CAICT 中国信通院

集智 专题报告

人工智能算力基础设施赋能研究报告

(2025 年)

中国信息通信研究院产业与规划研究所

2025年11月

版权声明

前　言

在人工智能落地破局与赋能千行百业进程中，以智算中心为代表的人工智能算力基础设施，被赋予更重要的定位和使命，成为支撑人工智能技术及产业发展的重要基石。然而，目前人工智能算力基础设施利用负载情况差异较大，尤其是地方政府或国资平台主导投建的智算设施，赋能价值有待进一步提升。因此，需要厘清智算中心赋能的需求场景、需求场景与所需关键能力的匹配，以及推进赋能落地的生态模式，助力人工智能算力基础设施真正实现赋能价值。**本报告聚焦智算中心赋能，围绕需求场景、关键能力、落地生态这三个关键环节，阐述最新发展趋势，致力于进一步释放智算中心的赋能效应，助力人工智能与实体经济深度融合**。

需求场景方面，大模型预训练、微调、推理场景对计算需求不一，当前各方主体已针对性开展各类场景的支撑。推进基础预训练大模型的训练，需要具备 E 级（EFlops）计算能力的高端万卡集群中心支撑。推进小模型（百亿级及以下参数）的训练、微调，或推进模型的推理，依托百 P 级（PFlops）计算能力的中小型智算中心即可支撑。

关键能力方面，训练场景与微调/推理场景在底层支撑能力、创新服务能力、运营保障能力要求有较大区别。在算力市场高阶技术服务需求暴涨的当下，智算中心亟需针对性夯实关键能力，支撑数据处理服务、算力调度服务及推理应用服务等。

落地生态方面，智算中心赋能需要分场景聚合 AI 能力主体，推进智算中心赋能所需核心要素主体的协同。智算中心需求场景和关键

能力需要有落地生态的加持才能落地，而落地生态需要推进算力、数据、算法、场景、产业等要素间协同协作，助力 AI 大模型赋能行业场景落地。

当然，报告仍有诸多不足，恳请各界批评指正。

目 录

一、人工智能算力基础设施演进态势 1
（一）技术创新：三位一体智算设施升级 1
（二）布局优化：全国智算设施协调发展 3
（三）产业升级：智算全产业链协同发展 4
二、人工智能算力基础设施赋能的重要走势 6
（一）需求场景逐步明晰，促使智算资源优配 7
（二）关键能力逐步聚焦，提升智算服务水准 7
（三）落地生态逐步聚和，加速智算价值释放 8
三、人工智能算力基础设施需求场景 8
（一）大模型预训练场景 9
（二）大模型微调场景 10
（三）大模型推理场景 11
四、人工智能算力基础设施关键能力 13
（一）基础支撑能力 14
（二）创新服务能力 17
（三）运营保障能力 19
五、人工智能算力基础设施落地生态 21
（一）智算与数据要素协同 22
（二）智算与模型算法协同 23
（三）智算与跨域智算协同 24
（四）智算与行业场景协同 25
（五）智算与区域产业协同 26
六、发展展望 28
（一）需求场景更加多元、复杂 28
（二）关键能力更加集约、软性 29
（三）落地生态更加聚合、协同 30

一、人工智能算力基础设施演进态势

人工智能算力基础设施，是基于人工智能专用算力芯片及加速芯片等组成异构计算架构，以智能计算设施为核心设施，以智能算力集群为核心载体，面向人工智能应用场景，提供所需算力服务、数据服务和算法服务的公共算力基础设施。大模型加速技术迭代与工程创新步伐，对人工智能算力基础设施技术创新、布局优化、产业升级也提出了更高阶的要求。

（一）技术创新：三位一体智算设施升级

当前，我国人工智能算力基础设施正处于系统性升级与架构演进的关键时期，正加速向集约高效、低碳共生、高速泛在的融合形态纵深发展，全面步入以**超大规模集群化、绿色低碳化与高速互联化**为核心特征的新发展阶段。在顶层设计与政策引导协同驱动下，我国智算中心规模持续扩大、技术体系日益自主、能效水平不断提升、互联能力显著增强，逐步构建起支撑数字经济高质量发展和现代化产业体系建设的算力底座。

超大规模集群化实现突破。超大规模集群指由万或超十万颗高性能 GPU/NPU/DPU/CPU 等异构算力卡、HBM/DRAM/SSD 等异构存储单元构成的异构计算/存储集群，通过高速互联网络实现近乎线性的算力扩展，以支撑千亿乃至万亿参数大模型的训练与复杂推理任务。**超节点**作为智算集群化发展的关键技术之一，正逐渐成为行业焦点。以华为昇腾 384 超节点为例，通过高速总线实现 384 卡高速总线互联，跨节点通信带宽提升 15 倍，并通过全局内存统一编址技术，将

分散在各节点的内存池虚拟为统一地址空间，支持跨节点直接内存访问，配置 8 节点存储超节点集群，具备超大带宽、超低时延、超强性能的三大优势。中兴通讯提出 AI 加速器高速互联开放架构，构建国产化 GPU 卡大规模高速互联的系列 Nebula 星云智算超节点，支持 Scale-Up 与 Scale-Out 双重扩展模式，高带宽域可扩展至 2048 卡，为 AI 训练及推理场景提供软硬协同、开放解耦、高效高稳的算力底座。

绿色低碳化发展成效显著。当前 AI 爆发式增长带来巨大的算力能耗挑战，人工智能算力基础设施的绿色低碳发展趋势正驱动商业模式创新。绿色低碳不仅是用户选购智算中心及算力服务的重要考量，也是企业服务方案的核心竞争力。目前业界广泛采用液冷、算电热碳一体化、模块化建设及智能化运维等先进技术，持续优化智算中心能效水平。中国移动长三角（苏州）云计算中心机房通过利用液冷技术，使散热能耗降低 50％～60％，数据中心 PUE 值降低至 1.25 以下。目前主流冷板式液冷仍需风冷辅助散热，未来将通过优化冷板设计、推广单相浸没式或全覆盖冷板液冷，逐步减少风冷依赖，提升单机柜功率密度并简化数据中心架构。此外，在智算中心平台侧，可通过算力编排管理系统、碳排放监测与统计平台等，适应不断变化的应用需求和能效要求。

高速互联化加速探索。智算中心的发展不仅取决于单点算力性能，更依赖于跨节点、跨地域的高效互联，分布式训练和跨中心协作成为当前探索的技术热点领域。我国正通过构建“物理网络+虚拟网络”双层协同技术体系，加快推进算力基础设施的高质量互联化发展，提

升算力资源的整体效能。目前 400G 技术体系基本完善，我国运营商逐步启动干线场景规模部署，对于更高速率的传输技术，业界加快 800G/1.6T 技术标准研制，OIF 和 ITU 等标准化组织正在开展相关项目研究工作，产业界也已经启动试点验证。此外，G.654.E 光纤、空芯光纤等新型传输媒介加快应用步伐，为构建高性能智算网络基础设施提供有力支撑。

（二）布局优化：全国智算设施协调发展

政策引导推动智算中心高质量发展。我国高度重视智算中心建设，自 2020 年发改委将智能计算中心纳入新基建范畴以来，国家相关部门相继出台了《新型数据中心发展三年行动计划（2021-2023 年）》、《算力基础设施高质量发展行动计划》等多份文件，提出统筹建设高性能智算中心，对于智算中心从鼓励建设转向深化布局，指引方向更加明确。2025 年 5 月，国家数据局印发《数字中国建设 2025 年行动方案》，提出逐步实现各地区算力需求与国家枢纽节点算力资源高效供需匹配。2025 年 8 月国务院印发《关于深入实施“人工智能+”行动的意见》，明确提出强化智能算力统筹，加快超大规模智算集群技术突破和工程落地，优化国家智算资源布局，完善全国一体化算力网，加强智能算力互联互通和供需匹配，创新智能算力基础设施运营模式，推动智能算力供给普惠易用、经济高效、绿色安全。

智算中心建设加快布局，总体规模进一步增长。近年来，我国持续加大对计算、存储和算力网络基础设施的投入，算力作为支撑人工智能发展的核心生产力，呈现出稳中有进的发展态势。据中国信通院

《2025 综合算力指数》数据显示，截至 2025 年 6 月底，我国在用算力中心机架总规模达 1085 万标准机架，智能算力规模达到 788EFlops[1]（FP16），为海量数据计算提供智能底座。**与此同时，区域智能算力向统筹化和集约化部署布局。一方面，**在全国一体化算力网建设、“东数西算”战略等大背景下，新建智能算力中心逐渐融入八大枢纽节点建设。截至 2025 年一季度，我国“东数西算”八大枢纽节点算力总规模达到 215.5EFlops[2]，智能算力规模占枢纽节点算力规模的 80.8%，枢纽节点间 20 毫秒时延圈已基本实现。**另一方面，**重点城市区域智能算力供给能力持续提升。北京市 2024 年新增算力达 8620PFlops，累计智能算力规模超过 33EFlops[3]；截止 2025 年 7 月，上海市智能算力规模已达到 100EFlops[4]；截至 2025 年 3 月，深圳已建和在建智能算力规模超过 62EFlops[5]。

（三）产业升级：智算全产业链协同发展

智能算力产业发展提质增速。智算中心作为集算力、存力、运力于一体的新型基础设施，为越来越多的行业数字化转型注入新动能，产业规模持续跃升。据国际数据公司（IDC）《中国人工智能计算力发展评估报告》显示，2024 年我国智能算力市场规模达 190 亿美元，同

[1] 数据来源：中国信通院

[2] 数据来源：国家数据局

[3] 数据来源：北京市科协

[4] 数据来源：上海市经信委

[5] 数据来源：深圳发布

比增长 86.9%。产业链各环节深度融合。**上游**核心硬件国产化突破加速，硬件产品性能实现跃升。据 IDC 数据显示，2024 年我国加速芯片市场规模超过 270 万张，GPU 卡占据 70%的市场份额，我国本土 AI 芯片出货量已超过 82 万张，市场渗透率从 2023 年的 15%提升至 30%。**中游**算力设施建设规模化推进，成熟大模型的运营有望为我国带来持续的智能算力需求。**下游**算力应用向金融、医疗、教育、交通、工业、传媒娱乐等诸多产业加速渗透，从通用场景迈向专业领域的特定场景。据中商产业研究院数据显示，从当前应用占比情况来看，互联网占比 53%，服务行业占比 18%，政府占比 9%，电信、工业制造、教育、金融等行业均占比 4%。

三大运营商的算力布局紧密围绕国家“东数西算”与“人工智能+”战略展开，已形成覆盖全国的立体化算力网络。中国移动在智算领域“N+X”智算能力布局不断完善，京津冀、长三角、粤港澳大湾区、成渝等区域首批 13 个智算中心节点投产，呼和浩特、哈尔滨两大万卡级超大规模智算中心高效运营，为 AI 应用的发展提供了强大的算力支撑。其中，黑龙江哈尔滨的智算中心节点，是全球运营商规模最大的单集群智算中心，智算卡算力规模达到 6.93 EFlops。**中国电信**适度超前开展智算建设，重点规划“2+3+7+N+M”的智算布局，建设“中心集群+边缘 DC”一体化的 AIDC，在内蒙古和贵州打造两个公共智算中心，在京津冀、长三角、粤港澳大湾区、成渝等地建设大型智算中心和超算集群，重点承载 AI 训练、高性能计算等需求。**中国联通**智能算力为“1+N+X”梯次布局，即建设 1 个超大规模的单体智算中心、

N 个智算训推一体枢纽，布局属地化的 X 个智算推理节点，最终构建以算力为核心的一体化算网融合生态体系，打造数字经济“第一算力引擎”。中国联通上海临港智算中心，凭借在技术创新、绿色低碳与产业赋能方面的突出表现，成功入选《中国信通院 2025 年智算中心典型案例》，成为长三角区域智算中心建设的重要标杆。

AI 大厂纷纷加速布局智算领域，聚焦大规模智算中心建设，注重技术创新与场景融合，助力 AI 技术落地应用。阿里云构建了完整的 AI 基础设施，来满足训练和推理的规模化发展需求，打造了灵骏超级智算集群，提供可扩容到 10 万张 GPU 卡规模的能力，主要包含四个重要组件：灵骏计算集群、HPN 高性能网络、磐久 AI 计算服务器，以及 CPFS 高性能存储集群。**火山引擎**近年来在智算中心领域布局迅猛，通过“自建+合作”模式，在内蒙古、安徽等重点区域投建大型绿色智算中心，并自研底层技术栈，目标是构建支撑 AI 大模型训练与推理的高效能算力网络。**百度智能云**聚焦在阳泉、沈阳、盐城等城市构建普惠的 AI 算力基础设施，通过“AI 大底座”输出整体解决方案，赋能地方产业数字化转型和智能化升级，特别是在自动驾驶等领域有较深积累。**商汤科技**突出“AI 大装置”基座能力，追求超大规模集群的性能和效率，并通过“算力 Mall”等模式降低 AI 使用门槛，在支撑自身前沿研究和多元业务的同时，也向行业输出能力。

二、人工智能算力基础设施赋能的重要走势

目前人工智能算力基础设施利用负载情况差异较大，尤其是国资平台主导的地方智算中心，赋能价值有待进一步提升。面向“十五五”，

人工智能算力基础设施推进科学赋能，围绕需求场景、关键能力、落地生态，有三个重要的发展趋势。

（一）需求场景逐步明晰，促使智算资源优配

需求场景定位日益清晰，助推智算中心精准赋能。“十四五”以来，各地政府及相关主体积极探索、系统推进，人工智能算力基础设施建设正逐步由“建得好”向“用得好”转变。地方及央国企在推进智算中心建设过程中，正逐步厘清核心服务对象与投资建设主体的边界，深化对地方特色经济和央国企核心业务智能化转型痛点、真实算力需求类型及应用优先级的系统性洞察，推动智算建设与行业应用紧密结合，实现从“以建促用”到“以用带建”的科学路径转变。同时，政府、央国企、技术提供商、应用开发商等各方权责体系正在不断明晰，协同机制持续完善。这种系统性定位的日益清晰，正有力促进资源优化配置，避免重复建设与服务偏离，保障智算中心高效运行，显著提升投资回报水平，为数字经济发展注入强劲动能。

（二）关键能力逐步聚焦，提升智算服务水准

关键能力供给持续强化，提升智算中心服务效能。“十四五”期间，人工智能算力基础设施的服务向高层次、全栈化的支持体系快速演进。在基础支撑方面，智算中心正从基础算力资源供给，向全面支持异构计算资源智能管理、大规模集群高效调度、高带宽低延迟网络传输及高可用容灾体系等核心能力加快演进，为复杂业务场景提供更坚实支撑。在创新服务方面，对前沿 AI 框架、工具链、行业大模型开发与软硬件协同创新的支持力度不断加大，预训练模型库、行业知

识库与低代码平台等应用加速普及，显著降低 AI 应用门槛，有效激发本地创新活力。在运营保障方面，运营保障体系日趋完善，正在扭转“重建设轻运营”现象，专业化的算力调度优化能力与模型全生命周期管理服务逐步落地，用户体验和运营可持续性稳步提升，助力实现智算中心的价值闭环与长效发展。

（三）落地生态逐步聚和，加速智算价值释放

生态体系加速整合，协同机制持续完善，有力促进智算赋能价值规模化释放。“十四五”期间，人工智能算力基础设施建设正逐步由基础算力供给，向“算力+算法+数据+场景+服务”一体化解决方案能力方向演进，与地方特色产业及央国企核心业务的融合不断深化。可持续、高价值的合作伙伴网络初步构建，一批具备行业专业知识、能够提供垂直领域解决方案的独立软件开发商（ISV）、系统集成商（SI），以及关键数据供给方和算法研究机构等核心主体加快集聚，合作模式由项目制向更稳定、长期的协同关系过渡。有效生态协作机制初步形成，利益共享、协同创新与风险共担的规则体系逐步健全，有助于降低协作成本、提升主体互信。整体生态建设正朝着更加系统、稳健的方向发展，与实际需求场景及关键能力供给的衔接更为紧密，为智算中心实现长期健康、可持续发展提供了有力支撑。

三、人工智能算力基础设施需求场景

大模型计算需求场景主要包括训练、微调以及推理，模型参数规模与对算力的消耗成正比，参数规模越大，对智能算力的需求越大。不同体量的智算中心支撑不同的大模型计算场景。推进基础预训练大

模型（千亿级以上参数）的训练，需要具备 E 级（EFlops）计算能力的高端万卡集群中心支撑。推进小模型(百亿级及以下参数)的训练、微调，或推进模型的推理，依托百 P 级（100PFlops）计算能力的中小型智算中心即可支撑。

（一）大模型预训练场景

万卡集群推进支撑基础大模型（千亿级以上参数）预训练。大模型训练阶段消耗的资源主要集中在预训练阶段，需要数千至上万块 GPU 并行运算、处理千亿级至万亿级 Token 数据、耗时数周至数月，占总算力消耗的 90-99%。随着基础大模型参数量从千亿迈向万亿，大模型预训练过程对底层智能算力的诉求进一步升级。头部基础大模型的训练算力需求已达到十万亿兆量级，且仍以每年 4.1 倍的速度快速增长。据相关数据显示，OpenAI 依托 2.5 万张英伟达 A100 GPU，处理了 13 万亿个 token，用时 100 天才完成 GPT-4 模型预训练。Meta 的 LLaMA-3 则动用约 1.6 万张英伟达 H100 GPU 在 54 天内训练 15 万亿 Token。由此可见，基础大模型预训练迫切需要高质量万卡智算集群支持。

国内通信运营商、AI 头部厂商积极建设万卡智算集群，持续研发推出基础通用大模型。中国电信人工智能研究院依托天翼云上海临港国产万卡算力池，并基于天翼云自研“息壤一体化智算服务平台”和电信人工智能公司自研“星海 AI 平台”的支持，可以实现万亿参数大模型的常稳训练，自主研发了国内首个全尺寸、全模态、全国产化的万亿参数“星辰”基础大模型体系。**中国移动**依托国产万卡级智算集群，

与多款国产芯片完成了深度适配优化，预训练数据量达 15 万亿（T）tokens 数据，完成了九天大模型（2000 亿参数）高效训练。**阿里**依托阿里云飞天平台的万卡 GPU 集群，完成通义千问 Qwen3（2350 亿参数）预训练，预训练数据量达 36 万亿（T）token，是前代 Qwen2.5 的两倍。**百度智能云**推出国内首个自研昆仑芯三代万卡集群，采用昆仑芯 P800 GPU，目前该集群已通过中国信通院测评，成为首个获“五星级”认证的国产万卡集群，可同时承载多个千亿参数大模型的全量训练。

（二）大模型微调场景

小体量智算中心可有效推进行业模型微调训练。大模型微调训练是连接预训练模型与下游应用场景的关键环节。90%的训练场景主要集中在 L1/L2 大模型微调，即开展百亿级以下参数的行业模型训练与微调。L1/L2 大模型微调是指基于预训练好的大型语言模型，通过调整模型参数以适应特定任务或数据集的过程，以使模型在特定任务上的性能得到显著提升。与基础大模型的训练相比，L1L2 大模型微调对智算资源需求规模成指数级下降。地方推出的小体量（百 P 级）智算中心在百亿级以下参数的行业模型微调方面具有显著优势。一方面，地方可以通过快速整合当地计算资源满足模型微调需求；另一方面，在保证性能的同时，还能够显著降低计算成本和时间成本。整体来看，地方推出的小体量（百 P 级）智算中心，以其高效、灵活的计算资源，足以满足当前主流行业模型训练微调需求，实现对特定区域进行精准服务。

当前国内多数智算中心着力支撑行业模型微调训练。南京智能计算中心已完成超过 150 种主流大模型的调优与适配工作，贯穿模型训练构建、高质量数据应用直至实际场景部署的全过程，可量身实施大模型微调策略，整合包括分布式训练效能提升、模型轻量级实施及边缘节点计算在内的核心能力，现已为上百家科研机构、高等院校及创新型企业提供高性能算力资源服务。**杭州人工智能计算中心**依托全国产软硬件平台，帮助企业用户在行业模型训练中显著缩短时间成本，提供不同规模的 DeepSeek 蒸馏版本模型，可按需灵活选择，覆盖金融、医疗、教育、制造等多行业场景，企业用户无需从头搭建底层模型，即可快速进行业务数据的迁移学习或微调训练。目前，已服务本地校企单位 500 余家，培育行业大模型 30 余个，孵化行业应用与解决方案 200 余个。

（三）大模型推理场景

当前推理需求场景中，云侧推理需求占据主导。推理智算需求场景包括网页端智能助手、移动端智能助手和企业侧应用等。其中**网页端智能助手**需要大量的实时计算资源来支持用户的高并发请求和快速响应，当前主要依托云侧推理完成。由于需要处理大量并发请求，对网络带宽智能资源调度和优化延迟也有较高要求。具体场景包括图像处理、信息检索、智能问答等。**移动端智能助手**，通常对实时性和功耗有较高要求，一般多采用轻量化模型并结合云端推理的方式，移动端主要处理前端数据采集和初步处理，复杂计算任务则交由云端处理，具体场景包括手机语音助手、语音识别、自然语言处理等。对于

企业侧应用，对于推理精度、稳定性和实时性要求均相对较高，企业需要构建专门的智算中心或采用云服务提供商的智算解决方案，另一方面需要低延迟、高带宽的网络连接以满足场景实时性要求。具体场景包括智能客服、智能制造等。

不同推理应用场景对于推理模型及智算中心需求各不相同。在文本对话、智能客服等互联网实时性要求高的推理场景中，时延要求一般在 50ms，可依托 GPT3 或 GLM 等百亿级大语言模型实现低时延应用推理；对于机器人语音对话等推理场景，时延要求一般在 100ms，可基于 Llama 等亿/十亿级别模型来实现推理；在文生图、视频等非实时性交互推理场景中，时延在 200ms，基于 Stability 等模型即可实现推理。整体来看，应对不同的推理场景，智算中心在加速卡选型方面有针对性的方案，以实现最佳的性能和效率。针对大模型推理应用场景，智算中心倾向于选择配备较大内存的训练卡来支撑推理过程，或者采用训练与推理一体化的解决方案，根据推理工作负载的需求，动态调整算力资源，通过“削峰填谷”的方式，来实现推理算力资源的高效利用，以及智算资源的错峰利用。对于实时性要求较高的小模型推理场景，智算中心同样需要选用训练卡来支撑推理工作，以满足快速响应和高效处理的要求。对于实时性要求低的小模型推理场景，智算中心可以选择专用推理卡来支撑推理任务，以优化成本效益并满足基本的处理需求。

专用于推理的智算中心持续涌现。杭州灵汐类脑智算集群已于 2025 年 7 月底实现了大模型快速推理 API 的企业服务试运行，该智

算集群部署异构融合类脑芯片，具有兼容 PyTorch 框架的类 CUDA 软件栈，可直接服务于多类开源大模型的快速推理、而不需要转换类脑算法，并通过存算一体、众核并行、稀疏计算、事件驱动等特性，实现将单用户的推理延迟控制在毫秒级别，首 token 延迟可降至百毫秒乃至十毫秒级，大幅降低智算中心的功耗。**山东移动千卡资源池**采用中兴通讯全栈全场景的智算解决方案，硬件层面使用中兴高性能智算服务器和自研 ROCE 交换机，算力资源包括 304 张天垓 150 GPU 和 720 张天垓 100 GPU 卡，实现同厂家异代算力统一纳管，软件层面部署了中兴通讯 TECS 资源管理平台和 AIS 平台，通过推理引擎二次优化，大幅提升推理资源池性能。**广东电信**基于中国电信粤港澳大湾区（韶关）算力集群已上线**昇腾超节点智算集群**，采用中国电信研究院自研的“翼芯”智算测试与适配优化平台，针对多种推理场景开展了大模型性能优化及测试，通过对主流模型与昇腾超节点的适配调优，不同场景下的推理性能均实现了大幅提升，通过尝试采用算子融合替换、PD 分离调度、KV cache 优化、集合通信优化、并行策略优化等多维度的调优策略，DeepSeek 671B 模型在多种场景下的单卡推理吞吐性能较优化前提升 2.5~4.3 倍。

四、人工智能算力基础设施关键能力

整体来看，当前人工智能算力基础设施正从“重硬轻软”向“软硬协同、服务赋能”加速演进。在持续夯实底层算力支撑能力的同时，各方日益重视提升创新服务与运营保障能力，不断拓展服务边界、增强发展韧性。面对算力市场对高阶技术服务的迅猛增长需求，智算中

心正加快面向应用场景系统构建关键能力，有效提升应对市场波动和实现可持续发展的综合实力。

（一）基础支撑能力

基础支撑能力是智算中心基础技术能力的核心体现，为用户提供最核心的技术服务。训练场景主要关注集群算力有效性、集群稳定性、单体集群算力规模，以及主流计算框架的兼容性等。集群算力有效性主要指智算中心算力的利用率，决定了智算中心最终的有效算力供给能力。在实际应用中，算力有效性普遍不高，通过尽可能降低在多卡互联、多机互联中的算力损耗，能够提升集群算力有效性。**计算集群稳定性**是智算中心可稳定支撑模型训练长时间运行的能力，直接关系到 AI 大模型训练的连续性和效率，目前可以通过冗余设计、负载均衡、数据备份等方式来提升智算中心的稳定性。**单体集群算力规模**是可支撑模型训练的单体集群算力规模上限，主要是对大规模计算需求的支撑能力。此外，底层算力卡可兼容主流计算框架的能力、支持多种主流通用基础大模型的能力、支持多种主流通用数据集及行业数据集等能力等，也是智算中心需要关注的基础支撑能力指标。**推理场景主要关注 token 吞吐率、时延以及智算卡的异构多样性。吞吐率**是智算中心支撑推理服务在所有用户请求中每秒可生成的输出 Token 数，高吞吐率意味着可以更快地响应用户请求，因此也是衡量智算中心对推理场景支持的重要指标。**端到端时延**为用户生成完整响应所需的总时间，同样，时延也影响着对用户的响应情况。**异构多样性**是指智算中心提供多元异构智算加速卡供用户选择，这是响应不同模型对算力

个性化需求的支撑能力。

表 1 智算中心基础支撑能力重点指标

能力	训练场景		推理场景	
	指标	含义	指标	含义
基础支撑能力	**集群算力有效性**	尽可能降低在多卡互联、多级互联中的算力损耗	**吞吐率**	智算中心支撑推理服务在所有用户请求中每秒可生成的输出词元(Token)数
	计算集群稳定性	计算集群可稳定支撑模型训练长时间运行	**端到端时延**	为用户生成完整响应所需的总时间。
	单体集群算力规模	可支撑模型训练的单体集群算力规模上限	**异构多样性**	提供多元异构智算加速卡供用户选择
	主流计算框架兼容性	底层算力卡是否可兼容主流计算框架	云服务高效性	可通过云方式提供高效的算力服务
	算法模型多样性	支持多种主流通用基础大模型		
	数据集丰富性	支持多种主流通用数据集及行业数据集		

大模型预训练对智算中心的绝对算力性能有强要求。反向传播中的梯度计算和参数更新均是计算密集型任务，因此模型训练性能是训练阶段最核心诉求，主要体现为在一定的智算资源下缩短训练花费的时间。训练消耗时间主要包括：数据加载时间、模型前反向时间、优化器时间、模型后处理时间、通信时间、调度时间等，能反映训练性能的主要指标包括吞吐率、单步时间、线性加速比、模型算力利用率等。其中集群线性加速比和模型计算利用率是集群算力性能的关键指标。**集群线性加速比**，指单机拓展到集群的效率度量指标，是集群算力性能的核心指标之一，高性能网络是让线性加速比尽可能逼近于 1 的关键，在高性能网络优化下，集群加速比可达到 90%以上。摩尔线

程夸娥（KUAE）智算中心实现了系统级全栈协同优化，覆盖硬件、软件、集群及云服务，提供全局综合解决方案。其在 70B 至 130B 参数的大模型训练中，线性加速比均可达 91%。**模型计算利用率** MFU（Model FLOPS Utilization）是一个用于评估人工智能加速器在模型训练期间利用程度的指标，表示在模型训练期间实际使用的浮点运算数（FLOPS）与理论上可用的 FLOPS 之间的比率，高 MFU 表明加速器在模型训练中被充分利用。从业界实践调研结果看，智算集群算力有效性能普遍不高，达到 40-50%属于较为优秀。部分主体在特定条件探索，可超过 50%。

模型推理对智算卡的内存和通信带宽有强要求。对于推理场景，模型推理目标是首 Token 输出尽可能快、吞吐量尽可能高以及每个输出 Token 的时间尽可能短，因此模型推理核心要求是高吞吐量和低时延。对于智算卡而言，推理场景的高吞吐量和低时延，对其内存和通信带宽有着较高的要求；一方面，智算卡需要具备充足内存容量，以满足推理过程中快速加载和存储大量数据以及模型参数的要求，为高效的推理提供存储基础。另一方面，通过高通信带宽确保数据在智算卡与其他设备之间能够快速传输，使得输入数据迅速得到智算卡处理，同时推理结果能够及时传输回应用程序，以减少数据传输时间损耗。由于推理过程主要是基于已训练好的模型对输入数据进行处理和输出结果，无需像模型训练需要进行大量复杂的计算操作，所以相比之下，推理场景对计算的需求相对较低。**高内存和通信带宽是实现推理应用高吞吐量和低时延的关键。**整体来看，**应对不同的推理场景，智**

算中心在加速卡选型方面有针对性的方案，以实现最佳的性能和效率。

（二）创新服务能力

创新服务能力是智算中心推进产业创新的核心体现，为用户提供高阶价值的技术服务。训练场景主要关注云服务高效性、模型迁移高效性以及数据治理多样性。云服务高效性，即可以通过云方式提供高效的算力服务的能力，也是智算中心场景应用支撑的重要指标，通过云来提供算力服务，是智算中心发展的重要趋势。**模型迁移高效性**指智算中心可高效完成用户模型的迁移适配，决定着模型是否能快速进入产业化阶段。**数据治理多样性**是指智算中心可以通过为用户提供多样的数据汇聚、共享、清洗等工具，帮助用户实现模型的落地应用。**推理场景主要关注智算资源池化调度能力、模型迁移部署高效性。池化调度能力**通过支持异构算力的统筹调度来衡量，集中管理和调度能够提高资源利用效率、降低成本、支持异构算力管理，并提供弹性计算服务的能力，推进池化调度。

表 2 智算中心创新服务能力重点指标

能力	训练场景		推理场景	
	指标	含义	指标	含义
创新服务能力	**云服务高效性**	可通过云方式提供高效的算力服务	**池化调度能力**	支持异构算力的统筹调度，并推进池化调度
	模型迁移高效性	可高效完成用户模型的迁移适配	**模型迁移高效性**	可高效完成用户模型的迁移适配
	数据治理多样性	为用户提供多样的数据汇聚、共享、清洗等工具	开发工具完整性	提供丰富完整的模型量化、剪枝、部署开发工具
	开发工具完整性	提供丰富完整的模型训练、推理、部署开		

能力	训练场景		推理场景	
	指标	含义	指标	含义
		发工具		
	场景方案丰富性	提供丰富的行业场景模型及行业场景解决方案样例		

模型训练突出智算中心的全栈软件能力要求，要求智算中心提供训练过程的全栈 MaaS 服务能力。全栈服务能力实现从硬件适配、资源池化到异构调度的完整全栈一云多芯，向下纳管异构芯片资源、向上屏蔽硬件差异，保障训练任务高效稳定运行。在模型开发阶段，需要智算中心提供包括模型训练、调优和部署等在内的全栈平台型服务，以支持低门槛的模型开发与定制，用户无需关注 AI 算力、框架和平台即可生产和部署模型。对于训练所需数据，需要智算中心支撑数据工程，提供包括大小模型及公私域数据集的丰富资产库服务，以支持模型和数据集的灵活快速调用，用户无需生产和部署模型即可调用模型和数据集服务。

智算中心的池化调度能力同样是推理场景关注的重要指标之一。在实际生产部署中，AI 推理往往与前端的业务/应用网络形成紧密配合，经由智算中心对外提供云服务，因此要求智算中心要能够支持提供各种异构算力（GPU、CPU、NPU）的能力，实现一云多芯调度。在具体应用中，智算中心应兼容华为昇腾、海光等国内外主流 AI 芯片，确保推理任务能够在由不同品牌、不同型号芯片组成的智算集群中顺利执行混合推理。此外，还需整合 GPU 硬分片和虚拟分片技术，实现 GPU 资源的池化管理以及跨集群调度能力，从而实现对多芯集

群的精细化运营，使得推理算力能够灵活应对各种不同类型的任务处理需求。此外，推理场景多应用于产业一线，对于底层算力的地理位置、端应用服务的快速连接性等要求较为严格，算力供给主体需具备海量的、可扩缩容的高性能算力资源，并确保算力能够稳定、可靠地交付给用户使用。

（三）运营保障能力

运营保障能力是智算中心实现科学运转的核心体现，为用户提供闭环商业服务。训练和推理场景都重点关注算力被调度的灵活性、算力租赁性价比、安全合规性等，此外训练场景还关注可协调合作主体的丰富性。**算力调度灵活性，**是指智算中心可依托外部算力调度平台被灵活调度，通过调度平台可对接更多区域外用户主体，从而提升智算资源利用率。**算力租赁性价比**也是训练和推理场景中用户都关注的重点指标。**安全合规性**，是指智算中心建设覆盖大模型全生命周期的安全服务能力，包括合规咨询、内容安全、数据防护及评测体系，保障用户安全合规地部署大模型应用。此外，服务的响应、服务质量跟踪等运营能力，也是用户较为关注的维度。

表 3 智算中心运营保障能力重点指标

能力	训练场景		推理场景	
	指标	含义	指标	含义
运营保障能力	**算力调度灵活性**	可依托外部算力调度平台被灵活调度算力池资源	**算力调度灵活性**	可依托外部算力调度平台被灵活调度算力池资源
	算力租赁性价比	可提供高性价比的算力租赁服务	**算力租赁性价比**	可提供高性价比的算力租赁服务
	安全合规性	保障用户安全合规地	**安全合规性**	保障用户安全合规

能力	训练场景		推理场景	
	指标	含义	指标	含义
		部署大模型训练		地部署大模型推理
	服务响应时效性	可快速响应用户的各类服务需求	服务响应时效性	可快速响应用户的各类服务需求
	协调主体丰富性	可协调丰富的产业主体资源	服务质量跟踪	支持面向用户的服务质量全过程跟踪
	服务质量跟踪	支持面向用户的服务质量全过程跟踪		

训练和推理场景均需要推进智能算力的灵活调度。智算中心算力资源被算力调度平台整合后，可实现资源跨区域调度，当某一区域的智能算力需求激增时，平台可以快速调用邻近区域的闲置智算资源，确保整体供给稳定、成本均衡，降低单一地区资源不足或故障的风险。通过纳入算力调度平台，智算中心运营方可借力平台对接跨区域算力需求主体，有效提升算力消纳利用率。**上海市智能算力资源统筹调度服务平台**以市场算力需求为导向，接入上海本地及市外各类空闲算力资源，可实现跨市域、跨资源池、跨厂商的异构算力资源交易和调度，截止 2025 年 7 月，平台已接入**上海仪电、上海电信、上海移动、上海联通以及阿里、百度等互联网云厂商、商汤、算丰**等第三方算力企业的算力，并实现新疆克拉玛依、湖南长沙的优质算力接入，上架总算力规模已超过 1.8 万 P。**无锡城市智算云平台**统筹调度全市多元算力，目前已汇聚包含**云工场边缘算力节点、滨湖马山、惠山尚航**等市内外 20 个算力中心共 3555P 智算资源，其中市外算力 524P，目前平台已经为 17 家企业提供算力服务，成交额达 2399 万元。此外，平台也于去年底成功入选工信部智算云服务国家级试点，成为 8 大试点之

一。

训练和推理场景都追求高性价比，推动智算中心加速卡选型持续演进。当前随着模型算法持续优化，低精度下可以部分支持大模型推理以及部分条件下的微调训练，智算中心的加速卡的选型也在演变，由配置高性能板卡，转向更匹配业务需求、性价比更高的技术选型。团队对各类智算中心主体的训练卡与推理卡服务价格进行了调研与梳理，调研对象覆盖电信运营商、地方国资平台及第三方算力服务商等，通常预训练服务以台月为单位报价，而面向微调与推理任务则以卡时为计价单位。

训练和推理场景均注重提升安全合规性。当前人工智能大模型快速发展的背景下，地方智算中心确实需要将提升安全合规性置于重要战略位置，以保障大模型从开发到应用的全生命周期安全。智算中心推进服务的安全合规主要包括数据语料安全、内容安全、网络安全防护等。百度在地方智算中心建设运营中，提供从模型训练语料清洗、输入输出内容实时过滤，到多模态风险识别和隐私保护的端到端解决方案。同时，百度支持智算中心提供大模型备案全流程服务，具备面向公众开放及企业内部不同场景的合规建设能力，并推出轻量化终端安全方案，满足低算力环境下的内容审核与风险管控需求。

五、人工智能算力基础设施落地生态

各类智算中心推进需求场景的支撑，对场景所需关键能力供给主体提出强合作需求，如数据要素主体、算法模型主体、跨域算力主体、行业场景主体、区域产业主体等，只有深度推进智算中心与各类要素

供给主体协同合作，才能助力智算中心赋能真正落地。

（一）智算与数据要素协同

推进与高价值数据的密切协同，是智算中心提升基础支撑能力的核心所在。数据不仅为智算中心提供模型训练与推理所需的核心资源，更为智算中心连接多元主体、凝聚产业合力提供了关键基础。各地智算中心在提供基础智能算力服务的同时，积极深化与公共数据资源主体、通用数据集建设机构以及专业数据治理单位的合作，通过接入政务、行业及公共领域等多源数据，吸引算法开发、模型训练与应用落地的各方力量，不断提升智算中心服务的精准性与多样性，促进高质量数据资源向智算生态能力转化。“算力+数据”双轮驱动不仅提升了智算中心本身的资源效率与服务能级，也为其构建开放、共生、可持续的智算生态提供了坚实支撑。

温州市人工智能计算中心由温州市数据集团投用，今年年初温州市数据集团联合浙江省大数据联合计算中心有限公司、每日互动公司发布全国首个基于可信数据空间和 DeepSeek 双重技术的自主创新大模型服务，依托数安港可信数据空间和温州智算中心等保 2.0 架构，结合 DeepSeek 模型数据传输和存储加密技术，为相关产业提供基于可信数据空间的可控大模型租用服务、私有化部署和精调服务。贵安新区正由过去的“存储中心”加速向“存算一体、智算优先”迭代，**全国一体化算力网络国家（贵州）主枢纽中心**截至当前累计建成 116P 高性能智算中心与 50PB 非结构化数据存储平台，并已面向省内外企业提供数算一体化服务，实现东部算力与存力向贵州转移。该中心作为

全国一体化算力网络八大国家枢纽节点之一，旨在打造国产算力适配中心、高质量训练数据集和面向全国的算力保障基地。通过不断挖掘数据价值，探索将大数据融入实际应用，加快政用、民用、商用数字场景融合，激活数据要素价值。

（二）智算与模型算法协同

推进与高水平模型算法的密切协同，是智算中心提升创新服务能力的关键路径。通过推动算力平台与多元模型架构的深度适配与优化，能够有效打通从底层硬件支撑、核心算法研发到上层应用服务的全链条技术体系，显著提升公共算力服务的普惠性与实用性。这种“算力+模型”一体化模式，不仅为科研机构与企业提供了从训练推理到场景化部署的端到端支持，加速垂直行业专属模型的开发与应用迭代，更通过构建多元模型协同的服务生态，形成功能互补、安全可靠的智算供给能力，有效满足差异化、多层次的智能应用需求。其对促进区域产业数字化转型升级作用显著，能够在智能制造、政务服务、文化教育、医疗卫生等重点领域形成规模化赋能效应，为培育新质生产力、实现高水平科技自立自强提供坚实支撑。

重庆人工智能创新中心将自身**昇**腾 Atlas 系列硬件深度适配 DeepSeek-R1 系列模型，适配范围包含参数量从 1.5B 到 70B 的六种不同模型。这些模型能够满足科研攻坚、行业实践、高效编码、内容制作等多方面需求，助力用户在垂直领域构建专属模型，构建起从基础模型、AI 算力到服务部署的全链条国产化服务能力。**武陵山（利川）人工智能计算中心**于 2025 年 2 月宣布完成国产开源大模型

DeepSeek 全版本推理服务部署，结合 2024 年已部署的讯飞星火大模型，构建出中西部地区首个实现“星火+DeepSeek 双底座”协同的县域智算中心，上线仅三个月，填充率便突破了 80%，成功在 AI+文旅、AI+教育、AI+政务、AI+医疗等领域实现赋能。同时武陵山（利川）人工智能计算中心也成功入选中国信通院 2025 年智算中心典型案例。**中兴通讯携手某电网企业共建国产千卡智算资源池**，联合完成电网基础 L0 系列大模型及 Deepseek、Qwen-QWQ 等 140+模型训推适配，通过迁移适配工具高效进行模型分析、算子开发、优化部署，加快模型适配速度，模型推理精度与英伟达持平，吞吐和推理时延等关键指标得到显著提升，助力电力行业高质量发展。

（三）智算与跨域智算协同

推进跨域智算互联协同，是智算中心运营能力跃升的重要探索。人工智能技术及生态迭代加快，大模型以及相关应用的发展对智能算力提出更强更大规模的需求。为满足万亿及以上量级参数量大模型训练需求，支撑多场景、多业务、大流量入算的智算业务，智算中心建设亟需突破单点物理中心规模受限、电力供应不足等瓶颈，智算中心间的互联成为重要补充，具备长距无损、高吞吐量、算间高效协同等互联能力的分布式智算集群应运而生。训练拉远场景下，通过智算互联进行分布式训练可以弥补单智算中心算力不足的问题，将闲置算力进行整合；存储拉远场景下，计算集群和存储集群互联可解决算力训练处理过程中会存在部分数据样本安全问题，满足数据本地化需求。当前，针对长距高可靠、任务式带宽、高效流量调度等智算中心需求，

产业界已进行了诸多探索。

运营商智算中心长距互联实现实践突破。中国联通临港智算中心通过智算、网络多项创新技术的综合运用，成功完成 AI 大模型 300 公里分布式协同训练技术验证，跨域分布式训练等效算力达到单集群的 95%以上，跨域带宽收敛比大于 16:1，为 AI 大模型训练模式提供了全新的解决方案。**中国移动**联合新华三、朗美通在河北移动鹿泉智算中心，完成业界首次 800G 以太网智算协同训练的现网技术试验，降低 40%单比特成本、35%功耗及 20%节点时延。在跨智算中心 700 亿参数大模型训练中，实现高达 98%以上的等效算力效率，实现探索跨智算中心互联的新架构和新技术的重要突破。

（四）智算与行业场景协同

推进与行业场景的密切协同，是智算中心生态持续演进升级的核心动力。行业重大场景是驱动智算中心构建协同开放、持续进化生态体系的核心依托。自动驾驶、智慧交通、政务服务等重大应用场景不仅为智算中心提供了持续的真实需求与应用验证环境，也推动其从提供基础算力服务向构建场景适配、功能专用、效能可评估的新型服务模式演进。智算中心依托真实具体场景不断验证、调优和迭代其算力与算法服务体系，吸引行业内技术供给方、解决方案开发商与行业用户共同参与，促进技术、资源与应用的深度融合，加快形成“以算促用、以用带算”的良性发展机制，形成更加开放、协同和可持续的产业生态。

长安汽车与百度智能云共建的长安汽车智算中心，是智算中心与

汽车产业深度协同的典型实践。以百度智能云的百度百舸 AI 异构计算平台为底座，已形成 142 亿亿次/秒的计算能力，支撑长安汽车构建起覆盖数据采集、处理、标注、训练到模型部署的全流程“星环平台”，实现跨集群算力与存储资源的统一调度。截至当前，长安汽车基于该平台已累积近亿帧的高质量标注数据，完成超 3 万次的 AI 算法模型训练，显著加速了智能网联与自动驾驶研发进程。双方在大模型领域深度联动，将百度文心一言大模型应用于车型智能对话及企业知识管理，并依托智算中心提供算力优化服务，助力长安汽车自研行业大模型。**云南交投集团依托近 20 年的交通数据积淀，建成云南交投智算中心**，其中通用算力超 4.5 万核，智能算力达 100P、综合存储能力 10PB，可支撑交通行业百亿参数大模型训练、上万个交通业务和政务系统运行。围绕交通场景的痛点和需求，云南交投智算中心本地化部署了盘古、DeepSeek 等大模型，集成自然语言、视觉、多模态、时空预测等能力，构建了统一高效的模型开发管理平台与工具链，具备数据工程、模型工程、智能体开发、监控运维等全流程支撑能力。同时，还构建了 600GB 行业知识语料库和 400G 企业专有知识库，形成覆盖“建管养运服安”全场景的评测数据集。截止到目前，云南交投智算中心已经打造了 36 个“人工智能+智慧交通”场景。

（五）智算与区域产业协同

推进与区域产业的密切协同，是智算中心实现多维度、全场景赋能的重要保障。依托算力资源可以吸引企业集聚，促进跨领域合作，联合孵化面向多场景的解决方案，通过推动“政产学研用”深度融合，

打通技术研发、成果转化、产业孵化和人才培育全链条，形成创新闭环和良好生态。加速人工智能与传统产业融合，赋能智能制造、智慧城市、数字政务等重点领域数字化转型，构建区域人工智能技术和产业高地，形成以智算引领技术创新、以平台支撑产业聚合、以生态驱动持续发展的新格局，为经济高质量发展注入新动能。

宁波人工智能超算中心打造自主可控、“智+超”融合的综合性人工智能创新平台，通过“超算+智算”的综合算力模型打造产业发展、科学研究、社会治理等领域更高算力适配和算法匹配优势。自 2023 年 1 月 10 日上线以来，中心算力使用率超 80%，累计运行任务作业超万个，已服务 40 余家用户单位。**武汉人工智能计算中心**于 2021 年 5 月投入运行，累计算力规模达 400PFLOPS，目前已吸引 400 多家企业入驻，联合孵化出 300+项场景化解决方案，打造覆盖智能制造、智慧城市、数字农业、自动驾驶、智慧医疗、智慧政务等应用场景，逐渐形成“两院四科研”的科研雁阵，在武汉打造多模态、遥感两大技术高地，组建双产业联合体，形成产业聚集区。**大连人工智能计算中心**重点打造“一中心四平台”，提供公共算力服务平台、应用创新孵化平台、产业聚合发展平台、科研创新人才培养平台，实现了“政产学研用”五位一体，形成东北乃至全国人工智能产业汇聚。已为 174 家科研团队、55 家企业提供算力服务，为用户开设主账户 229 个，孵化人工智能垂直解决方案 150 余个，行业大模型 5 个，为区域产业数字化、数字产业化提供强大算力支持和保障。

六、发展展望

人工智能算力基础设施作为数字经济的核心底座，正加速赋能千行百业智能化转型，为发展新质生产力注入强劲动能。**面对当前存在的赋能效应不凸显问题，政府部门、国资平台、投建主体、运营主体等智算中心相关方，亟需围绕需求场景、关键能力、落地生态三大环节，着力建立定位明晰、能力聚焦、生态聚合的赋能机制**。未来，随着全国一体化算力网络日益完善，智能算力资源将被高效调度与普惠使用，人工智能算力基础设施终将成为推动区域经济高质量发展、塑造区域竞争新优势的战略力量。

（一）需求场景更加多元、复杂

人工智能算力基础设施正从通用计算走向场景智能，其需求场景将更加多元、复杂且深度融合。随着技术发展和行业数字化进程的加速，其需求场景将呈现以下几个重要变化和拓展。**一是更高性能与更高效的算力与存力需求。**AI 大模型参数量的指数级增长对算力规模和处理效率提出了更高要求，而自动驾驶等实时交互场景则需要实现低延迟、高吞吐的推理智算服务。为了提升大模型训练效率，减少大模型幻觉，提升 AI 推理精度，需要依托与模型算法紧密协同、亲和性更好的 AI 原生存储设备，通过同步建设智算和数据存储基础设施，更高效管理、治理语料库和知识库，发挥存算一体化优势。**二是更广泛与更深入的行业融合。**除传统产业智能化改造对智算中心的需求从“上云”转向“用智”外，一大批新兴模型计算需求场景持续涌现，如 AI 数字人、元宇宙、科学计算等，对算力的异构性和专用性提出

更高要求。**三是更协同与更泛在的算力网络。**推理大模型逐步成熟，AI 应用需支持边缘侧和端侧部署，实现“云端训练、边缘推理”，亟需智算中心推进云边端协同深化，对网络延迟、带宽和成本控制要求更高。

政府部门应做好智算中心需求场景的引导员。一是顶层设计与规划先行，结合本地产业禀赋，制定区域性智算中心发展规划，明确重点支撑的行业方向，引导算力资源与产业需求精准匹配，防止“一拥而上”和低水平重复建设。**二是开放应用场景**，在智慧城市、智能政务、智慧交通、城市治理等领域开放大量应用场景，为本地智算服务提供“首购首用”的市场机会。**地方国投平台应做好智算中心需求场景的连接器。一是链接区域有迫切需求的企业**，推动智算中心与本地传统龙头企业对接，打造行业数智化转型标杆案例。**二是积极投资 AI 应用企业**，将视野从算力中心延伸，更重点投资下游有潜力的 AI 算法、应用解决方案企业，培育延伸产业链。**运营方应做好做好智算中心需求场景的跟踪者。**组建懂行业的技术团队，深入区域重点行业一线挖掘痛点与细分场景需求，提供定制化解决方案，并持续跟进 AI 大模型技术演进方向，推进智算中心需求场景的演进与设施迭代升级。

（二）关键能力更加集约、软性

人工智能算力基础设施正从粗放式的硬件堆砌转向精细化的服务提升。一是从集群到超大规模集群，持续推进集群内高速互联技术的升级，将成千上万个异构计算节点整合，突破小集群性能瓶颈，支撑万亿参数大模型的单一任务训练，实现算力规模和计算效率的同步

极致化。**二是从烟囱式到池化共享**，通过软件定义和虚拟化技术，将异构分散的算力、存储、网络资源抽象成统一的智算资源池，极大提升资源利用率，实现资源的全局优化和弹性供给。**三是从封闭到开放开源**，通过构建开放标准、开源软件栈和开放互联协议，打破不同厂商硬件和软件之间的壁垒，实现异构算力的统一管理和应用的无缝迁移，促进智算产业链协同创新。**四是从资源型到服务型**，依托封装好的算力服务、模型服务、数据服务甚至行业解决方案服务，自动匹配和供给所需资源，极大降低使用门槛，赋能更多非技术背景的行业专家进行创新。

行业组织应持续推进智算中心关键能力体系构建与完善。针对三大类场景，围绕底层支撑能力、创新服务能力、运营保障能力，细化智算中心的关键能力评估指标，丰富指标维度。推出标准化的建设参考依据，来帮助智算中心的供给与市场需求匹配度更高。**运营方应根据赋能场景不同针对性提升关键能力。**从产业实践来看，智算中心亟需夯实关键能力，支撑数据处理服务、算力调度服务（提供弹性的计算、存储、网络等资源）及推理应用服务等。此外，训推一体化方案成为智算中心提升关键能力重要举措，可基于算力底座的资源优化与高效调度，以 AI 模型训练与推理为应用目标，为用户提供 AI 模型开发、训练与推理加速服务。

（三）落地生态更加聚合、协同

人工智能算力基础设施赋能落地取决于其融入、组织和生态协同的意识与实践。一是从单打独斗到群策群力，需要形成由地方政府、

应用企业、技术提供商、高校科研机构、投资机构等共同参与的多元化、网络化共同体。**二是从简单采购到联合创新**，与智算需求方的关系正在转变，不再是简单的交付硬件或标准云服务，而是建立联合实验室、创新中心、产业联盟等载体，进行共创，共同开发方案，共同验证价值。**三是从提供算力到培育产业**，聚焦点不再是智能算力本身，而是聚合数据要素主体、算法模型主体、跨域智算主体，以及行业场景主体、区域产业主体，推动整个区域或行业的智能化升级和产业发展。

政府部门应强化智算中心落地生态的引导与资源对接。一是支持推进智算中心生态伙伴计划，征集、认证、集成合作伙伴的软件和算法，共同面向市场打造“联合解决方案”。**二是打造智算产业园区**，以智算中心为核心，配套建设 AI 产业园区，吸引 AI 算法公司、数据服务商、应用开发商集聚，形成“算力+数据+算法+应用”的产业集群。**三是探索地方政府与央国企业共建“智算中心+AI 数据”落地生态道路。**结合央国企行业数据和地方优势产业及公共数据，针对细分赛道和具体应用场景，依托智算中心，联合研制垂域大模型、智能体行业解决方案等，共建产业生态，并争取国家级政策支持，通过先行先试汇聚央地协同发展动力。**运营方应积极主动对接外部要素主体，深度推进合作。**鉴别需求场景所需的关键能力，遴选合作主体。训练场景提供创新服务能力，需要与智算基础软硬件厂商合作提升模型迁移适配高效性，与 ISV 厂商、场景主体合作提升场景模型、方法丰富性。训练场景提升运营保障能力，需要与区域算力调度平台对接，纳入大

区域调度平台，需要与产业平台智库合作提升可合作主体丰富性。推理场景提供创新服务能力，需要与智算虚拟化池化厂商合作提升池化调度能力。

中国信息通信研究院 产业与规划研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-68021375

传真：010-68033959

网址：www.caict.ac.cn