。马斯克旗下公司 xAI 已启动 10 万个液冷 H100 GPU 组成的超大规模智算集群。

AI 与超算基础设施融合，推动科研范式变迁与生产效率提升。超算中心广泛引入“处理器（CPU）+加速器（GPU 为主）”的超算异构架构，对超级计算机系统进行设计和改造，解决传统超算效率问题，提速科研领域新发现。根据 2024 年 6 月全球超级计算机性能权威榜单 TOP500 统计，排名前十的超级计算机中有 8 台使用 CPU+GPU 异构架构，其中超级计算机 Aurora、Summit 在系统设计时均面向人工智能场景进行优化。2023 年以来，随着以 ChatGPT 为代表的大模型的崛起，全球主要国家积极将 AI 技术融入超算应用场景，在生物、气

象、材料等领域取得显著成效。如，英伟达和美国制药公司联合开发的面向生物医药场景生成式 AI 模型，将计算性能提升 60%。进一步提升计算效率，拓展新的应用场景。

突破 IB 封闭生态及架构，加强以太网全栈优化。当前用于大模型训练的智能算力节点内部大多采用 InfiniBand 技术构建数据中心内高性能网络，提供高速连接， InfiniBand 技术为英伟达独家控制，成本偏高、开放性较弱，因此业界也在考虑用 RoCEv2 等无损网络技术替代 InfiniBand 技术。2023 年 7 月，由 AMD、Intel、Meta、微软、博通、华为、百度等头部云商、科技公司及硬件厂商等超过 30+家头部企业发起，成立超以太网联盟(Ultra Ethernet Consortium，UEC)，加强以太网全栈协议层及跨层的优化改造，弥补传统网络的不足，打造开放生态的 AI 无损网络，意欲实现 IB 替换。

图 12 UEC 联盟会员

趋势四：算力服务由资源租赁向平台化、一体化供给演进

算力服务由资源租赁向平台化模式演变。随着算力使用方从大型企业扩展到科研机构与中小企业，粗犷式的算力资源租赁服务面临运营门槛高、技术易过时等问题，云商普遍通过算力平台实现专业运营管理和调度等能力，为客户提供稳定、可靠、高效的算力服务，如阿里飞天智算平台，中国电信息壤算力服务平台等。与此同时，各地政府搭建公共算力服务平台，推进全国一体化算力体系和算力大市场的建设。政府算力平台通过广泛吸纳社会通、智、超多元算力资源，结合当地算力券等普惠政策，进一步降低中小企业的算力使用门槛，成

为推进政府算力监测、监督管理的基础底座平台。根据公开信息统计，国内已发布的算力平台超 17 个，其中政府牵头主导的算力平台近 50%，如重庆算力互联公共服务平台 7 月正式上线，首批接入算力提供商 18 家，计划 1 年内聚集算力资源超 1000PFlops。

MaaS 屏蔽底层差异，基于算力服务平台跨越模型供给侧与用户需求侧“鸿沟”，提速大模型应用普惠化。模型即服务（MaaS，Model as a Service）将人工智能算法模型以及相关能力进行封装，构建面向用户场景的一站式服务，有效降低使用人工智能的技术壁垒，弱化人工智能技术的专业性和复杂度，简化人工智能技术的使用和开发难度，大幅缩短开发周期实现更快速的商业落地。国外如 Google 的 AI Platform、微软的 Machine Learning 以及亚马逊的 Bedrock 等平台，提供了从数据处理到模型训练、验证、部署及监控的流水线服务。国内云商 MaaS 平台均支持多种机器学习算法和大模型，并提供低代码开发环境与高效的模型训练及部署能力，能够适应多样化的模型定制需求。火山引擎面向企业用户提供中立的大模型托管平台，通过“机器学习平台+算力”为大模型企业提供算力基础设施及平台，同时通过价格优惠策略快速切入市场。

趋势五：AI Agent（智能体）将成为智能交互的新流量入口

智能终端搭载多模态大模型，加速转型升级。依托轻量化大模型的原生智能终端将成为主流。各大厂商、云商、机器人厂商纷纷加快大模型从云端向到终端转移，如高通推出首个在

Android 智能手机上运行的具有超 70 亿参数的语言和视觉助理大模型 LLaVA。联发科联手 OPPO 和 VIVO，在搭载天玑 9300 芯片的手机上，运行语言大模型 Llama2 和视频生成 AI 模型 Stable Diffusion。华为、达闼等企业已推出 5G 云端机器人，通过云端大模型训练迭代与机器人端侧交互，实现机器人之间的学习和知识共享。小鹏汽车已在智能座舱场景中新增接入阿里云通义千问

大模型赋能终端应用及工具智能升级，AI Agent 逐渐成为大模型行业落地的主要方式。据 MarketsandMarkets2024 调研报告《Autonomous Al and Autonomous Agents Market》预测，至 2028 年全球 AI Agent 市场规模将达到 285 亿美元，2023-2028 年 CAGR 将达到 43.0%[18]。以 AI Agents 形式的模型应用将迎来爆发。目前 OpenAl 的 GPTS 和 GPT-Store 已成为 Agent 生态雏形。字节推出国内版 GPT-Store “扣子”，支持自身云雀、月之暗面（Moonshot）、通义千问等大模型，面向 2C 用户快速创建各类聊天机器人，并一键发布到不同社交媒体与消息应用当中，如飞书、微信公众号、豆包等渠道。阿里钉钉利用大模型重塑 20+条产品线，面向 2B 用户推出 AI Agent 产品提供处理文档、编写方案等一站式助理服务。目前钉钉已上线 200+位 AI 助理，覆盖办公、生活、娱乐等多个场景，已有超过 220 万家企业采用钉钉 AI 助理。

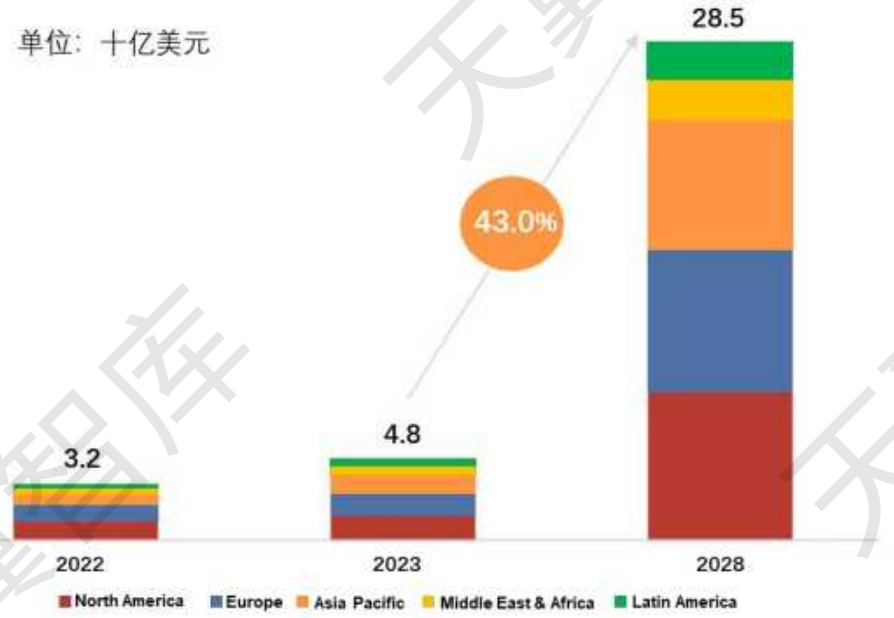

图 13 AI Agent 市场规模

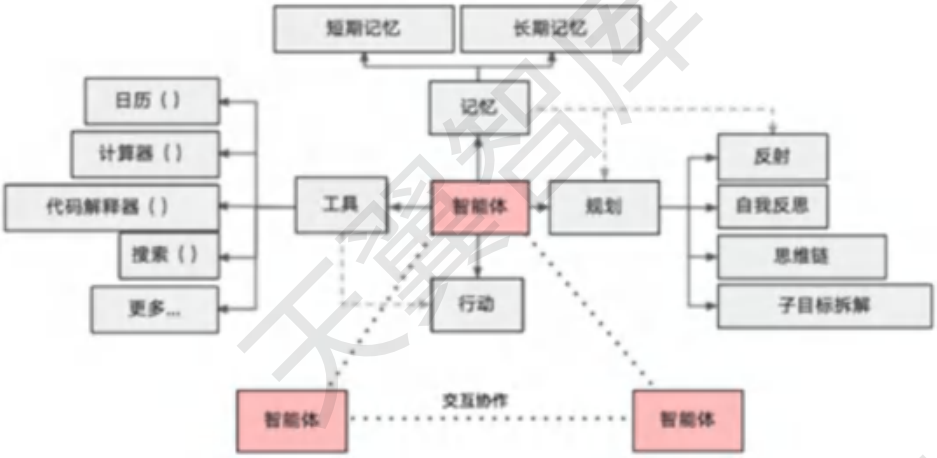

图 14 基于 LLM 驱动的 Agents 基本框架[19]

趋势六：AI 技术设备加速 AIDC 基础设施升级

AI 技术设备高功率、高密度、高弹性对数据中心能源基础设施提出更多新要求。AIDC 相较于传统 IDC，采用高性能 GPU、TPU 等 AI 加速芯片，支持大规模 AI 模型的训练和推理，引发数据中心基础设施**高功率、高密度、高弹性**的能源改变。据统计，现阶段单机柜电力规模已经从原本的十几、几十千瓦，提升至 120KW；部分独立机房的

千卡级集群训练部署平均密度可达 15-20KW，同时适配不同类型芯片和服务器功率，需要数据中心电源解决方案灵活支持高低功率机柜的搭配场景。高算力需求重塑数据中心形态，高功率、高密度、高弹性等特性对电力需求剧增，优化电力供应和配电架构、升级高效制冷散热方案，成为 IDC 向 AIDC 基础设施升级演进的重要趋势。我国近期发布了《数据中心绿色低碳发展专项行动计划》，对 AI 发展驱动下的新建和改造数据中心建设提出了 6 方面的重点任务。

配电制冷弹性建设、绿电储能创新部署、智能化运维管理成为 AIDC 基础设施升级改造的趋势。一是强化资源统筹匹配，推进配电、制冷等配套弹性建设模式。如全面推进智能小母线+弹性方舱的供电资源弹性适配，预留风冷+液冷等接口灵活调度冷量，并根据液冷和风冷补冷的匹配对应气流组织全面优化等。二是加大储能设施创新部署，增强稳定可靠供电能力。储能与供配电系统等协同配合，推动数据中心基础设施绿色低碳转型，实现数据中心备用电源、参与电力市场调峰相应等多重目标。三是引入智能化能效管理能力，推动数据中心精准节能。基于大数据与智能算法实现能耗监控与调优的智能决策引擎，可根据 AIDC 运行状态实时匹配合理的服务容量与资源，助于优化电力使用和规划未来功耗容量需求。

趋势七：算力与电力协同发展成为新态势

解决算力增长和电力消耗矛盾，推动算力和电力协同发展是必由路径。智能算力的高能耗特征日益显著，算力能源消耗呈现

快速增长趋势，预计到 2030 年我国算力用电需求将接近当前的 3 倍左右。超大集群供电承压、东部算电能源短缺、绿电使用占比低成为制约算力发展的三大用能结构“瓶颈”。以算力用电需求为导向，通过政策机制、技术架构等创新，推进算力和电力产业的协同融合发展，成为近年来我国算力发展的政策推进重点，如《算力基础设施高质量发展行动计划》、《数据中心绿色低碳发展专项行动计划》，均明确提出创新算力电力协同机制。

加大算力与电力在布局、市场和调度等方面融合发展成为算电协同重点方向。一是算电布局协同，解决电力资源与算力用能的空间供需不平衡，统筹东西部的电力和算力输送格局，加强电力和算力“双向奔赴”，构建面向算力中心的多层次可再生能源供给。二是算电市场协同，解决绿电交易市场与算力低碳运营不相适应，推进电力市场体系建设，为算力提供可靠绿电来源和有效价格激励，通过市场化机制实现绿电低成本供给。三是算电调度协同，解决新能源发电与算力用能的稳定性不匹配，通过 AI 技术推动两网间根据容量、季节等因素进行时间、空间匹配调度，通过联合调度实现低碳电力最大化消纳。

四、智算技术发展的六大关键词

关键词 1：MoE

混合专家模型（Mixture of Experts，MoE）是一种稀疏门控制

的深度学习模型，具备高效、灵活、可扩展等优势，适合处理大规模、多模态数据和复杂任务。MoE 主要由一组专家模型和一个门控模型组成，在处理任务时只激活或使用部分专家模型，且通过门控模型来控制专家模型的选择、激活和加权混合，从而提升模型整体性能。充分发挥门控模型的协调能力和专家模型的专业优势，MoE 将更适合处理复杂和特定任务。

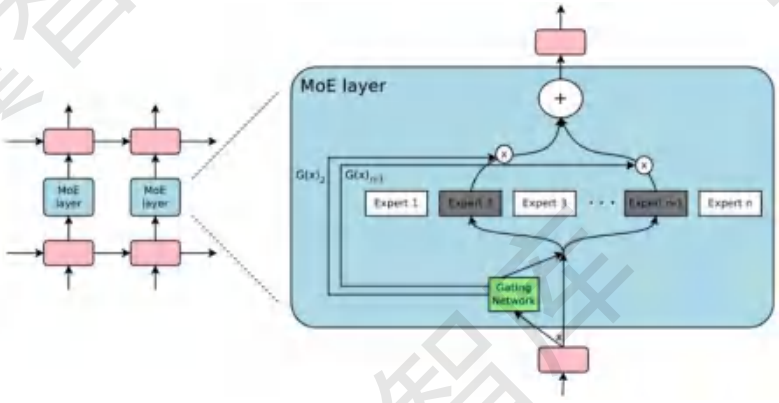

图 15 MoE 架构原理示意图[20]

大模型主流企业正在差异化推进 MoE 大模型布局和落地。一是谷歌、OpenAI 等国外头部企业，引领 MoE 大模型落地，优先推动 MoE 大模型全面布局，同时控制成本。谷歌从算法、工具、模型、硬件到生态全面布局，力争 MoE 发展领航者。OpenAI 推出国际领先大模型同时利用 MoE 权衡性能和成本，如 GPT-4 采用 MoE 优化架构，实现“性能/成本”提升 70 倍[20]。二是阿里、华为等国内头部企业，优先利用 MoE 提升大模型性能以追赶先进水平，同时增强模型可用性。阿里利用 MoE 升级多模态大模型 M6，与同期 GPT-3 相比，M6 实现同等参数规模，能耗仅为其 1%[22]。三是 Mistral AI、Databricks、昆仑万维、MiniMax 等国内外 AI 初创企业，依托 MoE 大模型起家，利用其低成本、高效优势，抢占市场先发优势。被誉为“欧洲 OpenAI”的 Mistral

AI，通过开源全球首个 MoE 大模型 Mistral 8x7B，迅速提升品牌影响力和市场地位，试图抢占欧洲市场。昆仑万维从技术研发、产品迭代、海外布局等多个方面布局 MoE 大模型，持续巩固国内先发优势，其“天工 2.0”已面向 C 端用户免费开放。

关键词 2：具身智能

以“大模型+机器人”为特征的具身智能快速发展，推动人形机器人闭环自智。具身智能(Embodied Intelligence)是一种高级的机器智能形式，它使机器人能够像人类一样感知和理解环境，并通过自主学习和适应性行为来完成任务。随着 AI 产业进程加速，大语言、视觉、多模态等大模型层出不穷，正加速与机器人技术相融合，与传感器、控制器等机器人关键部件共同组成人形机器人“大脑+小脑”，从感知、认知、决策、规划、控制、交互、学习、仿真等全方位提升机器人的智能化、通用化水平。由于机器人需通过传感器感知外部世界，因此无法直接照搬大模型中的 Transformer 架构，目前主要以 OpenAI 为代表的分层决策模型和以 Google RT-2 为代表的端到端模型两种算法路径实现。OpenAI、亚马逊、三星等多家巨头投资的 Figure 01 人形机器人基于大模型提供更高层级的视觉和语言智能，能流畅地与人交谈、理解人类需求并完成具体行动;特斯拉 Optimus 人形机器人内置 chatGPT 能在实际生产环境中纠正自己的错误。

关键词 3：分布式智算中心网络

大规模智算集群组网是 AI 大模型训练效率提升的关键。在大模型训练场景下，随着参数规模从亿级提升到万亿级别，算力需求呈现“爆发式”增长。据统计，2012~2022 年模型算力需求每年增长 4 倍，而 2023 年后模型的算力需求以每年 10 倍的速度增长。意味着训练超大 AI 模型需要数千/万卡 GPU 组成的集群高速互联。此外，机内 GPU 通信和机外集合通信将产生大量通信需求。例如，千亿级参数的大模型并行训练所产生的集合通信数据将达到数百 GB 量级。在极短时间内完成参数交换对 GPU 与 GPU 间、GPU 与网卡间、网卡与网卡间的超高带宽互联提出较高要求。网络拥塞和丢包也会严重影响 GPU 计算效率，据实验统计，0.1%的网络丢包率就会带来 50%的算力损失，因此提升通信性能可有效释放智能算力。AI 大模型训练/推理需要智算网络具备超大规模、超高带宽、超低时延、超高可靠等关键特征。如何设计高效的集群组网方案，提升 GPU 有效计算时间占比（GPU 计算时间/整体训练时间），对于 AI 集群训练效率的提升至关重要。

业界加快研究支持大模型跨数据中心分布式训练的长距组网技术，突破单智算集群算力规模、电力供给、机房空间的瓶颈。分布式跨智算中心网络是一项系统工程，涉及网络级负载均衡、异构集合通信优化技术、800G C+L 传输技术等关键技术。如谷歌基于自研 AI 芯片 TPUv4（AI 芯片）和光电路交换机，跨多个数据中心完成 Gemini Ultra、Gemini 1.5pro 等大模型训练。OpenAI 与微软计划建设十万级 GPU 算力集群，以满足 GPT-6 模型训练需求。但由于电力受限，

预计将 GPU 卡分布在几个或几十个地区，并利用开放 Ethernet 协议替换 IB 协议来实现跨区域 GPU 之间的互联。中国电信基于国产化算力完成跨百公里千亿参数模型在千卡规模下的分布式智算中心互联现网验证[23]，训练性能达到集中式单智算中心的 95%以上，证明了该技术方向的可行性。阿里提出“双上联+双平面+多轨”的 HPN7.0 网络架构，单 Pod 规模达到 15K GPU，不同 Pod 之间通过核心层互连，从而在单个集群中支持超过十万个 GPU 节点。

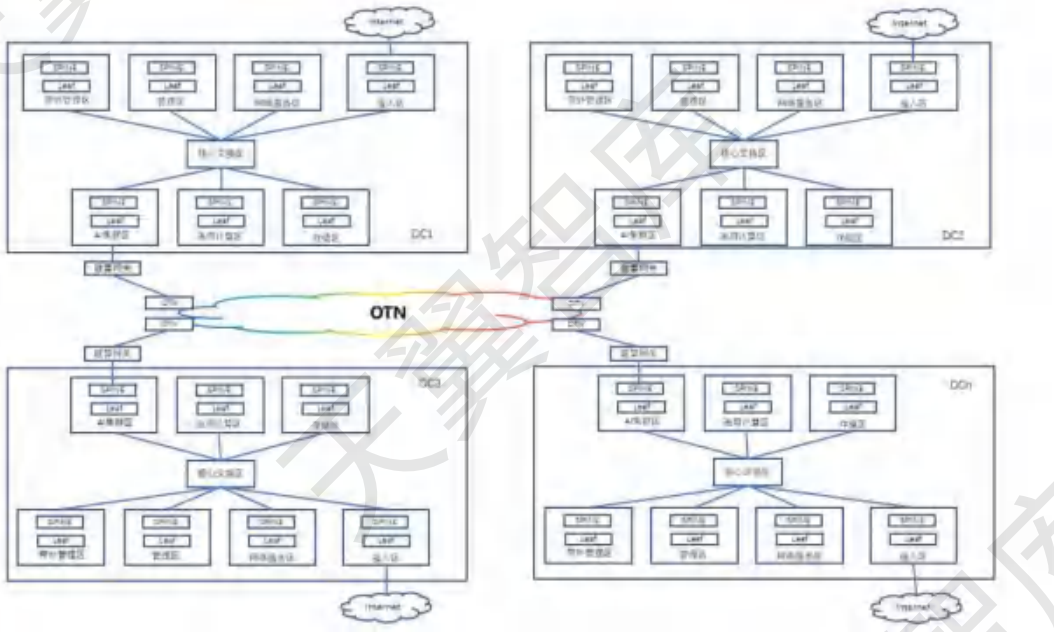

图 16 分布式智算中心无损网络总体架构

关键词 4：存算一体

存算一体是后摩尔时代的关键芯片技术，从根本上解决芯片发展面临“存储墙”、“功耗墙”、“工艺墙”的问题。存算一体是将存储单元和计算单元功能集成在同一芯片或系统中的技术，与先进制程、先进封装技术协同创新，具备高性能、高能效、高空间效率的特点。

根据 IBM 芯片级实验结果，诸如 ReRAM、MRAM、SRAM 等原器件，相比传统器件具备 10~300 倍的能效优势。

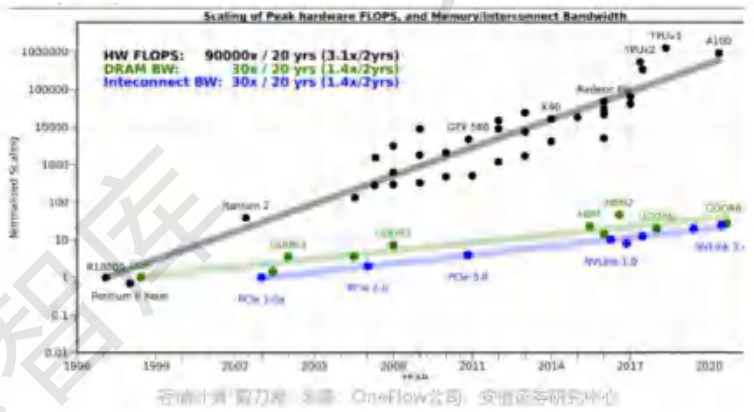

图 17 存储计算“剪刀差”

存算一体主要包括近存计算和存内计算两种发展路线。以 HBM 为主的近存计算技术基本成熟，已用于主流 GPU 和 CPU 芯片并规模商用，如英伟达 H200 使用了 HBM3e，显存带宽可达 4.8TB/秒，并提供 141GB 的内存[24]。以 NOR flash、SRAM 等传统存储器为主的存内计算技术相对成熟，已实现部分商用。以 ReRAM、MRAM 等新型存储器为主的存内计算技术正处于快速发展阶段，实验成果明显，如国外三星已完成 8nm eMRAM 的技术开发，国内清华大学基于 ReRAM 的存内计算芯片可将 AI 计算能效提升约 30 倍[25]。

大模型等 AI 应用为存算一体产业发展带来新机遇。HBM 凭借高带宽、低能耗已成英伟达、AMD、华为的等企业主流 GPU 的标配。随着大模型发展，HBM 呈现供不应求发展态势。海力士、三星等 HBM 主要供应商均表示，2024 年订单已售罄。存内计算技术，虽然目前仍处于早期发展阶段，但发展潜力巨大。后摩智能、D-matrix 等国内外初创企业正在积极布局存内计算芯片在大模型、自动驾驶等大算力场景的应用。据 QYResearch《全球存算一体技术市场报告 2023-2029》

显示，预计 2029 年全球存算一体技术市场规模将达到 306.3 亿美元，CAGR 为 154.7%[26]。

关键词 5：空心光纤

空芯光纤（HCF，Hollow-core fiber）突破实芯光纤的时延、距离和容量极限，被誉为下一代光通信的颠覆性技术。随着人工智能、大数据、云计算和物联网等新兴技术的发展，传统以玻璃等固体材料为光纤纤芯的实芯光纤，面临容量瓶颈和性能极限。空芯光纤和传统光纤架构一致，由纤芯、包层和涂覆层三部分组成，不同之处在于空芯光纤纤芯是空气。与传统光纤相比，空芯光纤具备低时延、长距离、大容量、低损耗、超宽工作频段等优势，如空芯光纤可实现传输时延降低 30%+，非线性效应低 3-4 个数量级，系统容量和传输距离提升 2 倍+，提供超 1000nm 的超宽频段等[27]。

表 1 空芯光纤应用场景

应用领域	具体应用	描述
长距离通信	跨国通信、海底光缆、卫星通信	空芯光纤具有高效、高速、大容量的特点，是未来长距离通信的首选方案
数据中心和云计算	数据传输	空芯光纤的高带宽和低损耗特性使其成为数据中心和云计算领域的理想选择
医疗领域	医疗设备制造	空芯光纤可用于制造医疗设备，提供更清晰、更准确的图像
工业领域	传感器和监测系统构建	空芯光纤可用于构建传感器和监测系统，实现对工业设备的实时监测和远程控制
其它领域	保密通信建设	空芯光纤可用于建设保密通信，提高通信的安全性和可靠性

诺基亚、西班牙 lyntia、Digital Realty 等国外运营商/IDC 厂

商，以及国内运营商、长飞等设备商正积极开展空芯光纤现网测试。2024 年 2 月，诺基亚贝尔实验室在巴黎进行了空芯光纤传输试验测试，展示了 800 Gb/s 和 1.2 Tb/s 的数据传输速率[28]。同月，lyntia 公司联合 Nokia、Digital Realty 等公司开展了空芯光纤实际线路环境中的现场传输试验，在 1.386 公里的链路上，实现往返延迟降低 30%。5 月，中国联通宣布携手北理工、上海诺基亚贝尔及长飞突破空芯光纤单波传输速率记录；6 月，中国移动宣布联合产业链在广东深圳-东莞开通首个 800G 空芯光纤传输技术试验网；同月，中国电信联合长飞、中兴通讯和华信设计院发布了全球首个单波 1.2Tbit/s、单向超 100Tbit/s 空芯光缆传输系统现网示范工程[29]。作为一项颠覆性技术，空芯光纤在批量生产、产品稳定性、市场成本等方面仍面临挑战，需产业界和学术界共同研究、攻克。

关键词 6：算电联合调度

算力和能源协同调度，是指应对算力节点负荷峰谷和可再生能源发电高波动特点，通过集成先进的监控系统、自动化工具和人工智能算法，统筹调度计算任务与能源供给，推进算力和能源在时间、空间维度的匹配调度，实现算力和电力深度融合发展。

图 18 支撑算网绿色低碳运行的电力网

算力和能源协同调度，从应用场景和发展方向，可划分三大类技术创新：**一是**以算力中心的源网荷储一体化为依托，构建多种能源资源和能源管理技术相结合的算力中心综合能源系统，统筹调度算力和电力资源，提升同一算力节点内源-网-荷-储的高效协同和能源高效利用，推动能源运营从被动用电向主动调度转型。**二是**以算力网络的调度平台为依托，结合电力系统能源结构和碳强度等预测数据，以及不同地区/时段分时价格信息，构建跨区域/算力中心的计算任务时间空间“负载转移”的动态调度平台，与各地区低碳电力供应进行匹配，促进算力可再生能源消纳和低成本低碳运营。**三是**以算力中心产生的大量低温余热利用为出发，利用热泵、吸收式制冷、蓄冷蓄热等技术，打造源网荷互动算力综合能源的数智化系统，实现算力与能源、热力的耦合互动，统一规划和调度，通过算力中心需求响应等方式有效减少用电量和热能的浪费[30]。

关键技术成熟度评估

参考 Gartner 技术成熟度曲线，综合分析新兴技术的标准化程度、科研热度、应用潜力、品牌效应/公众预期、市场竞争强度、创新能力，评估该技术在市场中的发展状态，技术成熟度共分为 5 个阶段：技术萌芽期、期望膨胀期、泡沫破裂低谷期、稳步爬升期和生产成熟期。

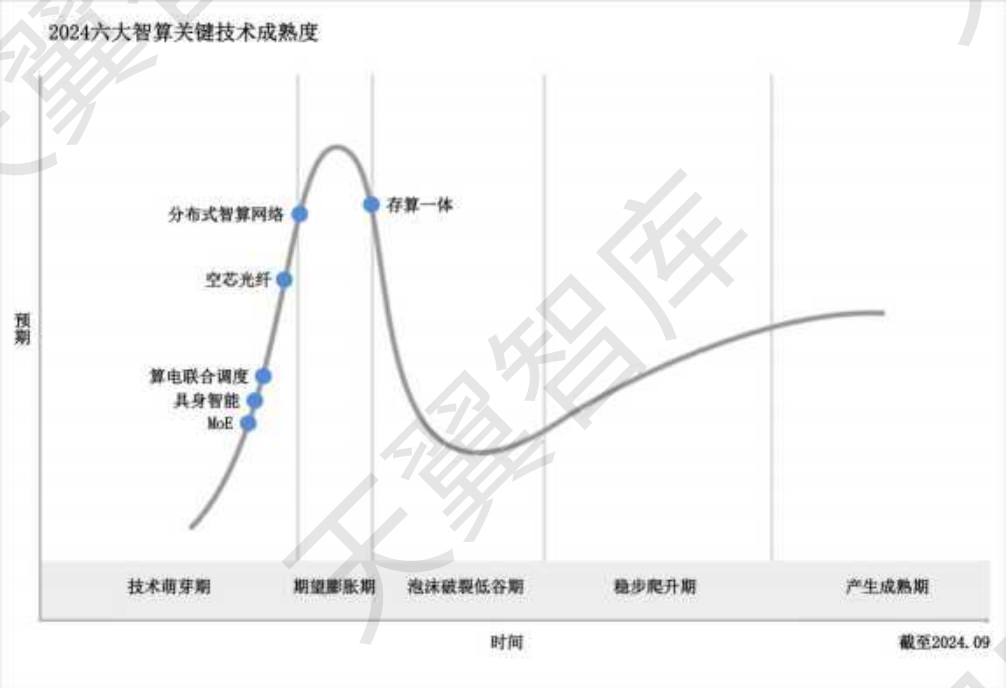

图 19 智算关键技术成熟度（TRL）

MoE、空芯光纤、具身智能和算电联合调度均处于技术萌芽期。MoE 模型技术仍处于前沿研究领域，尚未形成广泛的标准体系，仅有少量的研究和应用。空芯光纤概念提出时间较早，因缺乏应用场景而一直停留在实验室阶段，随着大模型的爆发式发展，空芯光纤在数据中心领域引起越来越多的关注，尽管这项技术还处于相对早期的商业化阶段，但已有一些初创企业通过融资支持其研发和测试。具身智能是人工智能发展的重要分支，该领域融合了人工智能、机器人、以及

感知和动作等多种复杂技术，仍处于前沿研究阶段，尚未形成规模应用和行业标准体系。算电联合调度处在方案阶段，仅有少量研究和应用，但是公众的预期较高。

分布式智算网络和存算一体处于期望膨胀期。分布式智算网路基于拉远 RDMA 架构，有较高的标准化程度和市场竞争强度，应用潜力较大，但是受成本等因素影响，目前尚未规模推广和应用。存算一体的近存计算技术路线已成熟，并被当前主流芯片厂商广泛采纳；存内计算技术路线目前已有代表供应商和商业案例，但仍处于发展早期。整体看，随着 AI 产业快速发展，推动智算规模的快速扩张和 GPU 等产品创新，公众对存算一体的商业化进程和市场前景有较高预期。

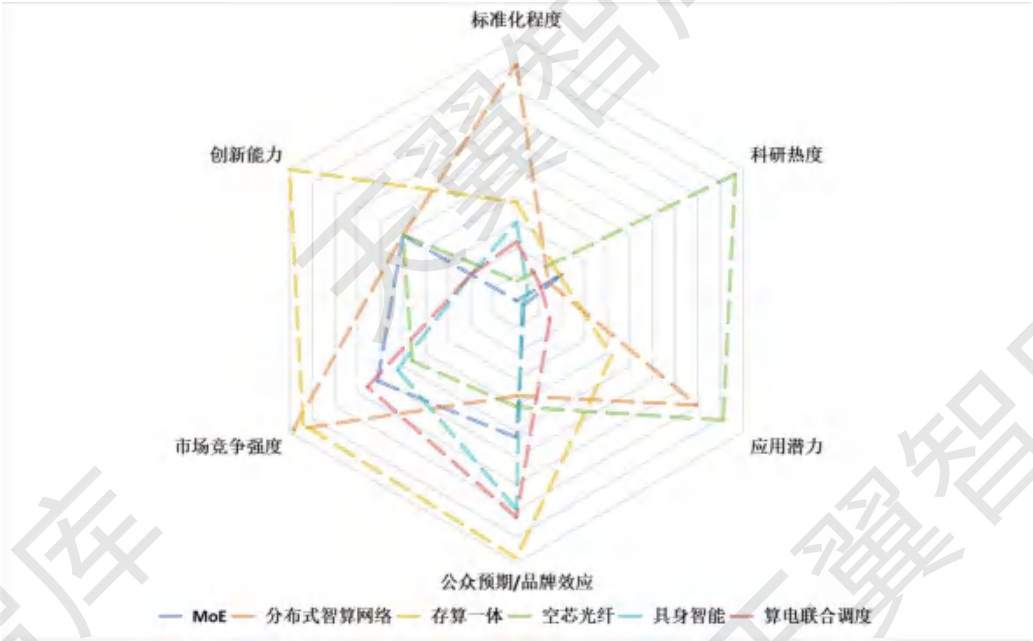

图 20 分维度关键技术成熟雷达图

五、智算发展潜力评估

1、评估方法

基于全国及各省智算业务相关政策、智算发展特点、行业专家意见，并结合国内外科研机构对智能算力的评估指标研究，借助统计学、指标筛选方法等构建智算发展潜力的评估指标。我们将智算发展潜力评估简称为 ICDP-EM（Intelligent computing development potential evaluation model）。ICDP-EM 如图所示，包括外部环境、基础设施、服务应用 3 个一级指标，以及相应的 8 个二级指标。

图 21 中国智算产发展潜力评估模型(ICDP-EM)

（一）模型分析

我们从外部环境、基础设施、服务应用三个方面对评估模型进行分析。

1）外部环境

AI 产业、智算中心、双碳等相关智算政策，将影响智算中心选址的具体位置。城市的商业电价、太阳能 / 风能 / 水等绿色发电能力决定了智算中心建设的总体成本，对智算中心的发展区域选择有较大

影响。员工薪资、GDP 等是经济发展水平高低的体现，对智算建设能力有一定影响。

2）基础设施

网络高带宽、低延迟是提升智能算力性能的重要因素，如光宽用户数、每万人 5G 基站数、IPV6 渗透率等网络基础能力作为智算中心算力、数据互通的基础，将影响智算对大模型等 AI 业务的训练推理速度、处理能力和结果的准确性。IDC 机架规模、总算力规模影响智算中心的建设和服务能力。

3）服务应用

大模型数量、AI 企业数量、AI 发明专利数等是衡量每个区域 AI 研发能力的关键，企业上云率、互联网网站数等体现了数字化能力，将影响智算服务未来的发展潜力。数字城市百强渗透率、人工智能产业园区数促进产业实践，影响智算服务应用能力。

（二）评估方案[2]

依据 ICDP-EM 模型分析，设计评估体系的评估方案，流程如下：

1）**指标构建：**通过 ICDP-EM 模型分析，构建中国智算发展潜力评估指标体系包括一、二、三级指标，详情见附录中表 3。

2）**指标赋值：**基于省人民政府、工信部、国家统计局等官网统计三级指标对应的最新数据，为三级指标赋值提供权威、客观的依据。

3）**权重确定：**基于 AHP 和熵权法主客观结合为各指标的权重设

[2] 详细的评估流程，见附录

计方案，其中一二级指标采用 AHP 方法确定权重，三级指标基于各省统计的指标赋值采用熵权法确定权重。

4）**评估指数结果：**最终根据指标的得分和权重得到各省相应的评估结果，包括综合评估指数、外部环境评估指数、基础设施评估指数、服务应用评估指数。

2、评估结果

基于评估方法确定的指标、权重和评估指数，本报告从综合评估指数、发展环境评估指数、基础设施评估指数、智算服务评估指数四个方面给出了我国 31[3]省智算发展潜力排序的建议。

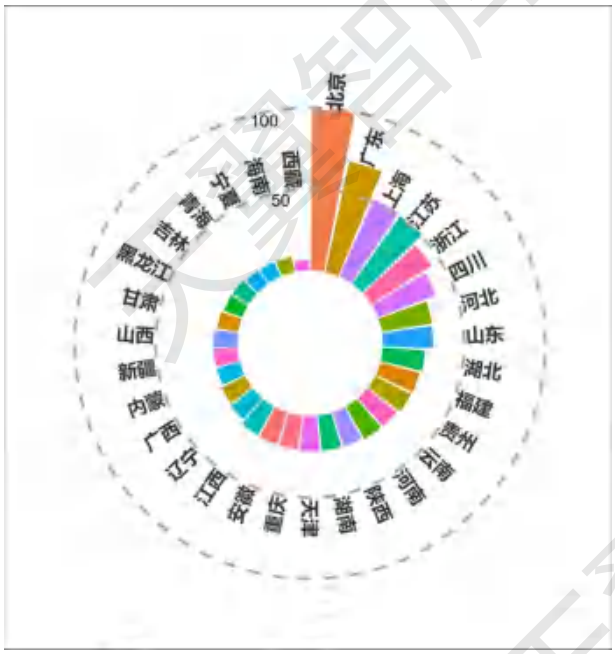

图 22 中国 31 省智算发展综合评估指数

（一）智算发展潜力综合评估指数

京津冀、长三角地区智算发展的综合评估指数均在中上游，是具有较高智算发展潜力的城市。

[3] 因数据获取难度等限制，本报告只统计中国 31 省数据，不包括中国香港、中国台湾和中国澳门

由图 13、14 所示，北京、广东、上海、江苏属于智算发展的第一梯队，综合指数在 50 以上。浙江、四川、河北、山东、湖北属于智算发展第二梯队，综合指数在 25 以上。如图 14 所示，以北京为代表的京津冀地区和以上海为代表的长三角地区人均 GDP 较高，拉动了智算整体的产业发展，在智算的发展建设上有更大的优势，助力大模型等 AI 业务快速发展。

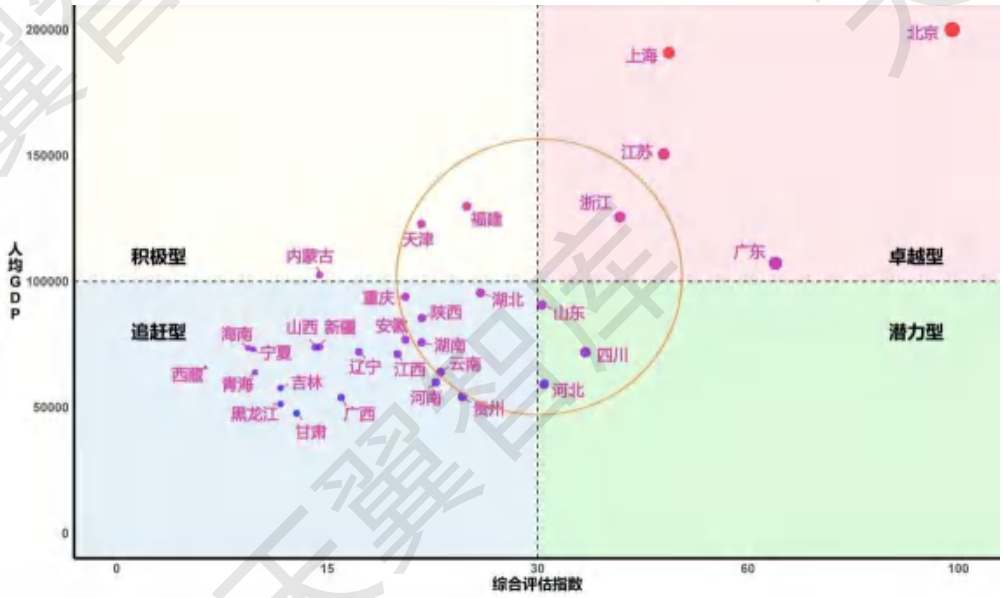

图 23 人均 GDP 与综合指数的象限图

（二）外部环境评估指数

中西部地区因绿电、建设成本低等特点，在智算发展的外部环境方面优势凸显。

如图 15 所示，四川、云南、湖北地区因水电等绿色能源供应量充足，新疆因工业电价低，均跻身第一梯队，适合发展绿色智算相关业务。北京、上海、江苏、广东因 GDP、高薪等因素在智算发展的外部环境方面也具有一定优势。

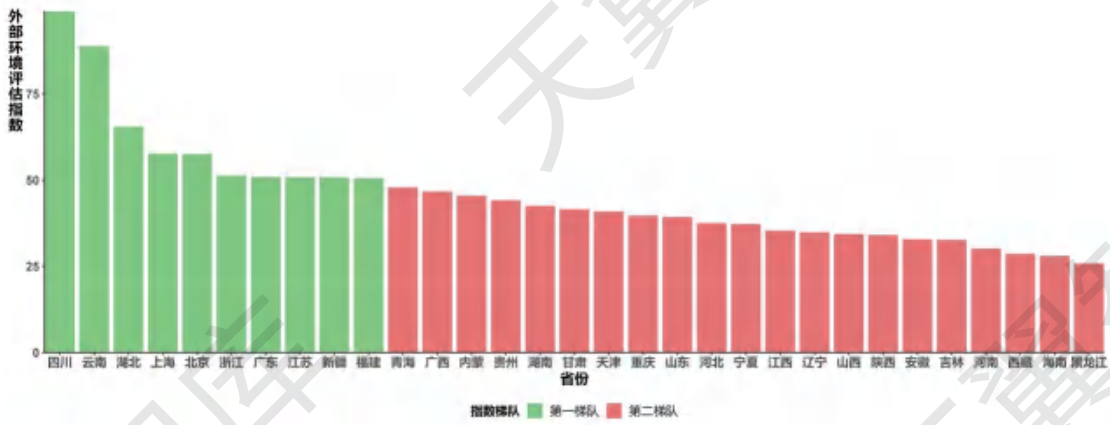

图 24 中国 31 省智算发展外部环境评估指数

（三）基础设施评估指数

全国智算基础设施布局不均，北京、上海、广东为代表的京津冀、长三角等地区在基础设施建设上具有城市集群效应，远高于中西部地区。

如图 16 所示，上海、江苏、浙江长三角地区均处于第一梯队，京津冀基础设施能力处于中上游水平，山东跻身第一梯队。西部地区在基础设施建设上还有很大发展空间，甘肃、贵州、四川等华西作为“东数西算”重要节点，智算中心规模发展迅速，挤入第二梯队。宁夏作为八大枢纽节之一，在光纤、5G 基站、IDC 机架建设等方面可重点发力。

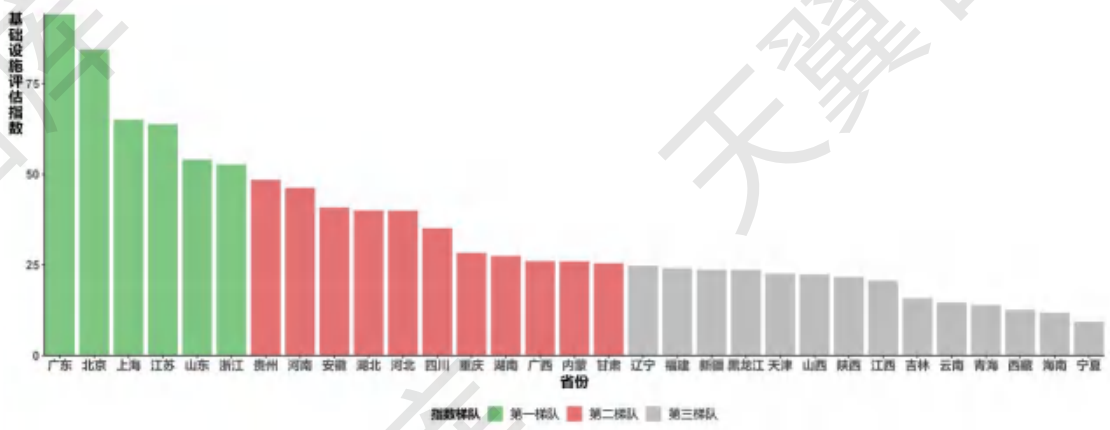

图 25 中国 31 省智算发展基础设施评估指数

（四）服务应用评估指数

智算服务应用能力主要聚集在经济较发达的一、二线城市。

如图 17 所示，北京、广东处于第一梯队，尤其北京在智算服务应用方面远高于其他省份。服务应用能力受基础设施能力的影响较大，服务应用评估指数的第一梯队（北京、广东）和第二梯队（上海、江苏、浙江），其均处于基础设施评估指数的第三梯队。

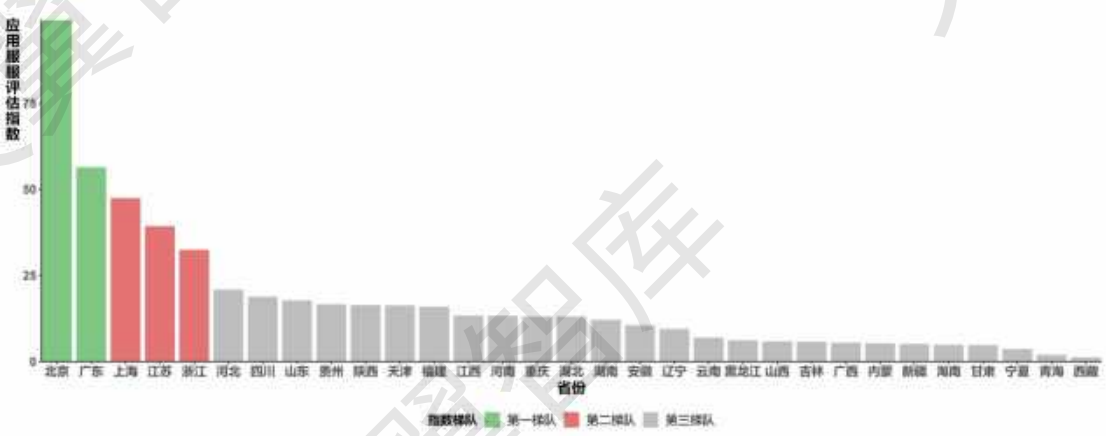

图 26 中国 31 省智算发展服务应用评估指数

基于以上评估指数排序，对综合评估指数 top10 的城市进行外部环境、基础设施、服务应用的能力分析。如图 18 所示，北京在综合能力和服务应用能力方面遥遥领先，广东、上海、江苏、浙江在基础设施能力方面占有优势，四川因出色的绿电供应（水电）使其在外部环境能力方面名列前茅。山东、湖北、河北、福建等在各方面处于中等水平，整体能力较稳定。

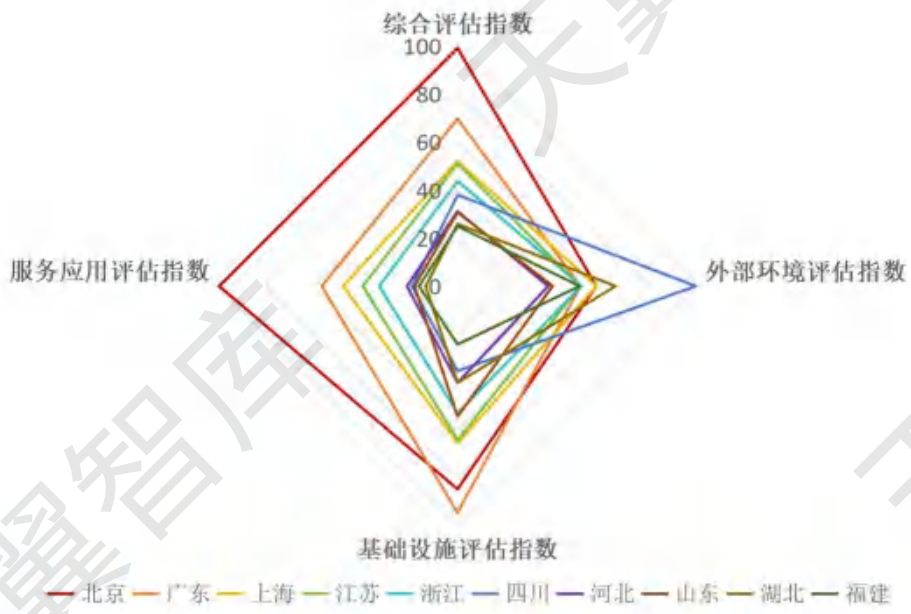

图 27 综合评估指数 Top10 省份细分评估指数对比

六、典型案例

1、中国电信上海万卡集群

2024 年政府工作报告明确要求开展“人工智能+”行动，打造具有国际竞争力的数字产业集群。随着大模型技术的快速迭代，智算算力供给面临着计算力指数级增长、更高性能存储以及高速无损网络需求等多项新挑战。与此同时，美国实施了更加严格的芯片限售政策，高性能 GPU 服务器的供应面临更为严峻的困境。对此，中国电信将上海、北京作为智算布局的核心枢纽节点，率先部署了国产万卡规模算力集群，并打造云骁、慧聚和息壤一体化的智算算力平台，以普惠的算力体系提供一站式大模型服务，助力 AI 快速创新，赋能千行百业。

2024 年 3 月 22 日，中国电信宣布天翼云上海临港国产万卡算力池正式启用，同时入驻首批用户。其中既有通用语言大模型公司百川

智能、稀宇科技、思必驰科技、天壤智能，也有深耕金融领域的行业大模型金声玉亮、国内领先的企业级 AIAgent 平台公司澜码科技、AI 创新生物制药公司赛陇生物，以及承担上海市人工智能研发与转化培育建设重任的上海人工智能研究院，基本覆盖基础层、技术层、应用层等人工智能完整产业链。这是国内首个正式投入运营的国产单池万卡液冷算力集群，通过全自研的智算平台，上海临港智算中心可面向多租户提供快速响应、灵活扩展、通智一体、安全可靠的智算云服务。

中国电信坚持超前布局，建设双万卡资源池，打造业界领先的智算 I+P 一体化的算力平台，加快推进天翼云向智能云升级，培育新质生产力。上海临港国产单池万卡液冷算力集群创新性地采用网络中置、算力分层的“魔方”型，实现了单一集群内万卡高速互联，满足万亿级参数大模型训练所需的多机多卡并行、高吞吐无损通信等需求。同时，为实现绿色低碳目标，全面采用融合液冷服务和 IDC 基础设施的新一代智算液冷 DC 舱，实现了数据中心的能效和智算集群的算效双提升，为“人工智能+”提供智能、弹性的绿色算力。

2、中国电信京津冀智算中心跨智算中心无损网络解决方案

伴随新一轮科技革命和产业变革，经济发展加速从工业经济向数字经济转变，也正在深刻影响全球经济竞争格局。全球数字经济标杆城市建设对提升国家数字经济核心竞争力至关重要。根据《全球数字经济标杆城市发展评价报告（2024）》数据显示，北京数字经济发展水平位列全球第二，其中人工智能企业约 2900 家，全国占比 28%，

位列第一，智算需求旺盛，是全国的智算高地。为满足未来北京市内及京津冀用算需求，以及解决单节点智算中心资源受限、不同智算中心资源使用不均衡等问题，中国电信率先在北京开展了分布式智算中心无损网络试验，验证跨数据中心合池训练的可行性，以提升区域内智算整体的供给效率。

中国电信提出“以网强算”的技术路线，通过将 IP 技术与光传输技术的协同创新，将相距百公里的多个智算中心连成一个更大规模的智算集群，补齐单点算力规模不足的差距。基于北京全光运力网规划，中国电信先后开展了现网机房的 64 卡以及 1024 卡组网验证。一阶段在京津冀智算机房进行 80km/120km 绕行拉远验证，模拟了两个数据中心组网。二阶段在武清、瀛海、永丰三机房开展百公里分布式大模型训练，验证当前分布式智算中心无损网络解决方案在真实业务场景下的效果，并探索分布式智算集群对大模型训练性能影响的关键因素。在前期百卡、百公里拉远验证基础上，三阶段在京津冀智算机房开展了千亿参数、千卡规模 120km 两点拉远验证，探索长距链路带宽收敛情况下模型训练的性能，目标是推动无损智算互联网络的技术进一步突破。系列试验均验证了在不同拓扑中分布式智算中心无损网络方案的有效性和稳定性。

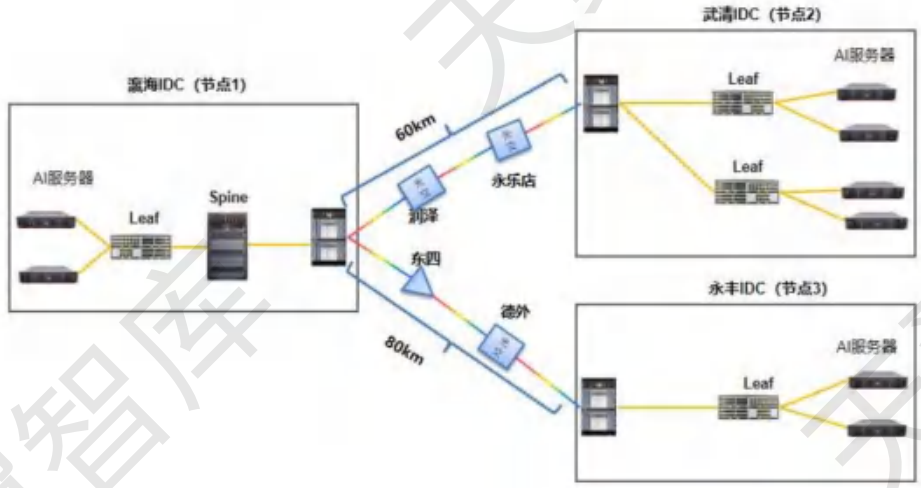

图 28 武清、瀛海、永丰三地 IDC 机房拉远验证组网

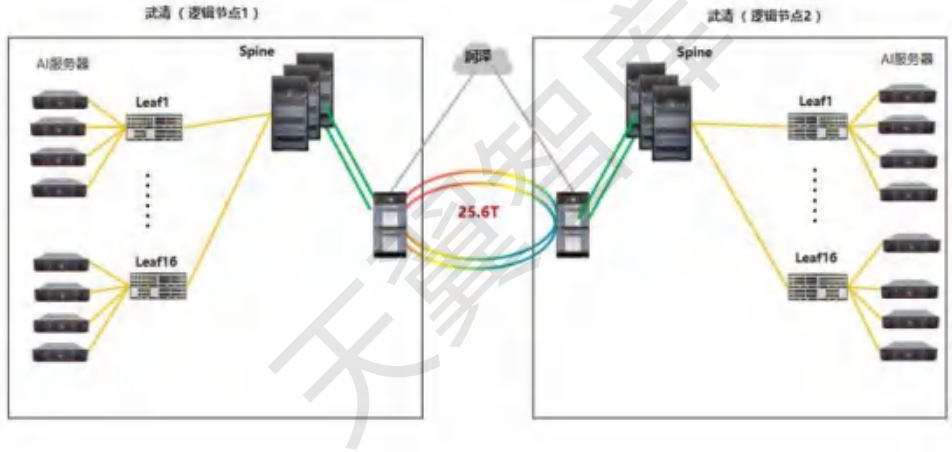

图 29 京津冀智算机房千卡 120km 绕行拉远验证组网

实验结果表明，训练效率方面，在不同组网拓扑下不同模型跨机房训练均可达同机房训练性能的 95%以上，证明分布式智算中心无损网络的可行性；网络稳定性方面，分布式智算中心无损网络可支持大模型一轮 5000 次迭代训练任务，均完成超 12 小时、约 80w 条样本数据的稳定性测试，具备支持大模型长期稳定训练的能力。分布式智算中心无损网络测试验证及相关创新研究将助力多方小规模智算中心并联成虚拟的大型智算中心节点，实现区域内智算中心协同计算

模式，解决临时性的大规模算力需求，推动端网算协同创新，解决供给与需求区域发展不平衡问题，促进京津冀战略协同，快速推进智算中心建设，夯实新一代算力底座，为区域算力互联网的建设打下坚实基础。

面向未来，中国电信将坚持“以网强算”的技术路线，打造面向智算业务的新型基础设施，以高性能智算网络作为提升集群算力性能的关键抓手，突破智能算力供给瓶颈，在赋能智算基础设施方面发挥更加重要的作用，为经济社会发展注入新的动力。

3、中国电信湖北中部绿色智算中心

中国电信中部智算中心位于湖北武汉东湖新技术开发区光谷八路中部智算中心（武昌）园区，园区占地 85 亩，是工业互联网标识解析国家顶级节点部署地、国家级互联网骨干直联点。2024 年 1 月中部智算中心正式对外发布，服务中部、辐射全国，为政府、企业、高校等提供公共算力、应用创新孵化、产业聚合发展、科研创新和人才培养等平台服务，重点满足湖北及周边省份智算业务发展需求以及“大模型”对信息基础设施需求。

图 30 中部智算中心（武昌）园区

园区规划方面，按照国家数据中心最高 A 级标准建设 3 栋数据中心、2 栋动力中心。其中，A1、A2 栋为通算中心，A3 栋为智算中心，可提供机柜数超 7000 架，远期可提供智算算力超 5000P。

网络系统方面，园区采用双路由直连国家 163 骨干网，出口宽带到达 9600G，配备 2 套 IDC 网络专用核心路由器。凭借强大的算力连接优势，园区在带宽、时延上可实现市内小于 1 毫秒，省内小于 3 毫秒，网络可靠性大于 99.99%。

弹性建设方面，采用弹性方舱方案适配弹性扩容需求，实现灵活部署，平滑扩容，实现弹性供电（小母线替代列头柜满足功率密度按需部署，业务灵活分布）、弹性制冷（弹性方舱融合多种空调形态，预留管路接口，兼容 8-50kw 冷量需求；匹配不同行业客户需求，制

定冷冻水、冷却水等多种制冷方式组合，灵活建设）和优化气流组织（CT 机房 IT 化，统一进出风方向，优化布局）等“两弹一优”特性。

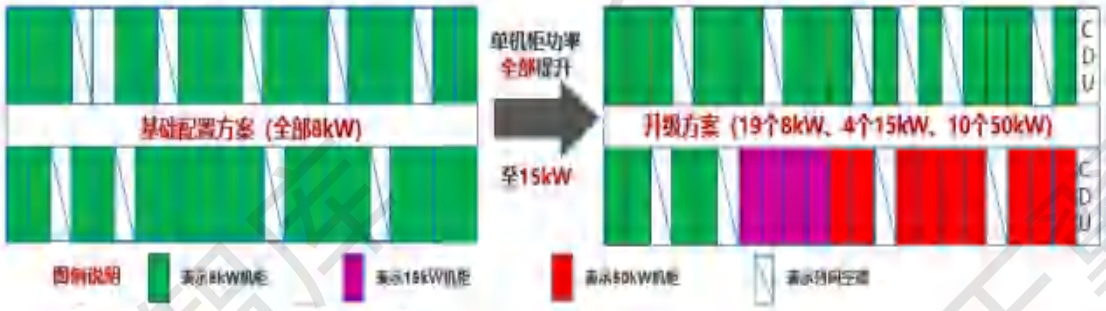

图 31 弹性扩容示意

绿色低碳方面，采用冷板式液冷、封闭热通道、间接蒸发冷却塔、近端制冷、混合补偿等节能技术全方位多维度实现绿色低碳节能，致力于打造中部绿色低碳标杆智算中心。其中采用的板式液冷技术无压缩机制冷，仅用冷却塔散热方式，极大降低了空调系统能耗，相比传统水冷空调系统，在节约冷冻站面积同时提高了供回水温度，液冷二次侧供回水温度提高至 40/48℃，极大的减少了空调系统能耗，年平均 PUE 可达 1.17，相比传统数据中心节能约 10%，年节约 1286 吨标准煤，相当于减少 3406 吨 CO2 排放。

图 32 冷板原理（左）和冷板式液冷系统图（右）

中部智算中心已成功为华中科技大学、武汉人工智能研究院、武昌区政府、抖音集团等实体及荆州城市生命线、十堰有云等项目提供大模型训练等智算算力或智算基础设施服务。

4、海兰云海底数据中心

海底数据中心作为一种新型绿色算力设施，在智算领域具备独特技术优势。**节能降碳：**海底数据中心利用海洋作为自然冷源，可以显著降低能耗。相比传统数据中心，海底数据中心省电 30%以上，而且不消耗淡水。**安全稳定：**海底数据中心算力舱中无氧无尘的环境提高了服务器等 IT 设备的可靠性。这一技术优势显著延长算力设备的使用寿命。对于价格昂贵的智算设备而言，可有效提高其经济效益。**算电协同：**海底数据中心具有与海上风电等新能源融合开发的天然优势，新能源使用率可达 80%以上，形成绿电消纳、产业协同、共建共维、降本增效的新发展模式。

海底数据中心产业应用稳步推进。全球首个商用海底数据中心于 2022 年 12 月在海南启用，目前已与中国电信、中国移动、崖州湾科技城、腾讯、拓尔思、广联达、陵水文化产业云等企业开展业务合作。2024 年 4 月，海兰云与合作伙伴共同发布全球首个海底智算中心平台。该项目将开展包括 AIGC 大模型、离岸教育科研、跨境影视制作、深海探测等特色业务，作为新型绿色算力设施加入全国算力布局。海底智算中心充分利用高能效的特性，可部署高功率密度的人工智能服务器。以 1MW 外电配给为参考，海底智算中心凭借高功率密度设计，

单舱可提供 1400PFlops 算力，算力效率提升 40%。此外，长三角海底数据中心+海上风电融合开发项目稳步推进，目前已完成相关审批，年内即将动工。该项目将打造绿电直供、算电协同的新发展模式。

海洋算力兴起推动陆数海算新潮流。海底数据中心凭借独特的优势，有望成为未来智能计算发展的新方向。通过结合液冷技术和与绿色能源，海底数据中心将为各行各业提供更加高效、可靠和环保的智算解决方案。

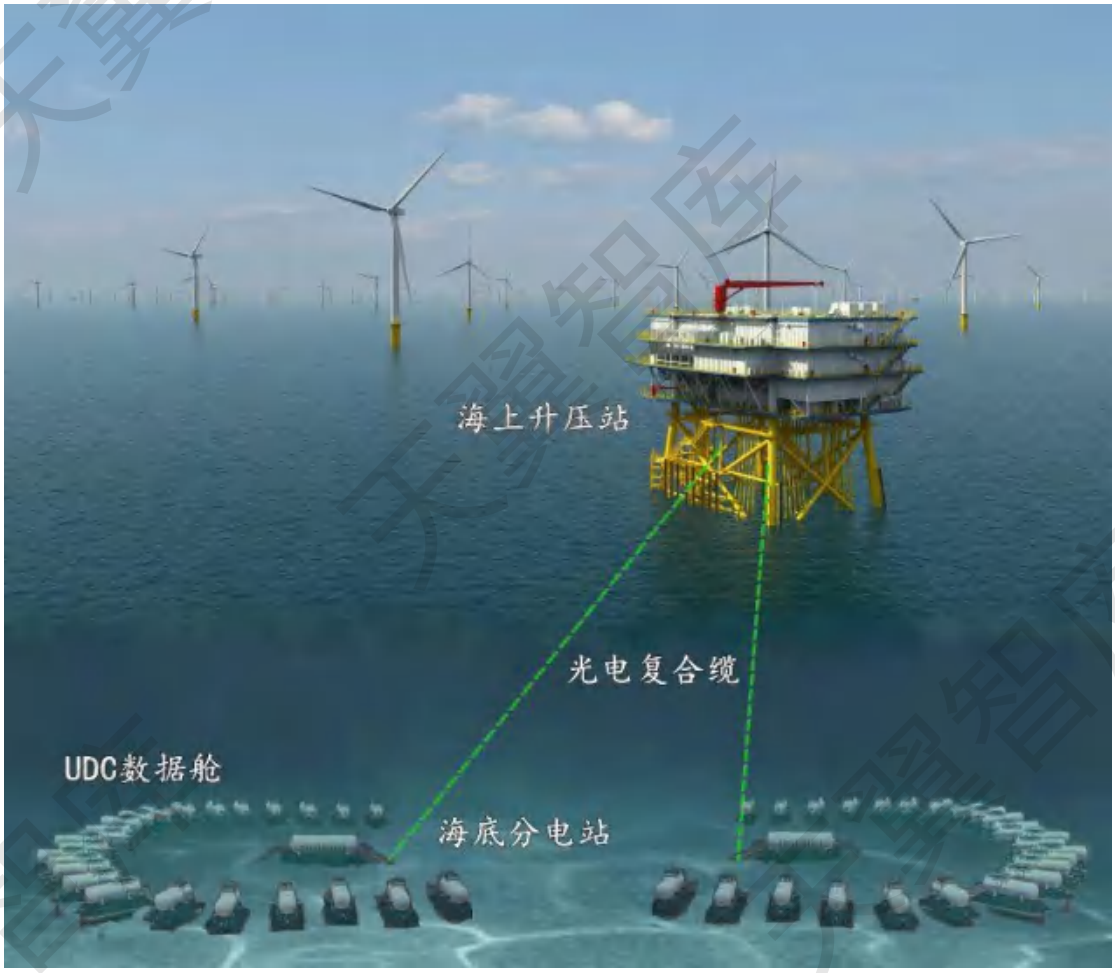

图 33 海兰云海底数据中心

七、总结与展望

当前人工智能已进入大模型时代，算力、数据、模型快速增长，相比之前的人工智能，大模型具有更好的通用性、更广的应用范围，具备赋能各行各业的潜力，颠覆传统的生产流程、创新模式，引领产业加快向智能化升级，是形成新质生产力的关键驱动力。

智算作为支撑大模型发展的基础底座，多重价值凸显，赋能全社会生产效率提升，带来产业数字化、智能化的经济价值，肩负着人工智能的核心产业价值。总体来看，智算行业呈现需求激增、价格高企、更新迭代快的特征，市场供给爆发式增长，类型多元化、异构化，万卡级及以上智算集群加速部署落地，为我国以生成式大模型为代表的人工智能奠定了坚实的基础。但也看到，我国智算布局分散、生态有待完善、规模化需求暂未显现、核心技术有待突破等问题依然突出。

面对新一轮科技革命和产业变革的发展机遇，智算规模化增长趋势及高质量供给的需求是确定的，未来需要产业界与学术界的紧密合作，共同探索智算的跨行业、跨领域的共享模式，加强训推一体的分时复用等关键技术的联合攻关，协同模型架构、通信拥塞控制等优化，探索分布式智算集群模式创新，为人工智能向全社会渗透提供支撑。

八、附录-智算评估实施方案

本白皮书对中国智算产业潜力发展评估的具体实施方案如下。

1、评估指标模型构建

结合模型假设的影响因素，我们编制了一级指标、二级指标以及对应的指标说明、评估单位，便于后续指标评估。

附表 1. 中国智算发展潜力评估指标体系

一级指标	二级指标	三级指标	单位
外部环境	政策因素	AI产业政策数量	个
		智算中心政策数量	个
		双碳相关政策数量	个
	经济因素	人均GDP	元
		人力成本（月薪水平）	元
	能源配套	用电成本（工业）	元/千瓦时
		太阳能、风能等绿电供应量	亿千瓦时
基础设施	网络能力	光宽用户数	万户
		5G基站数	个/万人
		IPV4地址数占比	%
		IPV6渗透率	%
	算力能力	IDC机架规模	万架
		总算力规模	EFLOPs
		智算中心规模	EFLOPs
服务应用	AI研发能力	AI产业规模	亿元
		大模型数量	个
		AI企业数量	个
		AI发明专利数	个
		高校、科研机构数量	个
	数字化水平	企业上云率	%
		互联网网站备案数	个
		政府网站数量	个
	产业实践	数字城市百强渗透率	%
		人工智能产业园区个数	个

2、评估指标赋值

基于省人民政府、工信部、国家统计局等官网统计智算相关三级评估指标的最新数据，为 31 省的三级指标赋值提供权威、客观的依据。为 31 省的 24 个指标赋值，并对所有指标数值 x 进行归一化处理，得到每个指标的标准化数值 x'。

3、评估指标权重设计

关于评估指标权重的确定采用主客观结合的方式进行，保证评估结果的专业性和客观性。对于一、二级指标，涉及指标全面性的确定，需专家参与判定，采用 AHP 的评判矩阵来确定指标的权重。对于三级指标，在已经确定指标全面性的前提下，采用熵权法确定指标权重，确保结果的客观性。

（一）一、二级指标权重确定

基于 AHP 方法对一、二级指标进行权重设计，借助评判矩阵得出一、二级指标的权重，权重确定流程如下：

1）根据指标分类制定评断矩阵模板。

附表 2. 智算发展潜力评估指标评判矩阵模板

	指标 1	指标 2	…	指标 n
指标 1	a_{11}	a_{12}	…	a_{1n}
指标 2	a_{21}	a_{22}	…	a_{2n}
…	…	…	…	…
指标 n	a_{n1}	a_{n2}	…	a_{nn}

备注： n 是一或二级指标的个数。矩阵中的值为对应纵向指标比横向指标重

要程度，例如，$a_{ij}=\frac{l}{m}$ 是第 i 个指标与第 j 个指标比较对智算发展重要程度比值，其中$l,m\in(0,9)$。0 到 9 表示两个指标比较对智算发展的重要程度，数值越大重要程度越大。

2）业内智算专家按步骤 1 规则对需要评估的 n 个指标进行打分，分别给出相应的 n × n 阶评判矩阵，我们将这些评判矩阵记为 $A_1, A_2, A_3, \ldots, A_m$。

3）通过公式 $CR=\frac{\lambda-n}{(n-1)*RI}$，对评判矩阵进行一致性验证。

4）若评判矩阵通过一致验证，计算最大特征值 λ，对应的特征向量，即为指标对应的权重。

（二）三级指标权重确定

基于三级指标对应的 31 省数据，采用熵权法确定三级指标的权重，主要思路是**根据指标变异性的大小来确定客观权重**。流程如下：

1）根据三级指标的 31 省数据，构造矩阵B，模板如下。

	指标 1	指标 2	…	指标 24
省份 1	b_{11}	b_{12}	…	$b_{1,24}$
省份 2	b_{21}	b_{22}	…	$b_{2,24}$
…	…	…	…	…
省份 31	$b_{24,1}$	$b_{24,2}$	…	$b_{24,24}$

2）对矩阵B数据进行标准化处理，对于正向指标采用

$$b_{ij}^{'}=\frac{b_{ij}-\min\left(b_{ij}\right)_{j=1,2\ldots,24}}{\max\left(b_{ij}\right)_{j=1,2\ldots,24}-\min\left(b_{ij}\right)_{j=1,2\ldots,24}}$$

对于负向指标采用

$$b_{ij}{}^{'}=\frac{\max\left(b_{ij}\right)_{j=1,2\ldots,24}-b_{ij}}{\max\left(b_{ij}\right)_{j=1,2\ldots,24}-\min\left(b_{ij}\right)_{j=1,2\ldots,24}}$$

3）算每个指标 j 的熵值

根据矩阵B标准化后的数值计算信息熵：

$$H_j = -\frac{1}{\ln 31}\sum_{i=1}^{31}\frac{b_{ij}}{\sum_{i=1}^{31} b_{ij}}.\ln\sum_{i=1}^{31}\frac{b_{ij}}{\sum_{i=1}^{31} b_{ij}}$$

备注：**信息熵**是对一个信源所含信息的度量，即信息量的期望。

4）计算指标 j 对应的权重值

$$w_j = \frac{1 - H_j}{24 - \sum_{i=1}^{31} H_j}$$

4、各省评估得分

根据以上方法确定的一、二、三级指标权重和 31 省的 24 个指标的标准化分值，为各省进行综合评分，并分别根据对应的二、三级指标为各省的一级指标外部环境、基础设施、服务应用三个类别进行评分。

九、参考文献

[1] EPOCH AI. https://epochai.org/data/notable-ai-models. 2024
[2] IDC. 《2024AIGC 应用层十大趋势白皮书》. 2024
[3] Dealroom,Flow Partners.《The Magnificent Seven – The Venture Capital frontier & the new AI Wild West》. 2024
[4] 世界经济论坛. 《生成式人工智能与国际贸易分析》. 2024
[5] IDC. 《The Global Impact of Artificial Intelligence on the Economy and Jobs》. 2024
[6] 华泰证券. 《AI 发展对电力存在哪些影响与机遇》. 2024
[7] Epoch AI. 《Stanford The rising costs of training frontier AI models》. 2024
[8] IDC. 《中国智算服务市场(2023 下半年)跟踪》. 2024
[9] 赛迪. 《IT2023》. 2024
[10]信通院. 《全球数字经济白皮书（2024）》. 2024

[11]赛迪. 《2024 年中国人工智能行业典型大模型 100 强企业》. 2024
[12]腾讯研究院. 《行业大模型调研报告》. 2024
[13]前瞻产业研究院. 《2024 年前瞻中国 AI 大模型场景应用趋势蓝皮书》. 2024
[14]麦肯锡. 《麦肯锡 2023 秋季刊：捕捉生成式 AI 新机遇》. 2023
[15]斯坦福大学. 《2024 年人工智能指数报告》. 2024
[16]Epoch AI. 《Will we run out of data? Limits of LLM scaling based on human-generated data》. https://arxiv.org/html/2211.04325v2
[17]Wayne Xin Zhao. 《A Survey of Large Language Models Computer Science (2024). https://arxiv.org/abs/2303.18223
[18]MarketsandMarkets. 《Autonomous Al and Autonomous Agents Market》. 2024
[19]Weng,Lilian. LLM-powered Autonomous Agents". Lil' Log. https://lilianweng.github.io/posts/2023-06-23-agent/
[20]Shazeer N , Mirhoseini A , Maziarz K , et al. 《Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer》. 2017.DOI:10.48550/arXiv.1701.06538
[21]http://finance.sina.com.cn/cj/2024-09-26/doc-incqphkq5237198.shtml
[22]https://baijiahao.baidu.com/s?id=1716150130515573927&wfr=spider&for=pc]
[23]中国电信研究院. 《分布式智算中心无损网络技术白皮书》. 2024
[24]https://mp.weixin.qq.com/s/myStZJgI6rXaqK1Hw0deiA
[25]https://mp.weixin.qq.com/s/G4c1uUzHOQurmuXA7h6GMg
[26]QYResearch. 《全球存算一体技术市场报告 2023-2029》. 2024.
[27]https://mp.weixin.qq.com/s/-X8jRlATOlh_6YOGcPLUOQ
[28]http://www.iccsz.com/site/cn/News/2023/10/11/20231011020854417484.htm
[29]https://mp.weixin.qq.com/s/g8rII9ngYsLQaMLIvBiaoQ
[30]中国信通院. 《中国绿色算力发展研究报告》. 2024