CAICT 中国信通院

集智 专题报告

智算中心综合评价报告

(2024 年)

中国信息通信研究院云计算与大数据研究所

2024年9月

版权声明

前　言

当前数字化时代，人工智能等新一代信息技术飞跃式发展，智能算力逐渐成为科技产业技术创新、成果转化与应用落地的关键驱动力。我国正积极应对这一技术变革，加强智能算力的布局、研发和应用，以提升国家竞争力。2024 年 3 月，“人工智能+”首次被写入政府工作报告。同年 9 月，工信部等十一部门发布《关于推动新型信息基础设施协调发展有关事项的通知》，再次强调逐步提升智能算力占比。

智算中心正面临着前所未有的发展机遇与挑战。构建一个全面覆盖技术先进性、安全性与可用性，又重点突出服务能力及可持续发展能力的综合评价体系，显得尤为重要且迫切。综合评价体系旨在通过科学的量化分析与客观的价值判断，评估智算中心的发展水平，前瞻性地引领其未来的发展方向。

通过综合评价体系，我们期望为智算中心的建设者提供决策依据，助力其优化资源配置，提升建设质量；为运营者指明管理方向，促进其提升运营效率与服务水平；同时，也为使用者制定透明、可信赖的选择指南，确保数据价值得以最大化实现。促进整个智算生态的良性互动与协同发展，推动智算中心从单一技术设施向集成化、全方位赋能的平台转型，为数字经济的蓬勃发展注入源源不断的动力。

因时间和能力所限，报告内容有所疏漏在所难免，烦请各界不吝指正。如有意见或建议请联系 dceco@caict.ac.cn。

目 录

图 目 录

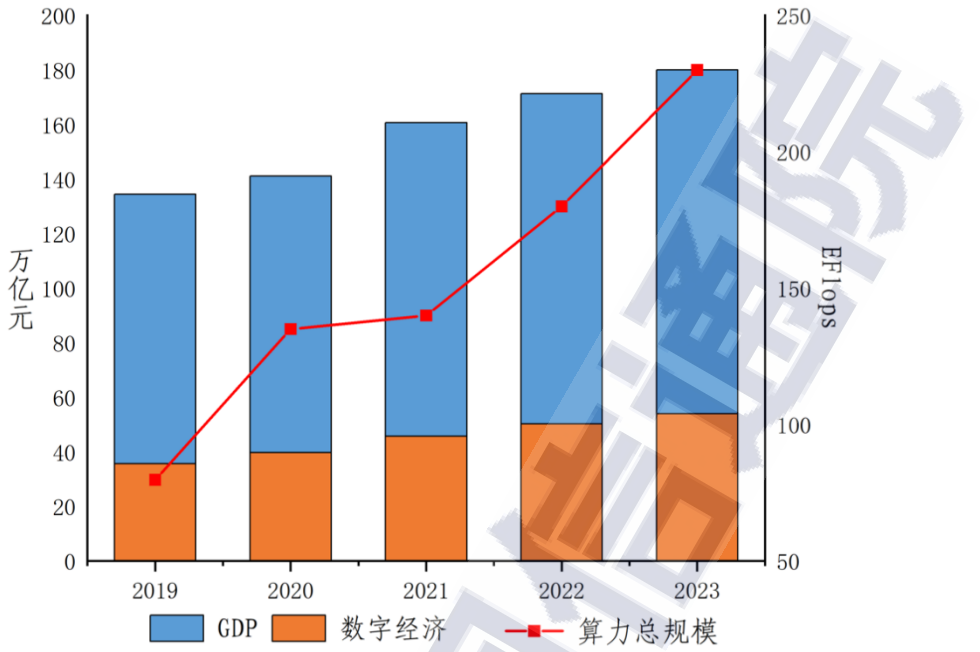
一、智算中心发展背景

随着人工智能技术的日新月异，特别是在大规模模型训练与应用领域取得突破性进展后，当前算力需求呈现出增长态势。党的二十届三中全会明确提出加快推进数字经济与实体经济深度融合，推动数字产业化、产业数字化，为智算中心的发展提供了更为广阔的发展空间和机遇。智算中心不仅成为支持人工智能、大数据等技术在制造业、服务业中的深度应用平台，也是在新型工业化进程中助力产业升级、实现高端化、智能化、绿色化的重要基础设施。在政府政策的积极引导和技术创新的持续驱动下，智算中心建设得到了强有力的支持。同时，市场对高效、稳定算力服务需求的激增，加之开源软件如 Kubernetes、Nomad 等技术的日益成熟，智算中心在产业智能化、企业数字化转型中的作用愈发重要，推动了其持续快速发展。

（一）智能算力成为经济发展新引擎

智能算力是数字经济时代新的生产力，带动数字经济发展和 GDP 增长。数字经济以数据为关键要素，以算力为核心生产力。智能算力支撑人工智能应用简化复杂任务，实现个性化服务，提升生产力水平。当前，算力正从互联网、电子政务等新兴领域向服务、金融、制造、教育等传统行业延伸，赋能传统行业数智化转型，激发经济增长新动能。此外，智能算力的发展也在一定程度上影响着 GDP 增长。从发展趋势看，算力规模与经济发展水平呈现出正相关，数字经济规模和地区生产总值较高的省份，算力发展水平也较高。同时，算力对经济具有辐射带动作用。截至 2023 年底，我国算力规模达到 230 EFLOPS，

算力总规模近 5 年年均增速近 30%，GDP 增长 5.2%。



来源：信通院、网信办、国家统计局

图 1 GDP、数字经济及算力总规模的发展趋势

（二）智能算力需求多层面快速扩张

宏观上，智能算力的发展是新的时代要求。在数字经济时代，让算力真正成为像水电一样的服务，使用户实现一点接入、全算贯通。大国博弈在算力、数据、算法等方面的竞争日益白热化，据不完全统计，自 2017 年起共有 50 余国家发布人工智能战略。习近平总书记强调，要把科技的命脉牢牢掌握在自己手中，在科技自立自强上取得更大进展。**微观上，**算法模型愈加复杂，数据量急剧增长，应用不断延伸，智能算力需求快速扩张。算法模型上，机器学习、神经网络等技术显著进步，算法复杂度和精度不断提高，大模型参数规模呈现指数级增长。从 GPT-3 到 GPT-4 大模型，参数量实现了从 1750 亿到 1.8 万亿的跨越。数据量上，我国数据规模持续扩大，2023 年数据生产总

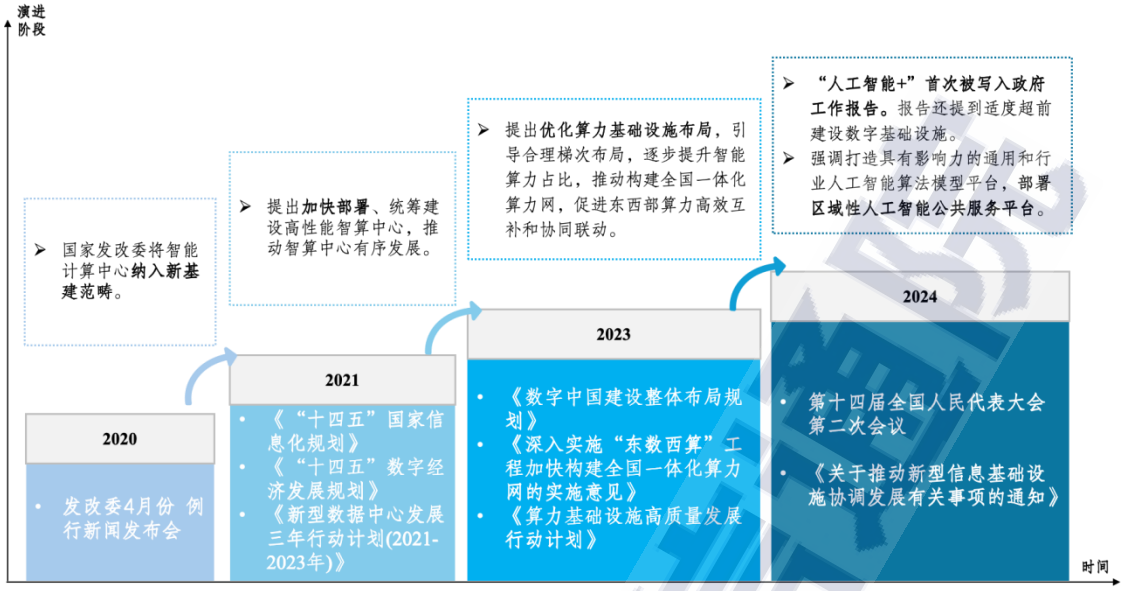
量达到 32.85 泽字节(ZB)[1]，同比增长 22.44%，非结构化数据爆发式增长，通用算力已经难以完成视频编解码、游戏渲染等新型应用背后的大量非结构化数据处理。应用场景上，人工智能在各行业应用程度不断加深，应用场景愈发广泛。自动驾驶、智能家居、医疗影像诊断等新兴场景对智能算力的需求日益旺盛。

（三）国家引导智算中心高质量发展

政策引导力度逐渐加大，推动智算中心高质量发展。2020 年，国家发改委将智能计算中心纳入新基建范畴。2021 年，国家相关部门发布了《“十四五”数字经济发展规划》、《新型数据中心发展三年行动计划(2021-2023 年)》等多项规划，提出加快部署、统筹建设高性能智算中心，推动智算中心有序发展。2023 年，《数字中国建设整体布局规划》和《算力基础设施高质量发展行动计划》相继出台，提出优化算力基础设施布局，引导通用数据中心、超算中心、智能计算中心等合理梯次布局，逐步合理提升智能算力占比。

当前，国家对于智算中心从鼓励建设转向规划布局，政策引导逐步深化，指引方向更加明确。在今年的两会上，“人工智能+”被首次写入政府工作报告，提到适度超前建设数字基础设施。9 月 4 日，工业和信息化部等十一部门重磅发布《关于推动新型信息基础设施协调发展有关事项的通知》，提到打造具有影响力的通用和行业人工智能算法模型平台，部署区域性人工智能公共服务平台。

[1] 《全国数据资源调查报告（2023 年）》



来源：中国信息通信研究院

图 2 我国智算中心相关政策演进阶段

二、智算中心发展现状

（一）智算架构不断丰富，评价体系由硬向软演进

在传统数据中心中，业界更多关注底层的风火水电等硬件设施。随着人工智能技术的不断发展，智算中心的概念逐渐明晰，其架构也在不断完善和进化，从关注硬件扩展到更加注重软件与硬件的协同设计与优化。工信部等十一部门在 9 月发布的《关于推动新型信息基础设施协调发展有关事项的通知》将智算中心定义为基于人工智能理论，采用人工智能计算架构，提供人工智能应用所需算力服务、数据服务和算法服务的一类算力基础设施。在智算中心发展阶段，供电、制冷等底层设施仍是关注的基础，同时业界的焦点逐步拓展至软件层和模型层。

在硬件层，从过去单纯地关注算、存、运单个系统的运行转向 AI 计算子系统、存储子系统、网络互连子系统的协同建设；在软件层，

除了操作系统、数据库、中间件等底层软件，AI 开发框架和软件加速库进入研究视野。随着 GPT-3 的出世，大模型的发展也是日新月异，大模型演进路径从通用模型（L0）演进至行业模型（L1）及垂直领域（L2）模型。智算中心展现出不同于传统数据中心的新变化，并且行业发展尚处于百家争鸣的阶段，这不仅要求评价体系从硬件向软件演进，同时业界期望通过标准化体系去判断各种产品或框架发展水平的高低。图 3 描绘了智算中心的总体架构，在关注硬件基础设施的传统评价体系的基础上增加了智算中心新框架的描述。

来源：中国信息通信研究院

图 3 智算中心总体架构

（二）算力结构不断优化，智能算力规模逐步扩大

算力结构上，智能算力需求日益增长，我国加快智算布局，智能算力的比例逐步提高。算力正从单一向多元化、智能化方向全面优化与演进，体现了信息技术的创新与进步。随着人工智能技术的发展，智能算力占比显著提高，并呈现出稳定的增长趋势。截至 2023 年底，我国智能算力规模达到 70EFLOPS，增速超过 70%，智能算力占算力

总规模比重超过 30%。算力结构优化不仅提升数据处理的速度与效率，还增强计算系统的灵活性和可扩展性，为各行业的数字化转型和智能化升级奠定了坚实基础。

算力规模上，智算中心建设经历了从百卡到十万卡的阶梯式发展。在智算中心发展的初期，市县级智算中心以百卡规模起步，地方政府秉持“小步快跑，不断尝试”的原则，积极推动百卡集群小规模智算中心的落地，如南京、武汉等地率先探索几十 P 至百 P 规模满足数字政务需求。千卡集群主要分布在省会城市的智算中心，运营商出租算力和大型央国企自用算力，例如工行、招行、深交所等建设千卡集群算力规模在百 P 到千 P 之间。随着调度技术的成熟和 AI 技术的广泛应用，智算中心步入万卡及十万卡集群，主要集中在 AI 企业和运营商，用于企业数字化转型自用、大模型和服务出租。OpenAI 和微软联已建成 10 万卡集群、Meta 发布了 1.6 万卡、2.4 万卡集群，特斯拉/xAI 在 2024 年 7 月搭建了壮观的 10 万卡的超级集群。国内企业紧随其后，腾讯、阿里也发布了超万卡集群。字节跳动搭建了一个 12288 卡集群，研发 MegaScale 生产系统用于训练大语言模型。科大讯飞在 2023 年建成了首个昇腾万卡算力平台“飞星一号”。

（三）国家引导布局优化，持续扶持智算中心发展

在布局方面，国家不断出台相关政策对智算中心在内的算力基础设施进行优化部署。2021 年，《新型数据中心发展三年行动计划（2021-2023 年）》发布，引导新型数据中心集约化、高密化、智能化建设，推动形成数据中心梯次布局。2023 年 10 月，《算力基础设施高质量

发展行动计划》提出完善算力综合供给体系，优化算力设施建设布局，促进东西部高效互补和协同联动。同年 12 月，国家发改委等五部门发布《深入实施“东数西算”工程 加快构建全国一体化算力网的实施意见》，提出了包括构建全国一体化算力网、算力的一体化布局、东中西部算力的一体化协同等在内的重点工作部署，以推动新增算力向国家枢纽节点集聚。

（四）企业成为重要主体，积极推进智算中心建设

在政府的积极引导下，电信运营商勇担使命，建设部署取得积极成效。三大基础电信运营商将智算中心与算力网络建设作为发展方向，并融合技术、应用等优势，加快落实布局东西部算力基础设施建设。中国电信通过建设全国“2+3+7+X”公共智算资源池，特别是在京津冀、长三角地区建设的液冷单集群万卡智算池。中国移动则依托算力网络“4+N+31+X”资源布局体系，统筹规划“N+X”智算中心布局。中国联通则致力于打造面向算力供给的数智新底座，加速推进智算中心建设升级，在数据中心“5+4+31+X”基础上打造“1+N+X”智算集群。三大运营商均致力于通过技术创新和服务升级，推动算力成为像水电一样“一点接入、即取即用”的社会级服务，为数字经济发展注入新动能。

第三方数据中心服务商紧跟“东数西算”布局规划指引。秦淮数据积极融入“东数西算”国家级零碳工程示范项目，依托甘肃庆阳丰富的可再生清洁能源，布局建设零碳数据中心产业基地，服务京津冀、长三角、粤港澳大湾区等区域的算力需求；万国数据充分利用西部地

区“风、光、天然气”等优势资源，打造绿色智能数据中心；世纪互联逐渐向西部地区扩大业务布局，计划用地 200 亩于乌兰察布建设云计算中心，并提升绿色能源使用比例。

（五）通算智算齐头并进，应用场景日趋多元丰富

通算与智算两种形态并存，共同构筑了现代计算的新格局。在数智化转型和智算中心建设的浪潮中，以智算为代表的算力规模稳步增长，但同时各类应用场景对不同算力的需求也日趋多样化和复杂化。2023 年 10 月，工信部等六部门印发的《算力基础设施高质量发展行动计划》强调了多元供给和优化布局的重要性。随着智算中心建设的不断深化，建设方逐渐认识到能力和目标单一的智算中心区域性和行业性依赖强，难以应对复杂多样的应用场景，无法充分发挥智算中心价值。

在此背景下，“一中心双引擎”（智算中心同时提供智算和通算资源）、超算中心的 AI 升级改造、云计算中心标配智算云服务等新建设场景不断涌现。新场景呈现出如下特点：**互补性**，通算、智算和超算合理配比并有机组合，使计算中心能够提供更灵活、全面和高效的算力服务；**协同性**，在某些复杂应用中，将通算、智算和超算的能力结合起来，以实现最优的性能和效率；**灵活性**，算力中心能够根据多样化的算力需求变化动态调整、灵活调度资源分配，甚至定制需求的算力资源；**可扩展性**，算力中心在特定算力不足的情况下，能够快速实施工程扩容或线上动态扩容，保证充足的算力资源。智算中心的多样化算力发展是响应不断变化的算力需求和经济挑战的必然结果，

有利于智算中心投资规划的理性化回归，有利于加速智算中心向大型化、智能化、绿色化方向的发展，有助于数字经济和社会的持续发展。

三、智算中心发展挑战

（一）智算规模持续扩大，倒逼底层技术加速变革

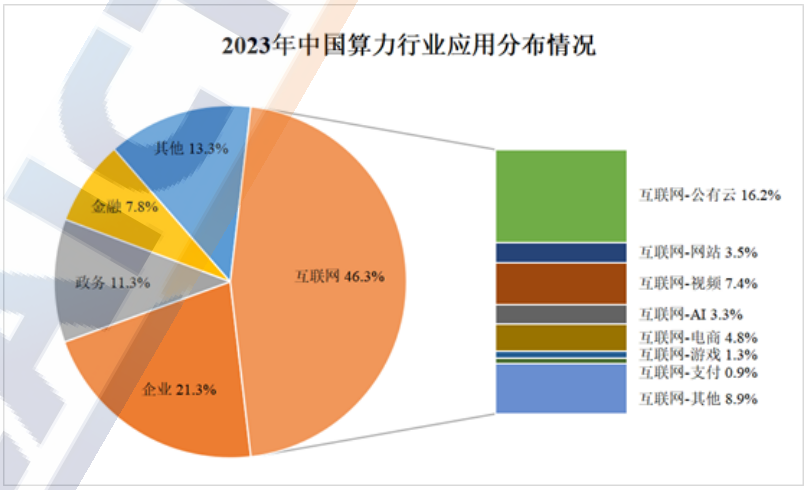
AI 应用场景对冷却的要求较高，风冷难以消解局部热点。大规模训练对于计算资源、存储能力及散热效率的要求尤为苛刻。一方面，在进行深度学习模型、大数据分析等复杂计算任务时，高强度的数据处理和运算会导致硬件设备产生大量热量，高温环境将严重影响硬件的性能稳定性和使用寿命，甚至可能引发系统故障。另一方面，智算中心业务流量峰谷波动显著，业务高峰期服务器集群需要满负荷甚至超负荷运行，局部热点问题突出。风冷技术受限于空气的热传导效率及风流组织的均匀性，难以快速且精准地应对服务器内部复杂的热量分布，局部热点难以有效消除。此外，风冷系统对于环境温度变化的响应速度相对较慢，难以实时匹配业务负载变化带来的热量波动，这在对温度敏感的高性能计算任务中尤为明显。液冷技术可利用液体高导热性，实现热量的快速吸收和转移，从而更有效地消除局部热点。

智算中心功率密度的不断提升，供电架构日益复杂。与传统数据中心相比，一方面，AI 芯片运行功耗峰谷特性明显，算力需求高时功耗达最高设计值，低算力需求时功耗较低。且大模型训练时间长，工作负载可以在峰值功率下，运行数小时、数天甚至数周。另一方面，智算中心对业务连续性要求高，供电系统平稳运行仍直接关系到智算中心核心功能的实时响应和执行效率。弹性供电系统可采用大容量、

模块化的高效不间断电源，形成电力资源池，配备储能系统，通过释放存储的能量来管理电力需求高峰，实现扛峰增载。

（二）算力应用门槛较高，普适普惠水平有待提高

算力应用以互联网为主，推动算力全行业普及应用尚存空间。据中国算力发展报告（2024 年）数据显示，截至 2023 年底，我国算力行业应用主要分布在互联网、企业、政务、金融等行业，占比分别为 46.3%、21.3%、11.3%、7.8%，互联网占比持续上升，政务占比进一步下降。其中互联网主要可细分为公有云、网站、视频、AI、电商、游戏、支付等领域，占比分别为 16.2%、3.5%、7.4%、3.3%、4.8%、1.3%、0.9%。算力应用正从互联网、电子政务等传统领域，向服务、电信、金融、制造、教育等多个行业拓展。随着智能算力在更多行业的应用前景不断显现，智能驾驶、影视渲染等典型领域有望充分发挥智能算力在提升效率与决策能力中的优势，应重点关注典型行业智能化转型需求，以点带面，助力全行业实现智能化升级。



来源：中国信息通信研究院

图 4 2023 年我国算力行业应用分布情况

中小型企业亟需成本优化以促进商用算力的深度应用与发展。一方面，智算中心的前期建设和后期运维成本高，还配备了高价值的 AI 服务器、高性能芯片和液冷系统等，总拥有成本居高不下。高成本通过价格传递机制导致智能算力租赁市场价格高。另一方面，尽管我国中小型企业有智能化升级的需求，但是目前追求降本增效更为迫切。而且大部分中小型企业处于行业价值链中低端，普遍存在专业人才缺乏、营收较低、抗风险能力弱等特点。对于中小企业而言，投入高算力成本进行智能化升级，难以带来足够收益，达到预期效果。

（三）智算业务灵活部署，算存运能力需全面增强

1. 智算深挖芯片潜力，算力调度与管理待优化

集群扩展对模型利用率（MFU）指标带来挑战。MFU 描述了在给定集群规模条件下，模型训练时有效利用计算资源的性能指标。一般来说，随着集群规模的扩大，MFU 是呈现次线性的。在超大集群中，通信带宽不平衡，随着集群扩大无法掩盖的集合通讯占比会增加，导致 MFU 逐渐降低。同时集群规模变大后，其稳定性、可用性降低会导致 MFU 下降。当前，以 GPU/NPU 为代表的通用加速芯片不断更新架构工艺、持续升级性能，同时专用加速芯片仍在不断发展。

算力资源的全局调度和高效管理有助于提高算力利用率。从集中式计算系统到分布式计算系统，算力调用方式经历了从固定资源到动态资源、从本地到云端的转变。这一演进不仅提高了计算资源的利用效率和任务性能，还为企业和组织提供了更加灵活、可扩展的计算服务解决方案。通过虚拟化技术，可以将物理资源转化为虚拟资源，实

现资源的动态分配和灵活调度。随着人工智能技术的不断发展，混部技术和 AI 弹性容量的智能化程度将不断提升。例如，中国移动智算中心（青岛）通过引入自研智算平台和先进算法，优化算力调度，可以更准确地预测应用负载和资源需求，从而实现超大规模训练场景下的精细化资源管理，有效提升计算效率，荣获算力性能 4A 等级认证。

2. AI 大模型算力需求大，存储能力需同步升级

全闪存储、分布式存储、冷热数据分离存储等技术，推动智算中心的存储向高效化和智能化发展。随着数据类型由单模态向多模态和全模态转变，数据量爆发增长。同时，大模型的训练过程需要随机读取海量小文件，以及快速保存模型数据集。高频词的读写使数据存储系统必须提供高达 100M IOPS 的读写能力和上百 GB/s 的带宽。过去广泛采用的共享存储搭配本地 SSD 盘的存储架构，因受限于容量不足、易受计算节点波动影响以及缺乏容灾备份机制等缺陷，已难以适应当前大模型发展的需求。而全闪分布式存储可扩展至上百节点，单集群存储容量可达数百 PB 以上，单个存储节点能达到数百万至上千万 IOPS，10-20GB/s 带宽，一般 10-20 个全闪存储节点即可满足 AI 大模型下的性能要求，同时提供完善的数据保护机制和安全防护措施，实现数据在多个计算节点间的共享访问，且管理运维简便高效。

3. 网络性能需求提升，技术创新刻不容缓

智算集群中计算节点的海量数据传输亟需高性能、超低延迟且支持无损传输的网络互联技术。为处理大规模智算集群带来的海量计算任务，多维度并行被广泛应用，包括数据并行、流水线并行、张量并

行和专家并行。通过多个 GPU/NPU 节点构建超大规模的计算集群，这几种并行方式将数据或者模型切分到不同 GPU/NPU 训练，并行数据需要在各个节点之间高速传输，以确保计算任务的高效完成。张量并行通信量是流水线并行和数据并行的 50 倍以上，业界通常采用机内定制的高速总线技术承载。流水线并行和数据并行需要跨多节点通信，通过超宽无损的网络提供超大的带宽和超快的数据传输速度，从而确保数据在各个节点之间流畅地传输。

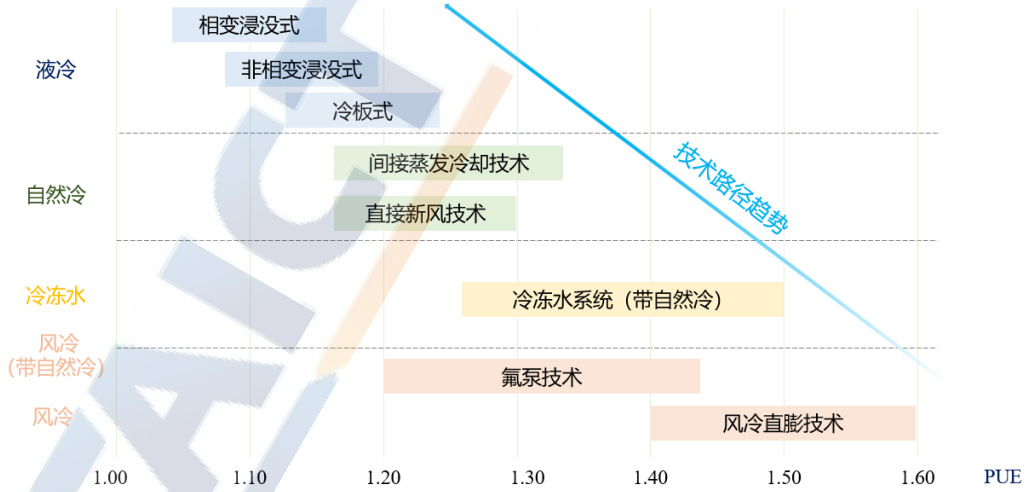
大规模 AI 计算要求有效利用和高效分配网络资源。传统网络的资源分配不均衡可能导致部分节点过载，从而降低整体网络性能。随着 AI 大模型应用的普及，多节点协同进行模型训练的频率和数据计算规模将显著增加，集群规模扩展使得网络资源负载不均的问题变得更加突出。高利用率的网络能更好地管理和分配资源，确保每个节点充分利用，从而提高整体计算效率。提升网络的资源感知能力将有助于更好地分配计算和网络资源，实现网络级负载均衡，提高整个集群的计算训练效率，从而处理更多的计算业务，减少资源浪费和成本。

（四）AI 服务器功率骤升，绿色低碳发展面临挑战

芯片 TDP 不断攀升和集群部署方式导致总功耗不断增加，能耗成为智算中心迫在眉睫的问题。一方面，AI 计算任务的复杂性和数据量的爆炸性增长，要求 AI 芯片具备更强的处理能力和更高的运算效率。AI 芯片设计不断向更高集成度、更多核心数、更高频率的方向发展，提升计算能力的同时也带来了功耗的显著增加。另一方面，为了降低网络时延，智算服务器需要以集群的方式进行部署。将多台服

务器连接在一起，形成一个统一的计算平台，从而大大提高数据处理的速度和效率。据 Digital Information World 发布的报告，智算中心为训练 AI 模型产生的能耗将为常规云工作的 3 倍，预计到 2030 年，智算中心的电力需求将以每年 10%的速度增长，而这对双碳背景下的智算中心能效提升带来了巨大的挑战。

液冷技术作为智算中心的高效制冷方案，可显著提升散热效率并降低电能使用效率（PUE）。采用风冷直膨散热冷却方式的数据中心 PUE 一般在 1.5 **左右。**因液体的热导率较气体可提高一个数量级，目前全球高密集度、高供电功率的超大型数据中心已逐渐引入液冷设备。自然冷却也是一种具有巨大潜力的节能技术，能适应不同气候条件和地区需求。例如，在南方炎热地区，可采用高温大温差并联冷水机组的方式来降低数据中心的温度。在北方寒冷地区，可以采用直接空气自然冷却的方式，提高能源利用效率。



来源：中国信息通信研究院

图 5 制冷技术 PUE 发展趋势

（五）软硬件一体化融合，智算中心追求提质增效

人工智能服务场景中，快速部署成为关键需求，智算中心面临的建设和交付挑战日益增加。用户对交付时间的要求不断缩短，复杂的组网与设备调试进一步加大了项目实施的难度，集群系统的性能、能效、可靠性、安全性等各方面都提出了更高要求。此外，如图 7 所示，智算中心涉及的 L1 层（基础物理设施层）和 L2 层（网络、存储及虚拟化层）也面临严峻挑战，尤其是 L2 层的集成需求变得日益复杂，需要更高效的预制化建设模式。当前，集成过程的工具化不足，缺乏统一的高效集成工具平台，导致项目从规划到实施再到上线的周期被大大拉长，集成效率显著降低，这不仅增加了企业的运营成本，也限制了 AI 服务快速响应市场变化的能力。

来源：中国信息通信研究院

图 6 基础设施&物理资源管理范畴

软硬一体化融合架构具备多方面技术优势，将提升智算中心服务能力的质量和效率。在硬件层面，按照计算、存储、网络等资源类别的差异，整合硬件资源，形成同类资源池，实现 CPU（Central Processing Unit，中央处理器）、GPU（Graphics Processing Unit，图形处理器）、NPU(Neural Processing Unit)、FPGA（Field-Programmable Gate Array，

现场可编程门列阵）、ASIC（Application-Specific Integrated Circuit，专用集成电路）等多种异构算力的按需重组，能够满足不同场景中的应用需求。**在软件层面**，推进硬件资源自适应重构，实现资源动态调整、灵活组合和智能分配，响应多应用、多场景需求。软硬件融合架构发挥资源管理和调度系统的应用感知能力，建立起智能化融合架构，使软件层面的全部资源在可调度的范围内实现动态组合，能够满足多种应用场景的智能化需求。

（六）建设经营多元发展，统一评价体系有待构建

智算中心建设多元发展，评价体系碎片化，亟需构建统一标准，促进技术创新与产业升级健康发展。在当今数字化转型的大潮中，智算中心作为支撑人工智能、大数据等前沿技术的关键基础设施，其建设与经营正呈现出百家争鸣的繁荣景象。各地政府、企业及科研机构纷纷投入资源，探索符合自身需求的智算中心建设路径，推动技术创新与产业升级。在制冷方面，液冷包含了冷板式液冷和浸没式液冷等多种方案，冷却液介质也存在氟化液和硅基油等多种选择。在供电方面，企业也对供电冗余提出了多种技术路线，根据实际情况采用 N+X 冗余、2N 或者其他的冗余供电方式。除了底层的基础设施，上层网络等随着节点数量的增加有各种组网方式。此外，出于成本或者是可用性的考量，智算中心的各个系统的布局也存在差异。

然而，这种多元化的发展模式也带来了评价体系的碎片化问题。不同主体在智算中心的建设标准、运营效率、技术创新能力等方面存在显著差异，缺乏统一、科学的评价体系来衡量其综合效能。这不仅

增加了市场比较的复杂性，也可能导致资源错配和重复建设，影响行业的健康可持续发展。构建一套统一、权威的智算中心评价体系可以引导行业健康发展，促进技术交流与合作，推动形成优势互补、协同发展的良好生态。

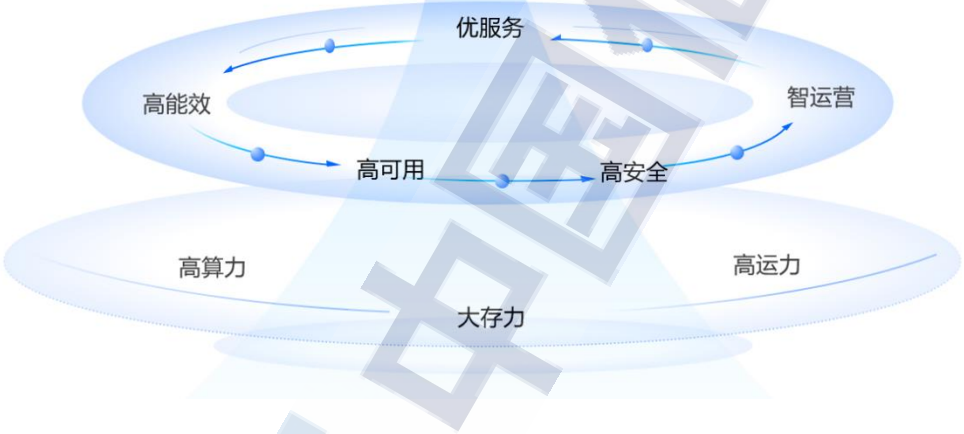
四、智算中心综合评价体系

（一）综合评价体系构建

目前我国算力评价主要可分为规格算力评价和算力综合评价两大类。前者主要关注硬件设备的计算性能，后者对算力系统进行多元的性能测试和分析。但是两者的共同点都是聚焦在硬件基础设施上，对上层软件的考量较少。然而，随着技术的快速演进和业务需求的复杂化，在应对人工智能应用场景，现有评价体系未能对软硬件设备提供全面的考量。在原有评价体系的基础上，本报告的综合评价体系不仅涵盖了智算中心软硬件总体架构，还综合考量了智算中心的特征。

与传统数据中心相比，智算中心具有高算力、大存力、高运力、高安全、高可用、高能效、智运营、优服务等特征。高算力、大存力、高运力构成了智算中心的算力底座。智算中心集成前沿的技术元素，如人工智能算法、存算分离、大数据分析及高速网络通信等，还配备高性能的计算硬件，应对大数据分析、深度学习、图像处理等复杂多变的计算需求。同时，智算中心融入了绿色节能理念，通过采用先进的能效管理系统和节能设备，实现了计算资源的高效利用与能源消耗的显著降低，展现出了高能效的特点。此外，在数据安全与隐私保护方面，智算中心构建了多层次的安全防护体系，包括数据加密、访问

控制、安全审计等，全方位保障数据资产的安全性与用户隐私，体现了其高安全性的显著优势。在业务连续性方面，智算中心通过全面的冗余设计、自动化的故障恢复机制等措施确保服务的高度可用性，即使在面对突发故障时也能迅速恢复。通过深度整合从底层硬件资源到上层应用软件的全栈技术能力并辅以先进的运营理念，智算中心能够全方位地为用户赋能，提供高度定制化、灵活可调的优质服务体验。



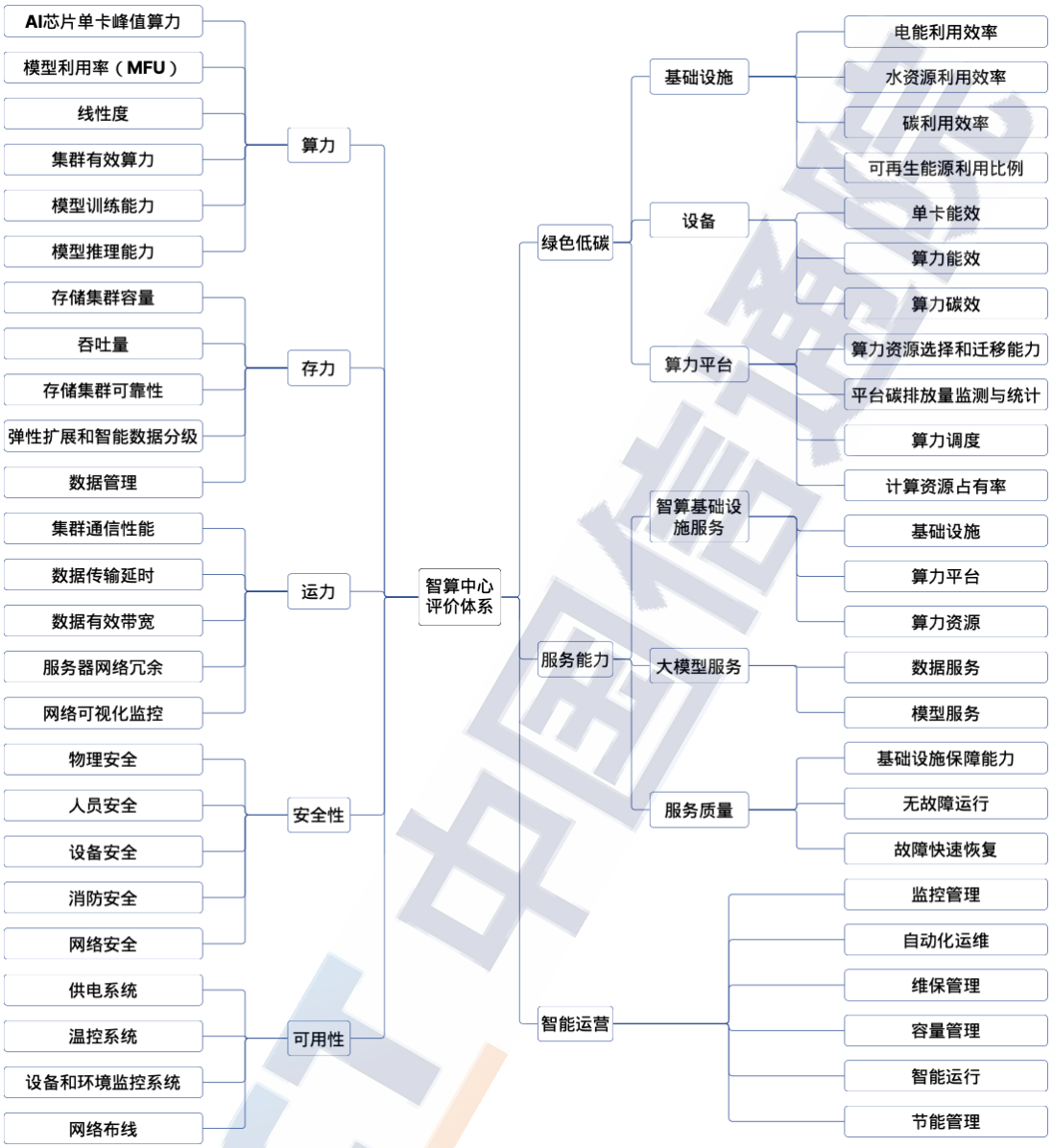
来源：中国信息通信研究院

图 7 智算基础设施特征

构建一个全面、科学、前瞻性的智算中心综合评价体系不仅是对智算中心现状的一次全面审视，更是对其未来发展潜力与方向的一次深刻洞察。对应“5+3+1”特征[2]，报告从算存运能力、安全可用性、绿色低碳、智能运营以及服务能力等多个维度出发，甄选了具有代表性的关键指标，旨在从多层次对智算中心进行综合评价，甄别智算中心的优势与不足，为其后续的优化升级提供明确方向；还能够促进智

[2] 1 是指风火水电，3 是指高算力、大存力、高运力，5 是高安全、高可用、高能效、智运营、优服务。

算中心之间的良性竞争与合作，推动整个行业的健康发展。



来源：中国信息通信研究院

图 8 智算中心综合评价体系

（二）算力

算力是衡量智算中心处理能力的核心指标，直接关系到数据处理的速度与效率。算力是支撑“人工智能与大数据”产业蓬勃发展的重要“底座”，也是驱动经济数字化转型的新引擎。算力水平对智算中心整体服务水平起着决定性的作用。理论算力，即 AI 芯片的各类性能参

数的标称值，奠定了性能上限。但在实际的运行中，算力发挥不仅取决于芯片本身的计算能力，还受到显存容量与带宽、互联技术以及系统架构设计等多方面因素的影响，往往发挥不出全部的算力性能，有效算力低于理论算力。对业务模型场景的支持能力也是考验智算中心的重要标准，体现了智算中心适应不同应用需求、快速响应市场变化的能力。这不仅要求硬件平台具备广泛的兼容性，能够支持多种框架和算法的运行，还要求软件生态能够提供丰富的算子库、预训练模型及工具链，以便用户能够快速部署和优化自己的模型。以商汤科技人工智能计算中心为例，该中心提供大规模弹性算力，支持超大参数的大模型训练，旨在满足上海和长三角地区对低延迟、高效能 AI 服务的需求。创新的低时延网络设计和 RDMA 高速通信网络，进一步提升了训练和推理的效率，推理服务的性价比提升了 3 倍，展现了较优性价比的 AI 服务效果，为智能制造等多个行业提供了坚实支持。此外，中国联通上海临港智算中心配备 1.5 万架机架，是联通“1+N+X”高等级算力集群的核心枢纽节点，基于统一联通云底座构建多卡并行、多元共生、训推一体的智算集群，实现了万卡算力供给，荣获智算中心算力性能 5A 等级认证。

1.AI 芯片单卡峰值算力

AI 芯片的单卡峰值算力是衡量其性能的关键指标，它决定了芯片在处理人工智能任务时的最大计算能力。智算中心往往运行计算量大、数据海量密集的人工智能任务。而支撑任务运行的算力，最重要的组成部分是 AI 芯片，峰值算力越高，表示芯片理论上能更快地完

成复杂的计算任务，如更快的响应实时应用的处理，尤其是在 AI 训练和推理过程中需要处理的大量数据和并行运算。

2.模型利用率（MFU）

模型利用率指模型一次前反向计算消耗的矩阵算力与机器理论算力的比值，反映 AI 芯片的规划、管理与使用情况。高模型利用率意味着更高效的资源使用，减少对额外硬件的需求。模型利用率可以反映出整体算力利用效率。

3.线性度

线性度是衡量一个系统或模型输出与输入之间线性相关程度的指标。它表示系统在一定范围内，输出与输入之间的比例关系保持稳定的程度。线性度好的系统，其输出能够较为准确地反映输入的变化，呈现出较为明显的线性特征；而线性度差的系统，输出与输入之间可能存在较大的偏差或非线性关系。在智算场景中，线性度为单卡训练扩展到多卡，单机拓展到集群的效率度量指标。线性度的取值范围为 0~1，数值越接近于 1，其性能指标越好。

4.集群有效算力

智算中心通过集群方式对外提供服务。集群有效算力是指智算集群实际能提供的最大算力和理论最大算力的比值，表征智算中心的实际算力表现。在一个由多个计算节点组成的 AI 集群中，实际可用于执行人工智能任务（如模型训练、推理等）的计算能力的总量不仅取决于单卡峰值算力，还依赖于整个集群的网络配置、规模和算力利用率。有效算力更能反映集群在实际工作负载下的性能。

5.模型训练能力

智算中心对多元化训练场景的高效支持能力，是衡量其算力适应性广度的关键指标。由于人工调参的差异，AI 芯片适用的业务场景有偏好。智算集群应满足多种模型在各个应用场景的训练以及配套性能，比如，应能够支持计算机视觉、语音识别、机器翻译、推荐算法、大模型等应用场景下的代表性模型训练与数据集处理。

6.模型推理能力

推理也是智算中心的关键应用领域。从场景看，智算中心应能够实现处理计算机视觉、语音识别、机器翻译、推荐算法等常见模型的推理任务，支持包括文本、图像、声音等多模态数据的处理以及跨模态推理任务。通过实际应用场景的效果进行验证，并确保其能够在真实世界问题中提供有效决策支持多模态推理；此外，在从新信息中学习和适应的能力方面，应能够展示出在少量样本或零样本学习情景下的推理的灵活性和适应性。

（三）存力

存力关注的是智算中心的数据存储与访问能力，是数据持久化与高效利用的重要支柱。党的十九届四中全会首次提出将数据作为生产要素参与分配，数据已成为重要生产要素和资产。伴随着计算走向异构化、复杂化，AI 模型走向巨量化，存储层面面临着海量数据的采集、存储、传输、管理等一系列挑战与问题，存储集群的容量需足够庞大以支撑海量数据存储。此外，在数据成为核心资产的背景下，安全可靠性成为了不可忽视的基石。在呼吁绿色低碳的当下，存储设备

的能耗不可小觑。为应对这些挑战，业界积极拥抱技术创新，将存算分离、弹性扩展、冷热数据分治等先进设计理念融入存储系统中，以提升资源利用效率。“大存力”的实现不仅是容量上的突破，更是可靠存储、先进技术与绿色理念的深度融合。以科大讯飞 AI 先进存储中心为例，其通过创新存储技术，采用分级管理、多协议无损互通等先进设计，成功构建了高效、稳定的 AI 大模型训练平台，解决了存储性能与容量的平衡问题，有效提升了数据处理效率与可靠性。

1.存储集群容量

存储容量是支撑智算中心数据密集型智能计算任务的数据底座，指存储集群系统容量总和。智算中心所存储的海量数据，包括原始数据、训练数据、Checkpoint 数据、中间结果数据以及模型文件等。而足够的存储容量可以确保数据的完整保存，满足大规模模型训练以及对外进行推理服务的需求。OpenAI o1 和 Sora 的出现标志着 AI 大模型从 NLP 走向多模态，所需数据量也快速增长。一个 NLP 大模型训练所需数据集在 50TB 以上，原始数据约是数据集的 50 倍，存储集群系统容量在 PB 级；多模态、万亿参数大模型所需数据集急速增长到数百 PB，原始数据约是数据集的十倍以上，存储集群系统容量需求提升到 EB 级。

2.吞吐量

存储的吞吐性能作为衡量智算中心存储集群设备性能的重要标准，影响着智算中心运行的整体效率，因智算中心常需要进行人工智能模型的训练和推理，需要快速访问和处理大量数据，高吞吐率的存

储系统能够提升数据传输速度，加速模型训练和推理过程。高性能存储集群系统可以减少 AI 芯片的等待时间，提高整体计算效率。在处理和存储海量数据的过程当中，高吞吐量的存储设备可以快速读写大量数据，满足大数据应用的需求。千卡、万卡集群逐步向十万卡集群推进，数据集加载和 Checkpoint 读写这两个过程与存储子系统紧密关联，已经成为影响智算集群能力的关键。数据集加载过程和推理过程以海量小文件访问为主，IOPS 性能密度需要达到千万 IOPS/PB 存储容量，以最小加载时长减少 AI 芯片等待时间。Checkpoint 读写过程是带宽型场景，带宽性能密度需要达到 TB 级带宽/PB 存储容量，将集群故障恢复耗时缩至最短。

3.存储集群可靠性

存储集群系统应提供不低于 6 个 9 的高可用性。数据存储的稳定可靠，直接决定了智算集群的高可用性。原始数据、训练过程中产生的 Checkpoint 数据、以及输出的训练模型等都会被保存到数据存储系统上，如果出现丢失或者损坏，其损失不言而喻。6 个 9 高可用性存储确保集群服务持续在线，同时多种跨智算中心的高可用方案需将跨域的高可用性提升至 7 个 9 以上，在遇到供电中断等不可抗力因素时，存储集群系统应具备数据灾备和恢复能力，并实施定期的数据灾备和有效的恢复策略，以防数据丢失。如通过同步、异步复制，支持小时级、分钟级的 RPO（Recovery Point Objective）、RTO（Recovery Time Objective），以达到灾备数据恢复的效果，有效的保障数据的安全可靠性。应支持软件可信、数据加密、防勒索病毒、安全销毁等存

储安全技术。

4.弹性扩展和智能数据分级

存储资源弹性扩展是指存储集群系统能够根据用户的实际需求，自动调整存储资源的分配，包括增加或减少存储容量，以适应业务的发展变化。随着智算中心存储数据量的增长，存储系统需要具备良好的扩展性，高吞吐量的存储设备有助于平滑地扩展存储容量并支持在线弹性扩展。存储集群系统通过增加节点扩展集群规模，灵活地满足存储容量增长的需求，应对 AI 应用大规模数据存储的挑战。存储集群系统应支持冷热数据自动分级，模型训练推理过程中热数据被频繁访问，而温冷数据如早期保存的 Checkpoint 和历史模型参数则很少被访问。数据存储自动分级允许在一个存储池内使用不同类型的存储介质划分不同的硬盘池，可以灵活的兼容文件、对象等多种协议。通过对不同价值数据的自动搬迁，将冷热数据存放到各自合适的存储空间中，性能与成本实现更好的平衡。

5.数据管理

存储集群系统提供数据编织能力，基于存储元数据管理及检索能力，通过全局数据视图技术，实现全局数据可视可管，大幅提升数据流动效率，达成业务无感、业务性能无损的数据最优排布，满足来自多个源头的价值数据快速归集和流动，以提升海量复杂数据的管理效率，直接减少 AI 训练端到端周期。提供智能检索引擎和 RAG 知识库能力，支持张量、向量等多维数据快速检索，加快大模型推理效率。

（四）运力

运力关注的是智算中心的网络连接与数据传输能力，是数据流动与资源共享的基石。作为构建智算算力服务的重要一环，网络运力是以数据通信网络基础设施为基础，以自动化、智能化网络技术等为支撑，实现数据在不同用户、算力设施间以及算力设施内高效流动的网络运载力。智算中心通常涉及到大量的数据传输，如人工智能（AI）、机器学习（ML）和高性能计算（HPC）应用，这些应用需要快速、高效的网络来处理和分析大量数据。智算中心运力应具备高速率的集群间通信能力、较低传输时延以及可靠性保障。为满足北京高密度人工智能企业的需求，北京电信构建了以数据中心为核心的全光互联网络，实现了数据中心间的超高速互联和极低时延。通过 800G 波分技术和全局负载均衡算法，该网络在 0.5-2ms 内实现了京津冀区域算力的无损互联。该创新网络不仅降低了能耗，还推动了算力资源的无差异高效辐射，为区域数字经济一体化和智能化转型提供了强大支持。

1.集群通信性能

智算中心的集群通信性能是评估算力与存力设备在集群环境中运行效率与效果的关键因素。为实现高效网络通信，数据中心常采用 all_reduce、all_gather、all_to_all、broadcast 及 pt2pt 等多种通信算法，以优化数据传输路径与效率。在全面衡量集群整体通信性能时，通常综合考虑数据规模、操作耗时、吞吐以及带宽，这些指标共同构成了评估集群通信性能与效率的综合体系。

2.数据传输时延

数据传输的网络时延影响着智算中心整体的应用效率。自动驾驶、虚拟现实、增强现实等人工智能应用，对响应时间较为敏感。因此，智算中心网络的设计和优化通常会考虑如何降低时延，提高数据传输的实时性和可靠性。对于实时性要求极高的应用来说，同交换机下点到点单向时延＜3.5us，三跳网络下点到点单向时延＜6.5us 是较为卓越的表现。

3.数据有效带宽

有效带宽是指在实际应用中，网络链路能够持续稳定提供的最大数据传输速率。它受到多种因素的影响，如网络拥塞、设备性能、传输协议等。RDMA（Remote Direct Memory Access，远程直接内存访问）技术支持高速、低延迟的数据传输，为大数据处理和分布式计算提供了强有力的支持。在某些特定场景下，高性能 RDMA 网络数据传输的有效带宽应达到或超过 90%的利用率，以确保数据传输的高效性和稳定性，为大规模并行计算任务提供坚实保障。

4.服务器网络冗余

智算中心需构建多路径网络连接，确保网络故障不会导致全系统的瘫痪。服务器网络的冗余可确保智算中心在运力侧实现高可用，降低故障对智算中心任务运行的实际影响。智算中心可通过服务器双物理端口，以多活方式接入 RDMA 网络，降低单 TOR 交换机故障影响，提升 RDMA 网络整体可靠性。

5.网络可视化监控

通过对智算中心网络运行情况的监控，及时跟踪排查网络丢包、

拥塞等相关故障，提升对内部算网的感知，实现高效运力运维。智能计算中心的网络分钟级、秒级的可视化监控，包括 RoCEv2 和 IB，网络吞吐、丢包、拥塞等监控，以及集群监控大盘，事件大盘，告警大盘等网络性能可视化能力，可帮助智算中心运维人员快速定位故障，及时预警，做到网络全流程的可视化高效运维管理。

（五）安全性

安全是智算中心运行的生命线。智算中心作为各个行业信息系统运行的物理载体，已经成为经济社会运行不可或缺的关键基础设施，应以发展与安全并重为原则，进一步强化安全管理和能力建设，构建完善的安全保障体系。具体来看，安全性应聚焦物理安全、人员安全、设备安全、消防安全、网络安全等五方面。物理安全确保了智算中心免受自然灾害、盗窃及非法侵入的威胁，保障了关键资产的物理完整性；人员安全则通过专业的培训与严格的访问控制，降低了人为因素导致的数据泄露或设备损坏风险；设备安全将防止因设备故障或老化导致的服务中断，确保智算中心能持续高效地提供服务；消防安全作为应急响应的重要组成部分，能够有效预防火灾事故的发生，并在火灾发生时迅速控制火势，减少损失；而网络安全则是智算中心面对外部威胁的第一道防线，通过加密技术、防火墙等手段保护数据传输安全，防范网络攻击，确保智算中心系统的稳定性和数据的机密性。例如，有孚网络的北京永丰云计算数据中心，采用了高性能的 UPS 和冷却设备，并实施了全面的物理安全措施，包括监控和门禁系统。这些措施有效保障了数据和设备的安全。此外，该中心还引入了流量清

洗和漏洞扫描等技术，确保数据安全性，为整个行业提供了良好的安全管理范例。

1.物理安全

高业务连续性要求智算中心故障发生率低，故障恢复时间快。为了尽可能地减少故障，需要对智算中心的硬件设备进行 24 小时全天候的无盲点监控，将关键硬件的运行状态以可视化的方式呈现给运维团队，使运维人员能够直观了解系统整体运行状况，确保设施设备的物理安全。对于智算中心的监控应在建筑各出入口、楼层出入口、运营机房内全面实现无盲点，全面关注各设备和环境状况。

2.人员安全

智算中心是涉及供电、制冷等系统的综合性建筑，对运维管理的专业性要求较高。人员配置上应覆盖电气、暖通、弱电专业，且应具备基本专业能力，维护人员获得国家及相关机构认可的电气及暖通职业资格证书。

3.设备安全

为了确保设施设备的正常运行与高效运维，运维装备配置比例维持在总运维人数的 10%或以上。这一策略不仅确保充足的技术资源和工具支持，以应对各种突发故障和日常维护需求，还能迅速响应问题，实现高效解决，从而有效维护智算中心整体运营环境的安全性。

4.消防安全

火灾自动报警系统通过探测器实时监测智算中心内的烟雾、温度

等火灾征兆，一旦达到设定阈值，立即发出报警信号，实现火灾的早期发现和预警。服务器对环境温度有严格要求，遇火或遇到与火伴随产生的热量、蒸气和烟雾时，特别容易损坏，且受损程度随温度上升而迅速提高。配置火灾自动报警系统可以及时发现并处理潜在的火灾隐患，降低火灾发生的概率和危害程度。

5.网络安全

智算中心承载着对于服务时效更为敏感的 AI 应用任务，需要实现网络端到端的安全灾备保护数据安全。安全域划分是实现网络安全的重要手段，可以将具有相同安全保护需求且互相信任的系统组成一个独立的区域，从而有效隔离不同安全域之间的风险，还可实施更加精细的数据访问控制策略，确保敏感数据仅在授权范围内流通，防止数据泄露或被非法访问。

（六）可用性

智算中心提供服务的首要条件是各系统具备可用性，能够支撑业务的运行。智算中心的建设涉及供电系统、温控系统、设备和环境监控系统、网络布线等多个系统的协调配合，对系统的可用性、稳定性和容错性有着极为严格的要求。不同于安全性，可用性表征在有一定容错或者并行维护的条件下的运行能力。系统冗余设计是保障智算中心高可用性的关键。通过多路径供电供冷、备份系统等冗余措施，可以在单点故障发生时迅速切换至备用资源，确保系统的持续稳定运行。中国雅安大数据产业园在供电系统设计上采用了双电源多回路环网供电的方式，确保在任何一个电源出现故障时，能够即时无缝切换到

备用电源，从而保证数据中心的持续运转，实现了高水平的可用性，极大降低了因单点故障带来的停机风险，为各类业务的连续性提供了有力保障。

1.供电系统

高可用的供电系统冗余要求为 2N。以市电为例，2N 冗余意味着有两条完全独立的电源线路为智算中心供电，即使其中一条线路出现故障，另一条线路也能继续供电，确保智算中心不会因电力中断而停机。对于一些关键业务系统和数据，企业和组织往往要求智算中心达到非常高的可用性水平。而市电进行 2N 冗余设计是实现高可用性的重要手段之一。通过提供两条完全独立的电源线路，可以最大限度地减少因电力中断而导致的停机时间，满足企业和组织对高可用性的需求。

2.温控系统

相比于传统数据中心，智算中心的变化之一是温控系统的冷却方式从风冷过渡到液冷。对于液冷设备而言，为了满足可用性的要求，同样需要对 CDU 换热单元作冗余设置以保证单条线路故障的情况下智算中心正常运行。2N 设置能够保证最高等级的可用性，但考虑到成本管控，企业也可采用 N+X 冗余配置，在单个设备故障的情况下仍能保障整体系统的正常运行。

3.设备和环境监控系统

设备和环境监控系统能随时采集各个设备的运行状态和健康状况，快速察觉故障点并做出反馈。监控系统的核心功能是实时采集和

传输视频、音频等数据，以及进行必要的控制操作。这些功能的实现都依赖于稳定的电力供应。由一路不间断电源+一路市电供电并保持末端监控设备的冗余能够保证监控系统的高可用性。

4.网络布线

智算中心涉及到大量的数据传输，网络的高可用性确保数据在传输过程中的连续性和完整性，减少因网络故障导致的服务中断和数据丢失。智算中心的网络包含园区至外部、园区内机房楼两部分。大部分智算中心具备 2 个及以上不重合的管道路由就可以满足可用性的基本要求。

（七）绿色低碳

在《算力基础设施高质量发展行动计划》中指出，坚持绿色低碳发展，全面提升算力设施能源利用效率和算力碳效水平。智算中心作为高能耗的基础设施，需从基础设施、设备到算力平台进行全方位的能效优化与碳排放管理。算力基础设施规模和复杂度日益增加，基础设施层的能耗问题逐步引发重视，推动了制冷、电力供应等领域的节能技术进步，早期关注点主要聚焦于基础设施层。随着对碳排放环节的识别，设备的能耗情况引起业界关注，尤其是服务器和存储设备的能效提升。当前，随着人工智能和大数据技术的广泛应用，智算中心的计算需求急剧增长，使得算力平台层的能效管理成为关键关注点。通过全链条的优化，智算中心可以实现真正的绿色低碳目标，支撑数字经济的可持续发展。蚂蚁消金自 2023 年起大规模投资绿色计算技术，通过智能运维、数据治理和业务优化，实现了节能减碳成效。根

据 GreenOps 的碳排放产品的计算，蚂蚁消金减少了 357 吨二氧化碳排放，且单笔交易的碳强度同比下降 58%。

1.基础设施

智算中心的基础设施层涵盖电力供应、制冷系统和建筑结构等系统，是支持其稳定运行和长期可持续发展的关键。基础设施层的能源使用效率与可再生能源的应用水平，直接决定了智算中心的碳排放强度，并在很大程度上影响了算力绿色低碳性的实现。同时，通过智能化手段加强能源和碳排放的管理，可以显著提升管理效率，确保基础设施在整个生命周期内的资源利用最大化和环境影响最小化。

（1）电能利用效率（PUE）

电能利用效率为智算中心总耗电量与智算中心 IT 设备耗电量的比值，一般用年均 PUE 值。统计除建筑办公设施外，智算中心 IT 设备、制冷设备、供配电系统和其它基础设施的用电量。PUE 越接近 1，电能使用效率越高。我国当前的部分智算中心不断强化绿色节能低碳技术应用，采用液冷等先进制冷技术，使得机房散热能耗降低 50%以上，PUE 值可降至 1.2。

（2）水资源利用效率（WUE）

从 PUE 到 xUE，能耗指标越来越丰富，业界意识到智算中心对水资源的消耗也很大。谷歌发布的 2023 年环境报告显示，AI 在 2022 年消耗了 56 亿加仑（约 212 亿升）的水，相当于 37 个高尔夫球场的水。水资源利用效率为智算中心总耗水量与智算中心 IT 设备耗电量的比值(单位：L/kWh)，一般用年均 WUE 值。WUE 数值越小，代表

智算中心利用水资源的效率越高。

（3）碳利用效率（CUE）

碳利用效率（CUE）是测量和计算智算中心碳利用效率的方法，为智算中心二氧化碳总排放当量(CO2eq)与 IT 设备负载能源使用量(通常以千瓦时为单位)的比值，单位是 kgCO2eq/kWh。需统计智算中心运行阶段，消耗电力、热力（蒸汽、热水）等能源所对应的二氧化碳等温室气体排放。IT 设备负载能源使用量为智算中心中 IT 设备耗电，包括智算中心中的计算、存储、网络等 IT 设备的耗电的总和。

（4）可再生能源利用比例

智算中心中的可再生能源电力耗电量与数据中心总耗电量的比值即为可再生能源利用比例。比率越接近 1.0，智算中心使用的可再生能源就越多，可再生能源利用比例越大，则表示该智算中心能源供给结构越优。

2.设备

算力设备层是数据处理和输出的基础，涵盖计算、存储和网络等 IT 设备。IT 设备尤其是服务器在智算中心的能耗与碳排放占比均很高，对关键设备的能耗监控与碳排放管理有助于推动智算中心整体的绿色节能。

（1）单卡能效

单卡能效是衡量智算中心 IT 设备层综合能效的关键指标，具体体现为单位用电量所产生的算力，即总能耗转化为算力的效率。随着芯片架构的不断突破，尽管单卡能耗不断提升，但算力能效也有明显改善。

（2）算力能效（CEE）

算力能效用综合算力与智算中心总耗电量的比值衡量。综合算力是指考虑了通用服务器、AI 服务器、数据存储以及网络交换实际使用比例与设备能力的乘积。算力能效的关注点从单个设备的优化逐步扩展到算存运一体的能耗优化。这一指标是对设备能耗情况的全面反映，尤其是通过归一化处理解决不同设备算力值量级不一致问题，实现各类设备能效的综合计算。计算公式为：CEE=C/E，其中 CEE 指的是智算中心综合算力能效值，表示单位电力所能转换的算力；C 是智算中心在一定时期内的综合算力总和，代表实际处理业务的能力；E 代表着测量期内智算中心单位时间的总耗电量（单位：千瓦时）。CEE 值越高，意味着智算中心在消耗相同电力的情况下，能够提供更高的算力输出，反映出更高的能效水平。

（3）算力碳效（CCUE）

算力碳效是衡量智算中心碳排放效率的关键指标，定义为智算中心综合算力与碳排放的比值，其公式为：CCUE＝C/CE，其中 C 表示智算中心在特定时间段内实际处理业务所产生的总算力，CE 则代表该期间内的总碳排放量，单位为 kg。CCUE 值越高，意味着在相同碳排放条件下，智算中心能够提供更强的计算能力，即在相同的碳排放量下能够处理更多的计算任务。

3.算力平台

算力平台层是智算中心在资源配置与管理中的关键环节，直接关系到计算效能与资源利用效率的提升。在推进绿色低碳发展中，平台

层应着力增强计算效用的监控评估能力，优化算力资源的全局调度和工作负载的消耗管理，确保资源配置的合理性与高效性。同时，平台层需具备灵活的算力资源选择和迁移能力，以适应不断变化的应用需求和能效要求。这些能力的提升和评价对于推动智算中心实现绿色低碳目标具有重要意义。

（1）算力资源选择和迁移能力

算力资源选择和迁移能力指的是企业在不同智算中心之间选择最合适的算力资源，并根据需求动态迁移计算任务的能力。这一能力确保了企业能够在多个智算中心中有效分配资源，从而优化计算效率和能耗。目前，随着多智算中心协同运作的需求增加，一些企业通过采用先进的调度算法和虚拟化技术，提升了资源选择的精准性和任务迁移的灵活性。

（2）平台碳排放量监测与统计

平台碳排放量监测与统计这一指标旨在对平台运行过程中产生的碳排放量进行精准跟踪和详细记录。该指标通过对计算资源消耗、电力使用、制冷需求等环节的实时数据收集与分析，确保能够准确评估平台的碳排放水平。现阶段，部分智算中心已开始采用智能监测系统和数据分析技术，实现了对碳排放的细化管理。通过持续监测和统计，平台能够及时调整资源配置和运行策略，以减少碳排放，实现低碳运营目标。

（3）算力调度

算力调度是智算中心通过智能化算法对计算任务进行动态分配，以实现资源利用效率最大化和能耗最小化。该指标涵盖了对实时业务

需求、资源可用性、能耗状况的综合评估，通过调度优化算法，动态调整任务执行顺序与资源配置，确保系统在高效运行的同时降低能耗。随着计算需求的多样化和绿色低碳目标的推进，算力调度的精确性和智能化程度越来越受到重视。部分智算中心已开始采用高级调度算法和机器学习技术，以期实现更为精确的资源分配和能耗管理。

（4）计算资源占用率

计算资源占用率指的是在一段时间内，计算资源被占用的百分比。高占用率意味着智算中心能够在处理大量任务时，最大限度地发挥处理器的性能，减少闲置资源，从而降低不必要的能耗。随着计算任务复杂性增加和资源需求的多样化，提升占用率已成为优化平台层运行的重要方向。当前，一些智算中心通过更精细的任务调度和资源管理，逐步减少计算资源的空闲时间和低效运行，从而在保持高性能计算的同时，降低能耗并推动绿色低碳发展。

（八）服务能力

服务能力是智算中心对外提供价值的关键。业务范畴上，具备卓越服务能力的智算中心能够迅速响应市场变化，灵活调整资源配置，为用户提供定制化、高性能、低延迟的计算服务，包括智算基础设施服务、算力资源服务、大模型服务和服务质量保障等。**层次划分上，**智算中心的服务能力涵盖了基础设施、大模型服务、服务质量三个层次。智算中心提供规划设计、系统集成等基础设施服务，全方位的大模型服务涵盖数据清洗、标注、增强以及模型选型、训练推理等环节，并且从无故障运行等方面保障服务质量。算力资源主要关注计算资源

的提供和优化，根据用户需求提供灵活的计算资源配置方案，包括CPU、GPU、NPU 等各类计算节点的组合与调度。在模型服务上，预训练模型和定制化模型开发等服务支持用户加速自身的应用开发和部署进程。中南智算中心依托天翼云“慧聚”平台。大幅降低大模型训练、微调、部署、推理的门槛，提供一站式、全链路、低门槛、高安全的大模型训推服务，为各行业、各场景提供从模型生产到应用闭环的解决方案。

1.智算基础设施服务

（1）基础设施

基础设施集成服务包含基础设施规划设计、建设以及工程安装等。基础设施集成服务应具备从基础设施集成发展到基础设施和 IT 设备联合勘测和规划设计能力，以及针对 IT 系统的特征对基础设施提供改造方案能力；智算中心建设主要涉及安全管理、物料管理、质量管理、进度管理以及测试验收等数字化管理能力；工程安装主要是门禁申请、设备签收、安装前检查、硬件部署、硬件初始化、固件升级、硬件压测、硬装验收。

（2）算力平台

算力平台集成服务主要包含子系统设计与实施以及集群系统集成服务。智算子系统规划设计实施主要是规划设计、OS 安装、软件部署、单机综合测试、单机训练测试、智算子系统验收。智算集群系统集成主要是集群系统需求调研、集群系统规划设计、集群系统对接联调、参数面集合通信测试、集群性能测试、集群稳定性测试、集群

初始化调优、集群试运行、集群验收、项目管理。算力平台集成服务应具备算存网协同整体规划能力，包括整体风险、质量、进度、沟通、问题、变更等一系列针对集群的基本功能和集群模型训练性能；

集群集成功能/性能测试验证能力，包括应具备算存网协同测试能力，包括整体风险、质量、进度、沟通、问题、变更等一系列针对集群的基本功能和集群模型训练性能测试能力，包含但不限于线缆链接、信号质量、配置部署和产品状态等方面的验证测试能力。以及万卡/十万卡风冷/液冷集群端到端规划设计、安装集成、测试验证能力。

（3）算力资源

算力资源服务包括算力租赁、资源管理和算力加速等服务。算力租赁将计算资源（如 XPU 等硬件资源以及相关的存储、网络资源）封装成服务，以租赁的形式提供给有需求的用户。这种业务模式允许用户根据自己的计算需求，灵活选择所需的算力资源量和使用时长。同时，还确保资源负载均衡，根据实时需求动态调整资源分配，避免过载与闲置。支持无缝动态扩容，随着业务增长迅速增加算力资源，确保业务连续性与高效运行。集群算力加速指智算中心部署后，使得有效算力进行进一步提升和加速的能力。针对集群训练场景应支持如并发调度能力、模型与算法优化能力、运行加速和编译优化能力；针对集群推理场景应支持如推理调度优化、模型与算法优化等能力。

2.大模型服务

（1）数据服务

数据服务主要涉及文档解析服务，数据清洗服务，数据标注服务，

数据增强服务，语料质量评估等服务。文档解析服务基于客户提供的非结构化/半结构化文档，通过文本和信息处理形成初始的训练数据集。数据清洗服务针对解析后的数据，进行数据清洗，从大规模的数据中解析出有效、合规的原子知识，生成 Token，并进行向量化。数据标注服务对清洗后的原始数据进行标记，以便于算法识别特定的信息。数据增强服务通过数据合成增加预训练和 SFT 数据，形成用于训练的 QA 对。语料质量评估服务对数据工程处理和生成的语料质量进行评估，保障模型训练的数据质量，以便达到预期的训练效果。

（2）模型服务

AI 开发环境部署支持主要根据客户的具体模型需求，结合产品形态，输出模型运行环境的安装部署方案，制作与推送容器镜像。具体服务内容包括模型训练环境部署支持，模型推理环境部署支持和开发工具部署支持。AI 开发使用支持具体包含训练开发支持，推理开发支持，模型适配部署支持主要基于客户具体业务场景与需求，完成适配模型的部署与调测。

模型服务包含模型选型适配、模型增量训练及模型微调服务。在智算服务中，模型选型至关重要。模型选型适配服务制定模型评估能力框架和指标，基于主流的开源模型，在实验室搭建验证环境，构建测试集从完整的视角评估模型能力以及和客户场景的匹配度，为客户提供模型选型的建议和模型增量预训练服务，支持用户打通模型增量预训练流程，拉起多卡任务，提供故障诊断等相关服务，通过增加预训练权重、优化超参等方法提升模型精度。模型微调服务进行分布式

全参微调支持、低参微调支持；帮助客户进行微调后权重合并转换并进行推理验证。

模型迁移调优主要包含模型迁移，模型调优和模型验收。模型迁移主要提供专属服务工程师，将客户模型从源平台迁移目标 AI 平台。模型调优针对已迁移模型优化模型的计算精度与计算性能至服务验收标准。包含模型精度调优和模型性能调优。模型验收基于客户提供的 AI 平台和验证环境，使用业界开源或客户提供的数据集对迁移后的模型进行测试。验收通过后，输出模型迁移测试报告，提供可直接在 AI 平台训练的模型代码、训练脚本与使用说明文档。

3.服务质量

（1）基础设施保障能力

基础设施保障能力是智算中心服务能力中的重要一环，它专注于对供配电系统、暖通系统等关键基础设施的日常巡检、维护与管理。基础设施保障能力的主要任务是通过建立科学、系统的巡检制度，明确巡检内容、频率及责任分工，确保对智算中心的核心设施进行全面、细致的监控与检查。

（2）无故障运行

智算中心的无故障运行时间（MTBF, Mean Time Between Failure）指的是在预定的运行周期内，系统能够持续、稳定地提供计算资源和服务，不出现影响业务连续性的中断或故障。这一指标是衡量智算中心服务质量、技术实力和运维能力的重要标尺。据 ODCC 调研，一个千卡集群在运行百亿参数的无故障运行时间在 7 天左右，万卡集群的

无故障运行时间约为 24 小时。

（3）故障快速恢复

故障快速恢复能力是指智算中心发生故障时能够快速进行故障恢复的能力，用平均故障恢复时间（MTTR, Mean Time To Repair 单位：ms）来衡量。具体表征智算中心在执行特定任务时，中心的某部分或整体 M 次(M⩾3)发生同一故障而无法继续执行任务的时间点，与该故障被修复，任务重新获得执行的时间点之间的差的平均值

（九）智能运营

运营管理正在逐步迈入以设施、平台、体系、服务为核心要素的智能运营发展阶段。智算中心以大规模、超大规模为主，海量的设备和复杂的系统为高效管理带来了挑战。如果缺乏与之相匹配的智算中心精细化运维手段，势必会造成电力和网络成本的浪费。智算中心需要在全自动、互联、自运维的基础设施环境下，通过全方位的监控系统感知并准确定位故障，通知智能决策系统下发变更、维护等指令，实现运维从数据输入到预测性维护全过程的数字化和自动化，形成智算中心运维全生命周期的服务能力。此外，智算中心单一的系统模块节能已经达到天花板，需全面转向系统化节能，芯片、服务器、机柜、制冷系统、配电系统、机房系统等环节缺一不可。AI 节能通过软件赋能硬件管理，利用人工智能技术来降低智算中心能耗。中国联通福州智云数据中心依托人工智能驱动的节能优化架构，实现了从设计到运维的全生命周期节能调优能力，通过能耗动态采集与数字化仿真技术，有效提升了能源利用效率，获得 AI 节能等级 4A 认证，为行业智能

节能树立了典范。

1.监控管理

监控管理是指智算中心应具备的一种核心能力，它允许通过统一标准的方式对资源进行监控与管理。这包括采用业界标准的带外管理协议（如 IPMI、Redfish）和带内管理方式（如 SSH、命令行），以确保跨平台兼容性。同时，资源池内同类型设备需提供统一的北向接口，以标准化方式上报状态、告警等信息，便于集中管理和快速响应。此举旨在提升运维效率，保障算力资源的高效稳定运行。

2.自动化运维

传统运维方式存在依赖人工操作、响应速度缓慢的问题，难以适应复杂多变的业务需求。为解决这些痛点，运维系统应具备全流程自动化功能，覆盖任务分配、状态监控及问题处理等环节。同时，此功能应支持 PC 端、移动端、大屏端等多平台访问，确保用户无论身处何地都能迅速掌握运维动态，高效执行管理操作，显著提升运维效率与应急响应能力。

3.维保管理

随着智算中心规模扩大，传统维保管理方式已难以满足高效运维需求。为确保机房设备稳定运行，需推进维保管理优化。优化应具备两大要点：一是维保计划电子化，实现任务自动分配与跟踪，确保维保工作如期进行；二是维保内容标准化，制定统一作业指导书，减少人为差异，保障维保质量。这两方面将有效提升维保效率与设备管理水平，确保业务连续性。

4.容量管理

容量管理是指对智算中心各层级（系统、机房、机柜、设备）的容量进行全面监控与分析。通过多维度监控，管理者能精准掌握资源使用状况，有效减少资源碎片化，确保资源分配的安全性与精细化。同时，基于实时数据，容量管理还支持前瞻性规划，助力提前预判并调整资源容量，确保业务连续性与扩展性。

5.智能运行

智能运行是指电气、暖通、安防等自动化运行设施结合软件能力，从快速地发现问题、及时地通报问题、准确地判断问题、高效地处置问题等方面，助力数据中心破除“人为主责”的局面，满足客户越来越高的 SLA（service Level agreement，服务等级协议）要求。安全性上，人为操作易出错，设施智能运行能实现更深层次的安全性。从效率角度来看，和汽车的自动驾驶一样，数据中心设施的自动化运行可以降低对人员的依赖，提升效率。

6.节能管理

借助大数据、AI 技术、数字孪生等技术，构建智算中心的 IT 设施与机房基础设施协同的智能化节能管理体系。运用 AI 算法预测 IT 设备运行工况、优化能源使用、智能调度资源，实现主动管理、精准调优。通过机器学习、大数据分析等技术，对智算中心的运行数据进行深度挖掘，提升能耗优化决策的准确性与效率。

五、智算中心发展建议

（一）强化创新引领，提升自主研发能力

持续推进智算中心软硬件基础设施研发投入和技术创新，把握技术主动权。重点加强 AI 芯片技术、开发框架、算子算法等的深入钻研，形成一批具有自主知识产权的核心产品和技术，增强算存运一体化能力。同时也应在基础设施上实现技术突破，比如柔直供电、液冷技术、微电网等，以应对智算业务日益灵活的需求。在智算需求增长和技术变革的背景下，软硬件加速融合的趋势显著，创新驱动尤为重要，加强产学研合作，加快科技成果转化，为智算中心高质量发展打下坚实基础。

（二）推动标准制定，促进技术规范发展

加强智算标准顶层设计，完善智能算力标准体系，明确标准化重点方向，系统开展智能算力标准制定。应系统开展智算中心建设、智算调度、计算架构、训练框架、数据接口、信息安全、软硬件规范等标准体系建设，特别是加快推进 AI 芯片等重点标准的研制。通过建立全面的评价标准，引导企业在研发、生产、管理等环节对标达标，有效促进智算资源的合理分配，帮助中小企业更好地利用智算服务，降低其应用成本。此外，标准化也有助于解决在高速发展中出现的不规范问题，推动智能算力产业的可持续进步。

（三）开展测试服务，助力评价体系完善

对智算中心的软硬件设施开展测试认证，通过评估问效不断丰富并完善综合评价系统。建立一套科学、系统且具备前瞻性的方法论，

紧密结合国内外智算技术的最新发展趋势，充分考虑我国智算中心的实际情况，涵盖硬件基础设施、软件平台以及整体系统的集成与优化能力。面向智能算力的供给方开展智能算力相关测试认证工作，树立智能算力应用典范，推广一批具有示范效应的智算实践。通过不断的测试实践，获取前沿的技术情况，了解实际地发展水平，持续完善综合评价体系的构建，使其更加地科学、客观、合理。

（四）构建智算生态，推动全产业链协同

构建“智算生态圈”，聚合产学研用力量，加快智算基础设施高质量建设。去年，在工信部的指导下，“算力产业发展方阵”成立，重点围绕产业研究、应用培育、协同创新、国际合作四方面开展相关工作，旨在促进相关主体间的交流和深度合作，促进供需对接、技术革新、知识共享，形成优势互补，加强应用推广，有效推进算力产业发展。汇聚产学研用多方力量，全面推进 “智算生态圈”建设，优化智算资源配置，提高智算服务效能，促进资源高效协同管理，打造智算资源共享、平台共建、价值共创的产业生态！

中国信息通信研究院 云计算与大数据研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62300095

传真：010-62300095

网址：www.caict.ac.cn