

一体化算力统筹云平台的架构设计、关键技术及应用研究

赵馨茹, 江琦, 贾明, 贾丽*

吉林省有云数智科技有限公司, 吉林 长春

收稿日期: 2026年7月2日; 录用日期: 2026年7月30日; 发布日期: 2026年8月6日

摘要

随着数字经济与人工智能技术的快速迭代, 算力作为核心生产力的地位日益凸显, 然而碎片化的算力资源分布、异构算力之间的协同壁垒以及供需错配等问题, 严重制约了算力的高效利用。算网融合作为新型基础设施建设的重要方向, 为打破算力孤岛、实现资源优化配置提供了关键思路。本文通过系统梳理一体化算力统筹云平台的总体架构与功能模块, 深入剖析异构算力统一管理、智能调度策略、P2P计算网络等关键技术, 探索平台在工业互联网、智慧城市、科学计算等场景的实际应用价值, 最终为构建高效、弹性、安全的算力服务体系提供理论支撑与实践参考。

关键词

算网融合, 一体化算力, 算力调度, 统一管理

Architectural Design, Key Technologies, and Application Research of an Integrated Computing Power Coordination Cloud Platform for Computing Networks

Xinru Zhao, Qi Jiang, Ming Jia, Li Jia*

Jilin Youyun Shuzhi Technology Co., Ltd., Changchun Jilin

Received: July 2, 2026; accepted: July 30, 2026; published: August 6, 2026

*通讯作者。

文章引用: 赵馨茹, 江琦, 贾明, 贾丽. 一体化算力统筹云平台的架构设计、关键技术及应用研究[J]. 计算机科学与应用, 2026, 16(8): 1-7. DOI: 10.12677/csa.2026.168256

Abstract

With the rapid advancement of digital economy and artificial intelligence technologies, computing power has become a pivotal productivity driver. However, challenges such as fragmented computing resource distribution, coordination barriers between heterogeneous computing systems, and supply-demand mismatches severely hinder its efficient utilization. The convergence of computing and networking infrastructure, as a key direction in modern infrastructure development, provides critical solutions for breaking down computing silos and optimizing resource allocation. This study systematically examines the overall architecture and functional modules of integrated computing coordination cloud platforms, delves into key technologies including unified management of heterogeneous computing resources, intelligent scheduling strategies, and P2P computing networks, and explores practical applications in industrial internet, smart cities, and scientific computing scenarios. The findings ultimately offer theoretical foundations and practical references for building efficient, scalable, and secure computing service ecosystems.

Keywords

Computing-Network Convergence, Integrated Computing Power, Computing Power Scheduling, Unified Management

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着全球经济向数字化、网络化、智能化方向转型升级，算力已成为数字经济时代的新质生产力要素。IDC、浪潮信息和清华大学全球产业研究院联合发布的《2022-2023 全球算力指数评估报告》¹指出，算力是经济增长的主要驱动力，算力指数平均每提高 1 点，国家的数字经济和 GDP 将分别增长 3.6% 和 1.7%。经中国信通院测算，2023 年全球计算设备算力总规模为 1397 EFLOPS，同比增长 54%，预计未来五年全球算力规模将以超过 50% 的速度增长，至 2030 年全球算力将超过 16 ZFlops，其中智算占比将超过 90%。算力如同农业时代的水利、工业时代的电力，已经渗透到生产生活的各个领域，是激活数据要素潜能、驱动经济社会数字化转型的新引擎。在国家政策扶持方面，如《工业和信息化部关于印发〈新型数据中心发展三年行动计划（2021-2023 年）〉的通知》²等强调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、边缘数据中心等进行合理的梯次布局。其中根据《关于加快构建全国一体化大数据中心协同创新体系的指导意见》³，全国一体化大数据中心体系规划了 8 个全国一体化算力网络国家枢纽节点和 10 个国家数据中心集群的总体布局，还明确到 2025 年全国数据中心要形成布局合理、绿色集约的格局，大型、超大型数据中心运行电能利用效率降到 1.3 以下。同时，以此为指导启动的“东数西算”工程，能够优化现有数据中心布局，按需分配东西部枢纽资源，助力数据中心规模化、集约化、绿色化发展，进一步推动算力资源的精准配置和按需获取，有效降低算力能源消耗，加快

¹<https://www.igi.tsinghua.edu.cn/info/1019/1321.htm>

²https://www.gov.cn/zhengce/zhengceku/2021-07/14/content_5624964.htm

³https://www.gov.cn/zhengce/zhengceku/2020-12/28/content_5574288.htm

推动异构算力调度技术创新与业务推广。在吉林省政策供给方面,吉林省人民政府办公厅印发的《吉林省大数据条例》⁴《吉林省大数据产业发展指导意见》⁵等明确指出推动传统数据中心智能化改造,探索建立算力调度机制,化解算力供需矛盾,积极发展算力产业生态,强化算力供需对接,推动算力产业区域协同发展。

国家全面推进算力建设与应用,本研究提出打造一体化算力统筹云平台的应用设计,汇聚通用与智能算力、高性能及边缘算力等多元算力资源,针对不同用户不同需求,设计异构资源池调度引擎,实现跨资源池跨架构跨厂商的异构算力资源调度[1]。平台服务于政务、企业、科研机构及个人用户,满足在人工智能训练、大数据分析、高性能计算等创新场景下的算力需求。用户可通过平台按需购买算力资源,无需自行投入昂贵的硬件设备,大幅降低使用成本,提升业务效率[2],也可以把闲置的算力资源加入平台,作为资源出租给有需求的其他用户。

2. 一体化算力统筹云平台的架构设计

针对国内智算中心数量快速增长,但算力需求与供应间存在信息隔离、缺乏统一管理平台及个人AIPC算力闲置等问题,本研究提出了一体化算力统筹云平台,引入国内众多智算中心作为算力供应方[3],通过容器技术和插件化驱动管理模式,实现多种异构智算资源的统一管理,允许第三方智算中心或小型IDC直接接入并售卖,还可作为管理平台直接使用。

2.1. 平台的总体架构

平台的总体架构设计主要包括资源层、架构层、服务层和应用层等,具体架构设计内容如下:

(1) 资源层是平台的硬件与基础设施核心。主要解决算力资源碎片化、异构化的行业痛点,实现全域算力资源的统一池化管理。管理公有云、私有云、混合云、数据中心、中心云、边缘环境等全形态资源,实现多云协同、云边一体的全域资源覆盖,打破算力孤岛。基础配套支撑体系主要包括:一是异构算力,兼容通用/智能/高性能/边缘等全类型算力芯片,适配CPU、GPU、NPU等异构硬件,满足AI训练、推理、高性能计算等差异化算力需求。正如Arm打造的异构计算平台,可在移动端通过Armv9.2 CPU集群搭配Arm Immortalis-G925 GPU,为图像分割、自然语言处理等AI用例提供加速;在物联网场景中,Arm Ethos-U85 NPU可与Arm Cortex-A系统结合,满足工厂自动化等场景的AI性能需求;合作伙伴NVIDIA更是借助Arm CPU定制出Grace Blackwell超级芯片,与NVIDIA H100 GPU相比,能耗降低了25倍,单个GPU性能提高了30倍,充分体现了异构算力在不同场景下的适配能力与性能优势。二是数据存储,提供分布式、存算一体、混合存储能力,适配AI大模型海量数据的存储、读取与处理需求。三是无损网络,基于RoCE/SLB/VPC技术构建高带宽、低时延的网络底座,保障算力集群的高效互联与数据传输。其中RoCE技术借助RDMA特性,可让数据传输时延理论上接近1微秒,同时实现高带宽传输,还能降低CPU负担;VPC网络模型因无隧道封装消耗,网络性能更优,还可通过链路冗余和带宽聚合提升网络可靠性与数据传输速度,为算力集群的高效运行筑牢网络基础。

(2) 管理编排层是平台中枢大脑,算力智能调度核心。基于云原生技术栈构建,核心技术底座为Kubernetes + Ray + Docker,向下屏蔽底层硬件差异,向上提供统一的运行环境与调度能力。覆盖资源感知、算力结构管理、智能推理、协同编排、智能匹配、灵活调度、统一管控,实现了算力资源的虚拟化、调度智能化、运维自动化。通过K8s实现容器与集群的统一编排管理,通过Ray适配AI训练/推理的分布式算力调度,通过Docker实现开发环境的隔离与一致性,解决异构算力的统一纳管、任务的按需调度、

⁴http://credit.jlcity.gov.cn/zcfg/snzc/202401/t20240110_2946879.html

⁵http://xxgk.jl.gov.cn/zsjg/fgw_136504/gkml/202305/t20230529_8716108.html

资源的弹性伸缩问题，最大化提升算力资源利用率。

(3) 框架层是平台的 AI 开发能力核心，全流程研发支撑。为算法研发提供开箱即用的开发环境与基础能力，是衔接底层算力与上层业务的关键环节。兼容主流框架，支持 PyTorch、TensorFlow、MXNet、Caffe 等业界主流深度学习框架，覆盖绝大多数 AI 开发场景的技术需求。其中 TensorFlow 作为谷歌开发的开源框架，在全球开源 AI 框架中的市场份额已超过 50%，凭借卓越的可扩展性和丰富的预训练模型，在移动端、云端和边缘计算中都能高效运行；PyTorch 则凭借动态图计算和优异的 GPU 加速性能，成为科研和工业界的首选，2025 年超过 70% 的 AI 研究论文采用其作为底层框架；Caffe 作为曾在 GitHub 上 star 数量最多的深度学习项目，在图像处理领域表现高效，拥有庞大的用户社区；MXNet 也凭借自身特性在深度学习领域占据一席之地。内置自动微分、张量运算、损失函数、计算图、优化器、分布式训练、精度调优等 AI 模型开发全流程核心能力，大幅降低 AI 开发的技术门槛，同时适配从科研级算法验证到产业级大规模模型训练的全场景需求。

(4) 服务运营层是平台技术能力向商业价值转化的核心环节。是服务交付的核心门户，也是技术商业化的重要载体。实现算力资源的服务化封装与全生命周期运营管理。面向用户的服务交付，提供算力目录、算力注册/封装、服务申请、场景感知、服务结算等能力，将底层算力与开发能力转化为标准化、可按需申请的服务，实现算力服务的一站式开通与使用。面向平台的运营管理，内置身份认证、用户管理、算力计费、运营分析等模块，为平台运营方提供用户全生命周期管理、服务计量计费、运营数据复盘的全流程能力，支撑平台的可持续商业化运营。

(5) 算力应用层是平台算力价值的最终落地载体，为场景化应用赋能。锚定千行百业的智能化转型需求，实现算力能力的场景化适配与赋能。覆盖金融科技、智慧园区、智能制造、生物医药、智能汽车、智能能源、智慧政务、智慧交通、智慧医疗、智慧教育十大核心行业场景，覆盖工业、民生、政务、金融等核心领域。将底层标准化的算力、AI 开发能力，与行业场景的差异化需求深度结合，实现算力价值的规模化落地：比如在金融领域，中国农业银行采用青云 AI 智算平台构建云边协同体系，为全国 430 余家网点和 40+ 业务场景提供边缘算力资源；某国有大行通过部署鲲鹏 + 昇腾算力底座，在信贷材料审核场景实现单日处理 50 万份，人力成本降低 35%，异构算力融合还使该行算力利用率提升 82%。在智能制造领域，宝钢基于华为 AI 平台部署昇腾训练与推理算力集群，结合盘古预测大模型实现热轧 1880 产线精准控制，解决了传统生产中检测不足、纠偏难的行业痛点。同时行业场景的需求反向驱动底层算力调度、服务能力的迭代优化，形成技术与业务的正向闭环。

(6) 支撑体系是全栈全域的安全保障与管理能力。构建了贯穿全栈的两大支撑体系，覆盖平台从底层资源到上层应用的全层级、全生命周期，保障平台的安全、稳定、可控运行。一是全链路安全保障体系，采用纵深防御的设计理念，构建了覆盖网络安全、主机安全、数据安全、运维安全、态势感知的全维度安全防护体系。针对算力平台的核心风险点，从基础设施安全、网络边界防护、数据全生命周期安全、运维操作台规、全局安全态势感知多个维度，实现端到端的安全防护。保障算力资源的可信访问、用户数据的安全合规、平台运维的可控可管，同时通过态势感知实现安全风险的实时预警与处置，满足等保合规、数据安全法等监管要求，适配政务、金融等强合规场景的安全需求。在金融场景中，算力安全与区块链技术的融合可重塑风控技术范式，像中国农业银行采用青云 AI 智算平台构建总分一体化云边协同体系，为全国 430 余家网点和 40+ 业务场景提供边缘算力资源，在提升运营效率的同时增强了安全性；广发证券通过 GPU 资源池化解决方案，在提升算力利用率的同时强化了安全能力。当前金融领域强监管态势持续升级，2023 年上半年原银保监会及其派出机构共开出 2425 张监管处罚罚单，总金额达 126,287 万元，这类强监管背景下，该安全能力更能满足金融机构的合规刚需。二是全周期监测管理体系，构建了“配置 - 管理 - 监控 - 运维 - 可用性保障”的闭环管理体系，覆盖资源配置、资源管理、资源监控、运

维管理、可用性管理五大核心维度。实现对平台资源、服务、应用的全维度可管可控，从资源初始纳管配置的源头管控,到运行期的实时状态监控、日常运维操作的过程管理,再到全平台高可用保障的兜底支撑，构建起完整的全周期管理闭环。为平台服务的高稳定性、资源的高可用性筑牢基础，同时依托精细化的资源监控与管理体系，进一步优化算力资源调度策略，持续提升资源利用率，有效降低平台运营成本。

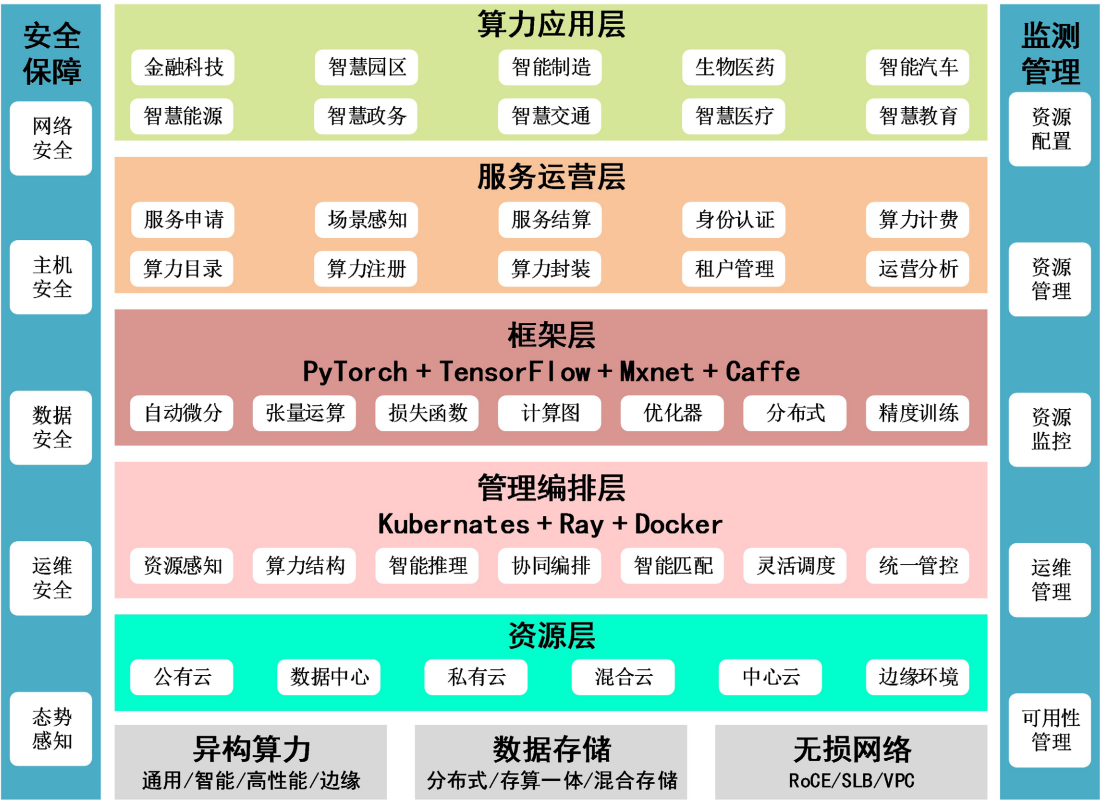


Figure 1. Overall architectural design of the integrated computing power coordination cloud platform
图 1. 一体化算力统筹云平台的总体架构设计

平台整体架构设计的核心优势如下：(1) 分层解耦，高扩展性。五层纵向架构职责清晰、模块化封装，支持各层能力的独立迭代与灵活扩展，可适配从中小规模算力集群到省级算力网络的不同建设规模需求。(2) 云原生架构，全域算力纳管。基于云原生技术栈，可通过 AI 数据编排与加速、资源统一抽象与管理等技术路径，实现异构算力、多云资源、边缘节点的统一纳管与调度，构建“云、边、端”一体的异构算力协同网络，结合“东数西算”工程构建的国家算力网络体系发展方向，彻底解决算力孤岛问题，实现算力资源的池化、弹性化、高效化利用。(3) 全栈闭环，一站式服务。从底层硬件资源到上层行业应用，配套全链路安全与管理体系，一站式满足用户从算力采购、AI 模型开发到行业应用落地的全流程需求。(4) 技术与商业双轮驱动。既通过底层技术架构解决算力调度、AI 开发的核心技术痛点，又通过服务运营层实现算力的服务化、商业化，实现技术价值与商业价值的统一。(5) 体系化保障，高可靠高合规。全栈贯穿的安全与监测管理体系，实现平台运行的安全可控、稳定可靠，可满足企业级、政务级等高端用户的合规与可用性要求。

2.2. 平台的功能设计

根据平台的总体架构设计，平台的功能主要体现在异构算力资源的统一管理、动态算力调度与负载

均衡以及去中心化的 P2P 计算网络, 具体功能描述如下:

(1) 异构算力资源的统一管理。通过虚拟化技术, 将不同品牌、不同类型的 CPU、GPU 等硬件资源抽象成统一的虚拟资源池[4]。平台自动发现或允许用户注册现有算力资源, 为每台设备分配唯一标识, 并记录其硬件配置、性能指标等信息。采用统一的资源描述标准和接口协议, 实现异构资源的标准化管理。

(2) 动态算力调度与负载均衡。开发智能调度算法, 实时监测算力资源的使用情况, 根据任务的性质、资源的需求和可用性智能分配任务[5]。通过负载均衡技术, 确保各节点任务量均衡, 避免过载或空闲。结合历史数据与实时反馈, 持续优化调度策略, 动态调整算力资源分配, 达成高效负载均衡。

(3) 去中心化的 P2P 计算网络。构建基于 P2P 模式的去中心化计算网络, 节点间直接通信与资源共享, 无需中心服务器[6]。通过节点发现与身份验证机制, 确保网络的安全与可靠。利用分布式检索和索引技术, 快速定位所需资源, 通过高效的数据传输协议, 实现节点间的直接数据传输。

通过采用上述技术和方法, 确保了本研究实施的可行性和高效性, 为用户提供一个高效、便捷、安全的算力交易与服务平台[7]。同时, 将持续优化迭代平台功能与性能, 以适配用户日益增长的需求与期望。

3. 一体化算力统筹云平台的关键技术

3.1. 异构算力资源的统一管理技术

平台支持英伟达、华为等国内外多种品牌、不同类型的异构算力资源, 通过自动发现或用户注册方式纳入统一管理, 实现资源的标准化与抽象化。虚拟化技术将物理的计算、存储、网络资源抽象成虚拟资源, 为不同类型的应用提供灵活的资源分配[8]。容器化技术提供轻量级、可移植的计算环境, 实现异构资源的统一管理和调度。微服务架构将复杂的应用分解为小的、独立的服务单元, 便于在异构环境中进行部署和调度。

3.2. 平台配置算力调度策略技术

通过代理探针, 可实时监测每个智算中心的资源闲置情况, 在为用户分配算力资源时, 可以根据策略选择调度到特定要求的集群[9]。边缘计算技术将计算任务分散到网络边缘, 减少中心数据中心的压力, 提高响应速度。分布式计算框架处理大规模数据集, 支持在异构环境中进行分布式计算。研发高效资源调度算法, 达成闲置算力资源的动态分配与实时响应。智能分析用户算力需求与资源供给情况, 动态调整智算资源的分配, 确保资源的负载均衡与高效利用。

3.3. 去中心化 P2P 模式的计算网络技术

已经完成去中心化的 P2P 计算网络的研发, 第三方智算中心、个人 AIPC 均可以把资源接入算力调度平台[10]。利用人工智能调度算法和机器学习算法动态优化资源调度策略, 实现更高的资源利用率和系统性能。基于 P2P 模式的去中心化计算网络, 实现节点间的直接通信与资源共享, 该去中心化的网络架构, 降低了系统对中心服务器的依赖, 提高了系统的可扩展性与容错性, 第三方智算中心、个人 AIPC 均将资源接入算力调度平台, 构建全国算力网[11]。实现算力出海技术突破: 搭建底层 P2P 网络, 在获取对应海外网络连接授权后, 可直接调用海外先进智算中心资源。

3.4. 服务透明计费与安全保障技术

本研究的计费系统已完成, 可提供清晰的按量计费模式与先进的加密机制, 确保资源使用的透明性与数据的安全性[12]。区块链技术构建去中心化的资源管理和结算平台, 实现透明、安全的分布式结算。

先进的加密、认证和访问控制技术,确保数据安全和系统稳定运行[13]。

本研究不仅能够提高资源的利用率,实现算力资源的负载均衡,还可以降低使用门槛,帮助算力供给企业降低前期投资和运营成本[14]。平台的研发也将有助于突破现有技术瓶颈,提升算力服务的质量与效率,推动人工智能技术升级与产业结构优化。

4. 结论

本研究有利于贯彻落实国家“东数西算”工程要求,通过算力资源的高效调度,推动东部数字经济产业链向西部延伸拓展;推动算力资源的精准配置和按需获取,比如参考中国移动落实国家“东数西算”工程部署,建成覆盖全部枢纽节点的全国性算力网络的实践,通过优化算力选址布局、提升数据中心能效等方式,有效降低算力能源消耗,助力区域协调发展和信息通信行业碳达峰、碳中和目标达成。一是技术成果转化方面。将核心技术,包括资源虚拟化、分布式计算、智能调度算法、P2P网络等,转化为实际应用产品。通过持续优化与迭代,保持平台在算力资源整合、调度与分配方面的技术优势。二是产品市场推广与应用方面。平台作为一种高效、低成本且灵活的算力应用解决方案,广泛应用于多个领域以满足对 AI 大模型设计和测试的需求。在高校院所,助力师生加速模型训练和优化。在医学图像、生物科技、化学能源、交通驾驶等专业科研领域,助力推动科学发现和创新。在角色设计、场景构建、特效制作以及数字人、直播等场景中,可有效缓解临时 GPU 激增压力,加快本研究的进展[15]。面对物联网对数据存储、分析、智能计算和图形处理的日益增长的需求,平台凭借稳定、低成本的算力资源成为理想之选。三是生态系统构建方面。围绕平台构建算力资源共享的生态系统,吸引更多的算力供应商与需求方加入。通过生态系统内的资源共享与协同,推动算力产业的区域协同发展。

基金项目

2025年吉林省科技发展规划项目成果转化“一体化算力统筹云平台的开发与应用”(20250303007CX)。

参考文献

- [1] 冯勇, 句赫, 段升位, 等. 绿色算力中心全生命周期用能管控平台设计与实现[J]. 电力勘测设计, 2025(9): 42-50.
- [2] 张雪芬, 王晓震. 以算力平台建设推动应用型大学高质量发展的策略分析[J]. 北京联合大学学报, 2025, 39(4): 7-13.
- [3] 屠佳琪, 朱晓伟, 张华. 高校通超智一体化 AI 公共算力服务平台建设研究——浙江大学的探索与实践[J]. 中国教育信息化, 2025, 31(7): 50-62.
- [4] 钱卫东, 胡兵, 张洋, 等. 钢铁行业大模型平台建设框架与关键技术[J]. 冶金自动化, 2025, 49(4): 59-71.
- [5] 尹小康. 以杭州算力资源调度平台为例探究算力中心互联互通技术路径[J]. 通信世界, 2025(13): 31-32.
- [6] 刘磊, 卢国彬, 杜科, 等. 大模型运营平台建设实践与技术优化[J]. 信息技术与标准化, 2025(6): 71-77.
- [7] 任彤, 吴美希, 郭亮. 打造中国算力平台加速构建全国一体化算力网[J]. 通信世界, 2025(11): 16-19.
- [8] 尹小康. 算力资源调度服务平台的现代化建设[J]. 通信世界, 2025(10): 41-43.
- [9] 董颐. 一种面向云原生 GIS 平台的算力网络计算任务设计[J]. 自动化与仪表, 2025, 40(5): 144-148.
- [10] 尹培培, 谢健, 赵文涛, 等. 算力赋能高校高质量发展——南航高性能计算平台建设实践[J]. 中国高校科技, 2025(5): 94-99.
- [11] 刘威辰, 杨华锋, 江民民, 等. 大模型应用服务平台建设研究[J]. 信息通信技术与政策, 2024, 50(12): 21-30.
- [12] 王思航, 匡先辉, 郭煜, 等. 加快构建全国一体化算力网分析[J]. 决策咨询, 2025(2): 12-13+41.
- [13] 孙宁, 孙霏翀, 金权, 等. 基于算网融合的工业仿真云平台架构与网络安全研究[J]. 中国宽带, 2025, 21(4): 50-52.
- [14] 唐毅, 张逍遥, 高森, 等. 基于 K8s 的异构环境混布智能算力调度平台设计与实现[J]. 通讯世界, 2025, 32(2): 178-180.
- [15] 洪涛, 程乐. 全国算力体系一体化建设的五大问题及治理对策[J]. 中国科学院院刊, 2024, 39(12): 2086-2095.