DOI:10.20033/j.1003-7241.(2026)04-0064-05

基于图注意力网络的算力资源调配优化研究

潘 远[1], 张华兵[2], 刘梓健[2], 陈晓江[2], 钟秋辉[3]

(1. 中国南方电网有限责任公司,广东 广州 510700;2. 南方电网数字电网集团有限公司,广东 广州 510663;
3. 深圳市远界管理咨询有限公司,广东 深圳 518000)

摘 要:为了提高网络资源的利用率,研究利用网络功能虚拟化技术来降低计算机网络功能部署的复杂性,并结合功能服务链(service function chain,SFC)的用户请求设计出了一个算力资源调配系统。该系统由多个虚拟网络功能(virtualized network function,VNF)提供服务,可将用户请求拆分为多段VNF任务并映射至层级化算力节点。同时,利用图注意力网络(graph attention network,GAT)构建用户请求预测模型,借助其节点注意力机制精准捕捉请求分布规律,并结合Adam优化算法,基于该模型预测的用户请求动态优化算力资源的调配方案。结果显示,研究策略的阻塞率为9.41%,较初始策略、LAMP策略、RAMP策略分别降低15.78%、24.08%、25.65%。中央处理器(central processing unit,CPU)利用率为73.54%,内存利用率为65.04%。基于图注意力网络的算力资源调配优化研究有效地提升了算力资源的调配效率,从而显著提高了网络整体的资源利用率,为算力网络高效运维提供了技术支撑。

关键词:图注意力网络;算力资源;资源调配;功能服务链;网络功能虚拟化;Adam优化算法;网络阻塞

中图分类号:TP311.13 **文献标志码**:A **文章编号**:1003-7241(2026)04-0064-05

Research on optimization of computing power resource allocation based on graph attention networks

PAN Yuan[1], ZHANG Huabing[2], LIU Zijian[2], CHEN Xiaojiang[2], ZHONG Qiuhui[3]

(1. China Southern Power Grid Co., Ltd., Guangzhou 510700, Guangdong, China;
2. Southern Power Grid Digital Grid Group Co., Ltd., Guangzhou 510663, Guangdong, China;
3. Shenzhen Yuanjie Management Consulting Co., Ltd., Shenzhen 518000, Guangdong, China)

Abstract: To improve the utilization of network resources, this paper studies the use of network function virtualization technology to reduce the complexity of computer network function deployment, and designs a computing resource allocation system based on the user request of service function chain (SFC). The system is served by multiple virtual network functions (VNF), which can split user requests into multiple vnf tasks and map them to hierarchical computing nodes. At the same time, the graph attention network (GAT) is used to build a user request prediction model, with the help of its node attention mechanism to accurately capture the distribution of requests, and combined with the Adam optimization algorithm, the user requests predicted by the model are dynamically optimized for the allocation of computing resources. The results show that the blocking rate of the research strategy is 9.41%, which is 15.78%, 24.08% and 25.65% lower than the initial strategy, lamp strategy and ramp strategy, respectively. The utilization rate of central processing unit (CPU) is 73.54%, and the memory utilization rate is 65.04%. To sum up, the research on the optimization of computing resource allocation based on graph attention network effectively improves the allocation efficiency of computing resources, which significantly improves the overall resource utilization of the network, and provides technical support for the efficient operation and maintenance of computing network.

Keywords: graph attention network; computing resources; resource allocation; functional service chain; network feature virtualization; adam optimization algorithm; network congestion

随着人工智能、云网络和在线视频等大规模密集型网络应用的发展,互联网行业面临着满足高速增长的算力需求的挑战[1]。算力网络作为一种新型信息基础设施,根据业务需求在云、网、边之间按需分配和灵活调度计算资源、存储资源和网络资源[2]。然而,传统通信网络依赖专用且封闭的硬件设备,导致网络扩容和新业务开发与部署变得异常困难[3]。虽然云计算具备强大的存储和运算能力,在一定程度上满足了传统通信网络的服务要求,但其核心虚拟化技术依赖于将数据必须先传输至云计算中心进行处理后再返回的模式。与传统网络通信和云计算技术不同,

收稿日期:2024-12-27;**录用日期**:2025-02-01

基金项目:中国南方电网创新项目(0000002023030301XX00107)

作者简介:潘 远(1986—),男,硕士,高级工程师,研究方向:电网数字化、信息系统运行、网络安全、电力调度。

引用本文:潘远, 张华兵, 刘梓健, 等. 基于图注意力网络的算力资源调配优化研究[J]. 自动化技术与应用, 2026,45(4):64-68. (PAN Yuan, ZHANG Huabing, LIU Zijian, et al. Research on optimization of computing power resource allocation based on graph attention networks[J]. Techniques of Automation and Applications, 2026,45(4):64-68.)

算力网络的分布特性使其节点遍布各地,而非集中于云数据中心,因此能够有效解决云计算技术和传统通信网络的问题[4]。然而,算力网络的分布结构也导致了该网络节点需要进行统一分配管理,若算力资源未进行合理调度和管理,就会造成大量的资源浪费。功能服务链(service function chain,SFC)是将网络服务按照特定顺序连接起来,以便实现复杂的服务功能的方法。而网络功能虚拟化(network function virtualization,NFV)技术是通过软件来实现传统上依赖于硬件的网络功能技术。功能服务链(service function chain,SFC)可以将算力网络的节点相连接,而网络功能虚拟化(network function virtualization,NFV)可以进一步降低网络功能部署的复杂性和资源调度成本[5]。因此在这一背景下,研究创新性地将 NFV 和 SFC 相结合应用于算力网络中,并通过用户请求感知进行服务映射,利用图注意力网络(graph attention network,GAT)构建请求预测模型以优化算力资源调配。进行基于图注意力网络的算力资源调配优化研究,以期提高网络整体的资源利用率。

1 基于 GAT 的算力资源调配优化

1.1 功能服务链映射策略研究

算力实质上是网络的计算能力。智能手机、电脑以及高性能服务器的反应速度越快,表明其算力越高。算力网络通常由大量的边缘节点和雾节点构成,每个节点实际上都是一台物理设备[6]。在算力网络中,用户请求从用户终端设备发送到算力适配器。适配器对请求进行初步分析和处理后,将请求拆分为多个 VNF 后发送到算力节点。通常,一个用户请求会被拆分为 3 条 VNF,这 3 条 VNF 之间通过多条虚拟链路相互连接[7]。因此,算力网络处理用户请求的过程就是将这 3 条 VNF 及其关联的虚拟链路映射到算力网络的过程[8-9]。这个映射过程可以转化为 SFC 的映射问题。基于用户请求特征的功能服务链映射策略如图 1 所示。

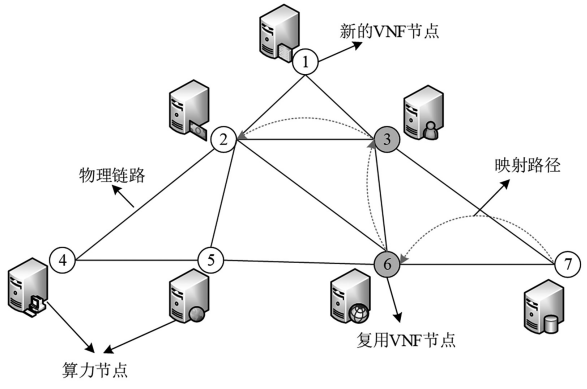


图 1 基于用户请求特征的功能服务链映射策略

Fig. 1 Functional service chain mapping strategy based on user request characteristics

由图 1 可知,在基于用户请求特征的功能服务链映射策略中通过合理规划映射链路,可以减少 VNF 的实际映射数量,从而提高网络资源的利用率。在处理用户请求时,VNF 映射链路优先选择网络中已复用的 VNF 实例,以降低请求处理时间并提高效率。当存在符合条件的复用实例时,系统会优先选择这些实例。然而,如果没有符合条件的实例,或者同类型实例的复用次数已达到上限时,则必须选择新的 VNF 节点进行映射。VNF 映射链路的选择属于多目标优化问题。在面临多个候选算力节点时,如果这些节点的映射代价相同,那么系统倾向于选择网络跳数较短的节点作为映射目标[10]。这样的选择旨在降低数据传输的延迟和成本。如果这些候选节点的跳数也相同,那么系统将采用随机策略从中选择一个算力节点作为映射节点[11]。这种处理策略可以确保系统在资源分配和网络效率之间取得平衡,同时可以满足用户请求的需求。最小映射代价即为最小化网络资源总开销,其中包含了内存开销、中央处理器(central processing unit,CPU)开销以及节点之间的通信开销等因素[12-13]。映射代价的数学表达式为

$$O = \frac{\alpha \times df_i^c}{R_n^c} + \frac{\beta \times df_i^m}{R_n^m} \tag{1}$$

式中,O 表示映射代价,α 和 β 分别表示 CPU 代价权重和内存代价权重,f_i^c 和 f_i^m 分别表示第 i 条 VNF 映射链路所需要的 CPU 资源和内存资源,d 表示 VNF 映射链路的传输距离,c 表示 CPU 计算资源,m 表示内存存储资源,R_n^c 和 R_n^m 分别表示第 n 个用户请求所耗费的 CPU 资源和内存资源。

网络资源总开销的数学表达式为

$$O' = \sum_{R_n \in R_{SFC}} \sum_{f_i \in F} \left(\frac{\alpha \times df_i^c}{R_n^c} + \frac{\beta \times df_i^m}{R_n^m} + \frac{\varepsilon \times df_i^b}{B_i} + \frac{\gamma \times T_i}{df_i^t} \right) \tag{2}$$

式中,O' 表示网络资源总开销,R_{SFC} 代表 SFC 对应的所有用户请求集合,ε 和 γ 分别表示带宽资源和网络传输的代价权重,f_i^b 和 f_i^t 分别表示第 i 条 VNF 映射链路所需要的带宽资源和网络传输资源,b 代表带宽资源,t 代表传输资源,B_i 和 T_i 分别表示第 i 条映射链路所使用的带宽资源和网络传输资源。

通过用户请求特征的功能服务链映射方式,算法能够在最小化资源总开销的同时,选择出最优的映射路径,从而提高整个网络系统的效率。

1.2 算力资源调配优化

在研究的基于用户请求特征的 SFC 映射策略中,系统可以根据资源开销来选择映射路径。然而,该策略存在部分节点超载导致的局部网络瘫痪的局限性。为了解决这一局限,研究引入了图注意力网络(graph attention network,GAT)来构建用户请求预测模型对现有的映射策略进行优化,以降低用户请求的网络阻塞率,使得算力资源分布更为均衡。GAT 是一种图神经网络模型,该模型通过在节点之间引入注意力机制,来更好地捕捉节点之间的关系和特征信息[14]。因此,利用 GAT 的这一特性可以提高用户请求预测的精度,从而减少网络阻塞。GAT 的

自注意力权重如式(3)所示。

$$\eta_{ij}=\frac{\exp(e_{ij})}{\sum_{k\in K(i)}\exp(e_{ij})} \tag{3}$$

式中，i 和 j 都表示节点，η_{ij} 表示自注意力权重，k 表示节点 i 的邻居集合中的任意一个节点，$K(i)$ 表示节点 i 的邻居节点集合，e_{ij} 表示原始注意力得分。

GAT 预测模型的目标函数表达式为

$$L=\| Y_t-\hat{Y}_t \| +\psi L_2 \tag{4}$$

式中，L 表示损失函数，L_2 表示正则化项，ψ 表示正则化项系数，Y_t 和 $\hat{Y}_t$ 分别表示在 t 时刻的输入请求数量和输出预测请求数量。

在 GAT 模型预测出用户请求之后，研究利用 Adam 算法来优化算力资源的调配。Adam 算法是一种自适应学习率的优化算法，能够有效处理稀疏梯度和非平稳目标[15-16]。并且，该算法在处理大数据和高维空间时表现尤为出色，因此研究选择该算法来优化大规模数据下的算力资源调配。基于图注意力网络请求预测模型的算力资源调配网络架构如图 2 所示。

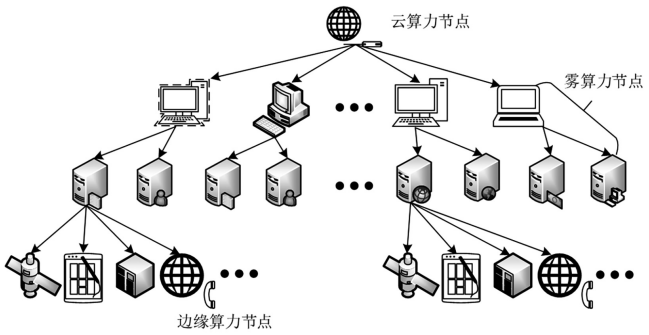


图 2 基于图注意力网络请求预测模型的算力资源调配网络架构

Fig. 2 Computational resource allocation network architecture based on graph attention network request prediction model

由图 2 可知，基于图注意力网络请求预测模型的算力资源调配网络架构采用树形结构进行多层节点连接，1 级 4 四级算力节点的数量分别为 128 个、32 个、16 个和 1 个。每一层结构根据资源属性进行划分，其中边缘算力节点的 CPU 性能为 4 000 MIPS，其余节点的 CPU 性能为 10 000 MIPS。该网络架构采用 SFC 范式构建，以便将 VNF 任务拆分，并将拆分后的多个 VNF 节点任务映射到合适的算力节点上。

2 算力资源调配优化策略验证

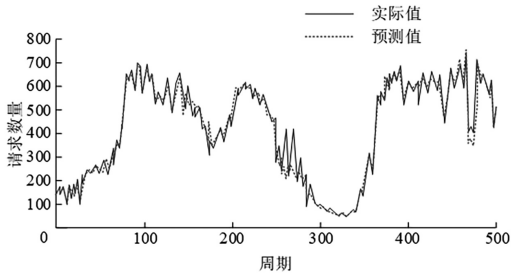
为了验证基于图注意力网络的算力资源调配优化策略的有效性，研究首先进行了实验环境的搭建和节点属性的配置。实验环境配置的是 macOS 12.4 操作系统，并配备了四核的 Intel Core i5 处理器。使用的深度学习框架为 TensorFlow 1.5.0，Python 版本为 3.8.8，硬盘和内存配置分别为 512GB 和 32GB。在构建的算力网络系统中节点按四个等级进行划分，该层次的算力网络结构以树形结构进行连接。研究使用的数据集为 PEMS-BAY 数据集和 Cityscapes 数据集。其中 PEMS-BAY 数据集包含了传感器在不同时间段的流量数据，适用于交通网络和算力资源调配的研究。而 Cityscapes 数据集具有大规模数据量，适用于测试算力资源调配策略的效果。为了评估和验证算力资源调配的优化效果，研究将这两个数据集按照 4∶6 的比例划分为了测试集和训练集。具体的实验环境和节点属性配置如表 1 所示。

表 1 实验环境和节点属性配置

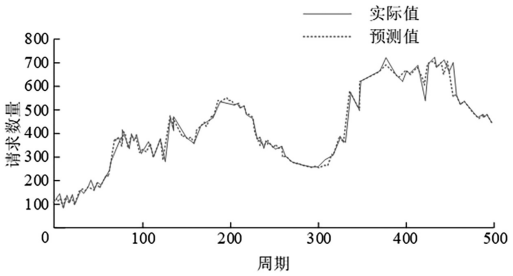
Tab. 1 Experimental environment and node attribute configuration

项目	配置	节点等级	节点名称	链路带宽/(MB·s^{-1})
系统	MacOS 12.4	1 级	边缘算力节点	100
处理器	Intel Core i5	2 级	雾算力节点	500
深度学习框架	TensorFlow 1.5.0	3 级	雾算力节点	500
Python	3.8.8	4 级	云算力节点	5 000

为了验证基于 GAT 请求预测模型的有效性，研究在测试集和训练集中进行了实验验证。基于 GAT 请求预测模型的预测值与实际值的对比如图 3 所示。从图 3 中可以看出，研究模型的预测值与实际值高度重合，证明该模型能够准确地预测用户请求。在图 3(a)中可以看出，在测试集中迭代次数接近 490 时，预测值和实际值之间出现了最大的误差。此时的实际用户请求数量为 400，而预测数量为 354，误差达到了 11.50%。在图 3(b)中可以看出，在训练集中迭代次数接近 430 时，预测值和实际值之间出现了最大的误差。此时的实际用户请求数量为 543，而预测数量为 600，误差达到了 9.50%。综上可以看出，基于 GAT 请求预测模型的用户请求预测效果在迭代后期上还有提升空间，但整体上能够对用户请求进行较为准确的预测。



(a) 测试集中预测值与实际值的对比

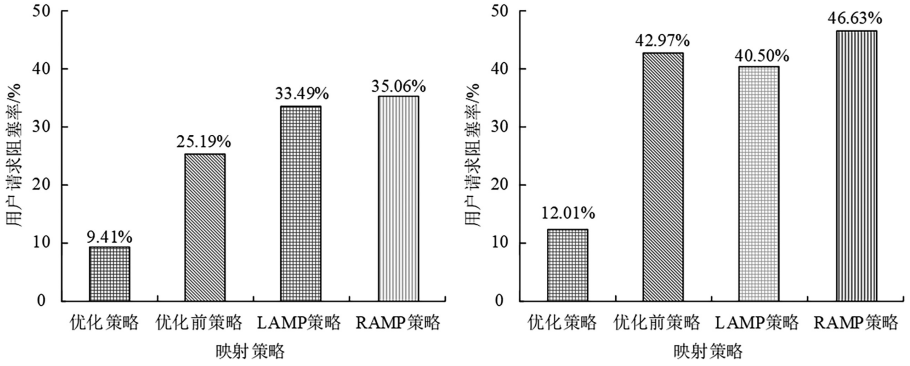


(b) 训练集中预测值与实际值的对比

图 3 基于 GAT 请求预测模型的预测值与实际值对比曲线

Fig. 3 Comparison curve of predicted and actual values based on the GAT request prediction model

为了验证基于 GAT 的算力资源调配优化策略的性能,研究将该策略与其他先进的映射策略进行了对比分析。对比策略包括 GAT 优化前的初始策略、局部性感知内存放置(locality-aware memory placement,LAMP)策略和资源感知内存放置(resource-aware memory placement,RAMP)策略等[17]。不同映射策略的网络阻塞率对比如图 4 所示。从图 4(a)可以看出,在 PEMS-BAY 数据集上,研究的资源调配优化策略的网络阻塞率明显低于其他 3 种策略,阻塞率为 9.41%。相比之下,优化前策略、LAMP 策略和 RAMP 策略的阻塞率分别为 25.19%、33.49%和 35.06%。研究的资源调配优化策略的阻塞率相较于这 3 种策略,分别降低了 15.78%、24.08%和 25.65%。从图 4(b)可以看出,在 Cityscapes 数据集上,研究的资源调配优化策略的阻塞率为 12.01%,相较于其他 3 种策略的阻塞率,分别降低了 30.96%、28.49%和 34.62%。综上所述,基于 GAT 的算力资源调配优化策略有效降低了网络堵塞率。

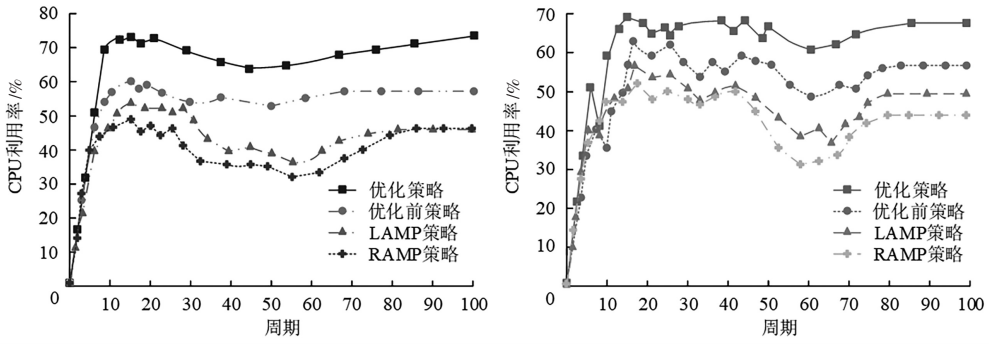


(a) PEMS-BAY 数据集下的用户请求阻塞率对比　(b) Cityscapes 数据集下的用户请求阻塞率对比

图 4　不同映射策略的网络阻塞率对比

Fig. 4　Comparison of network blocking rates for different mapping strategies

为了进一步验证该策略的性能,研究将该策略与其他策略的 CPU 利用率进行了对比验证。不同映射策略的 CPU 利用率对比如图 5 所示。从图 5(a)中可以看出,PEMS-BAY 数据集中,研究的算力资源调配优化策略的 CPU 利用率最高达到了 73.54%,相比优化前策略、LAMP 策略和 RAMP 策略的 60.97%、54.41%和 49.78%,分别提高了 12.57%、19.13%和 23.76%。从图 5(b)中可以看出,在 Cityscapes 数据集中,研究的算力资源调配优化策略的 CPU 利用率最高达到了 68.72%,相比优化前策略、LAMP 策略和 RAMP 策略的 63.01%、56.54%和 51.99%,分别提高了 5.71%、12.18%和 16.73%。综上所述,基于图注意力网络的算力资源调配优化策略有效提高了 CPU 利用率。



(a) 在 PEMS-BAY 数据集下的 CPU 利用率对比　(b) 在 Cityscapes 数据集下的 CPU 利用率对比

图 5　不同映射策略的 CPU 利用率对比

Fig. 5　Comparison of CPU utilization for different mapping strategies

为了全面验证基于 GAT 的算力资源调配优化策略的性能,研究对比分析了不同策略的内存利用率如图 6 所示。从图 6 中可以看出,基于 GAT 的算力资源调配优化策略在各个层级节点的内存利用率上表现非常均衡,且利用率均高于其他 3 种策略。研究的优化策略在 1 级到 4 级节点的内存利用率分别为 62.64%、61.83%、65.04%和 63.47%,节点之间的最大差值仅为 3.21%。相比之下,优化前策略在 4 级节点和 2 级节点上的内存利用率分别为 60.49%和 31.94%,差值达到了 28.55%。LAMP 策略在 1 级节点和 3 级节点上的内存利用率分别为 43.15%和 26.41%,差值为 16.74%。RAMP 策略在 4 级节点和 3 级节点上的内存利用率分别为 43.64%和 25.37%,差值为 18.27%。此外,研究的算力资源调配优化策略的最大内存利用率相较于其他策略的最大内存利用率,分别提高了

4.55%、21.89%和21.40%。因此综上可以看出,基于图注意力网络的算力资源调配优化策略不仅显著提高了内存利用率,还使得算力资源分配更加均衡。

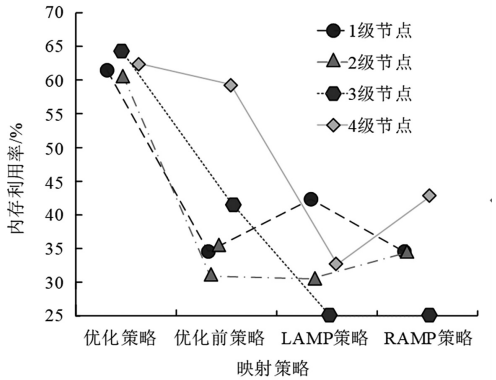


图6 不同映射策略的内存利用率对比

Fig. 6 Comparison of memory utilization for different mapping strategies

3 结论

为了满足用户对服务质量的要求,提高网络资源的利用率,研究将VNF和SFC相结合进行了用户请求的链路映射。同时,为了降低网络堵塞率并提高算力资源调配的均衡性,研究利用GAT的结构构建了用户请求预测模型。结果表明,研究策略的CPU利用率最高达到了73.54%,相较于其他3种策略分别提高了12.57%、19.13%和23.76%。从内存利用率来看,在PEMS-BAY数据集中,研究策略的内存利用率最大值为65.04%,相较于其他3种策略的最大利用率,分别提高了4.55%、21.89%和21.40%。在资源调配均衡度方面,研究策略的内存利用率差值为3.21%。综上可以看出,基于图注意力网络的算力资源调配优化策略研究有效提高了网络资源的利用率和均衡性。但实际网络请求场景复杂多变,研究未能涵盖所有网络场景来进行分析和讨论,因此研究结果不够全面,这方面还待进一步完善。

参考文献

[1]杨明烜, 洪学海, 唐宏伟. 基于任务资源需求预测的人工智能算力调度[J]. 高技术通讯, 2024, 34(5): 475-485.
[2]祝淑琼, 徐青青, 李小涛, 等. 算力度量与任务调度:物联网端侧设备策略研究[J]. 电信科学, 2024, 40(4): 122-138.
[3]陈晓红, 曹廖滢, 陈姣龙, 等. 我国算力发展的需求、电力能耗及绿色低碳转型对策[J]. 中国科学院院刊, 2024, 39(3): 528-539.
[4]喻少如, 唐成余. 迈向智能监察:人工智能赋能国家监察的逻辑与进路[J]. 中共天津市委党校学报, 2024, 26(2): 45-54.
[5]王素红, 唐煜星, 郭文豪, 等. 考虑优先级的智能电网业务调度与资源分配方案[J]. 南方电网技术, 2024, 18(4): 59-70,79.
[6]吕桐弢. 数字经济视域下算力盗用的双维风险及法治应对[J]. 山东师范大学学报(社会科学版), 2023, 68(5): 114-124.
[7]孙毅, 王会梅, 鲜明, 等. Kubeflow异构算力调度策略研究[J]. 计算机工程, 2024, 50(2): 25-32.
[8]张焘, 许长桥, 连一博, 等. 基于深度强化学习的算力网络主动防御方法[J]. 中国科学:信息科学, 2023, 53(12): 2372-2385.
[9]王晓辉, 张颉, 季知祥. 存在依赖关系的边缘计算多任务调度策略研究[J]. 自动化技术与应用, 2025, 44(5):28-32,56.
[10]衷璐洁, 王目. 区块链赋能的算力网络协同资源调度方法[J]. 计算机研究与发展, 2023, 60(4): 750-762.
[11]柴若楠, 郜帅, 兰江雨, 等. 算力网络中高效算力资源度量方法[J]. 计算机研究与发展, 2023, 60(4): 763-771.
[12]LI M, CHENG N, GAO J, et al. Energy-efficient UAV-assisted mobile edge computing: resource allocation and trajectory optimization [J]. IEEE Transactions on Vehicular Technology, 2020, 69(3): 3424-3438.
[13]王昊, 李晖, 宋端正, 等. 面向云-雾计算系统中的遗传算法任务调度研究[J]. 电子测量与仪器学报, 2023, 37 (8): 40-51.
[14]李越鳌, 彭业顺, 陆伟继, 等. 基于5G高精度融合定位技术的研究[J]. 环境技术,2024, 42(3):143-151.
[15]丁巧宜, 王梓耀, 潘振宁, 等. 面向电量-调频-容量市场的数据中心园区算力及电力资源规划 [J]. 电力系统自动化, 2024, 48 (1): 59-66.
[16]王永强, 李子龙, 王杜鑫, 等. 基于数据挖掘的光纤通信网络异常数据检测[J]. 自动化技术与应用,2024, 43(11):111-114.
[17]白露. 基于IP集群数据挖掘的网络行为异常检测系统[J]. 微型电脑应用,2023, 39(6):153-155.