基于强化学习的大规模多模Mesh网络联合路由选择及资源调度算法

朱晓荣* 贺楚闳

(南京邮电大学通信与信息工程学院 南京 210003)

摘 要：为了平衡新型电力系统中大规模多模Mesh网络的传输可靠性和效率，该文在对优化问题进行描述和分析的基础上提出一种基于强化学习的大规模多模Mesh网络联合路由选择及资源调度算法，分为两个阶段。在第1阶段中，根据网络拓扑结构信息和业务需求，利用一种多条最短路径路由算法，输出所有最短路径。在第2阶段中，提出一种基于多臂老虎机(MAB)的资源调度算法，该算法基于得到的最短路径集合构建MAB的摇臂，然后根据业务需求计算回报，最终给出最优的路由选择及资源调度方式用于业务传输。仿真结果表明，所提算法能够满足不同的业务传输需求，实现端到端路径的平均时延和平均传输成功率的高效平衡。

关键词：Mesh网络；路由选择；资源调度；多臂老虎机；强化学习

中图分类号：TN915.85 **文献标识码**：A **文章编号**：1009-5896(2022)YU-0001-10

DOI: 10.11999/JEIT231103

Joint Routing and Resource Scheduling Algorithm for Large-scale Multi-mode Mesh Networks Based on Reinforcement Learning

ZHU Xiaorong HE Chuhong

(*College of Telecommunication and Information Engineering*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210003, *China*)

Abstract: In order to balance the transmission reliability and efficiency of large-scale multi-mode mesh networks in the new power system, a two-stage algorithm is proposed based on reinforcement learning for joint routing selection and resource scheduling in large-scale multi-mode mesh networks, building upon the description and analysis of optimization problems. In the first stage, based on the network topology information and service requirements, a multi shortest path routing algorithm is utilized to generate all the shortest paths. In the second stage, a resource scheduling algorithm based on Multi-Armed Bandit (MAB) is proposed. The algorithm constructs the arms of the MAB based on the obtained set of shortest paths, then calculates the reward according to the service demands, and finally gives the optimal route selection and resource scheduling mode for service transmission. Simulation results show that the proposed algorithm can meet different service transmission requirements, and achieve an efficient balance between the average end-to-end path delay and the average transmission success rate.

Key words: Mesh networks; Routing selection; Resource scheduling; Multi-Armed Bandit (MAB); Reinforcement learning

1 引言

在通信环境复杂、设备种类繁多、接入形式多样的配电物联网场景中[1]，随着分布式电源大规模并网，电网调度运行模式向源网荷储协调控制、输配微网多级协同转变，催生出分布式光伏调控、精准负荷控制、配网保护等一系列新型电力网络业务应用。面对“新能源、新业务”大规模接入、“响应”由骨干向末梢延伸等业务应用新特点，新型电力系统需要构建调用灵活、按需互联的通信网络，优化配电网调控管理。然而，现存的电力线载波(Power Line Carrier, PLC)通信网络难以有效应对配电台区海量设备接入、广域覆盖与多样化业务应用需求等挑战。因此，通过结合成本低、扩展性好且覆盖范围广的无线Mesh网络技术[2]，可以构建融合PLC通信与无线射频(Radio Frequency, RF)通信的网络，以解决新型配电物联网的接入与传输问题。这一双介质融合的大规模Mesh网络不仅能够

收稿日期：2023-10-10；改回日期：2024-02-04；网络出版：2024-02-26
*通信作者： 朱晓荣 xrzhu@njupt.edu.cn
基金项目：国家自然科学基金(92367102, 92067101)，江苏省重点研发计划(BE2021013-3)

Foundation Items: The National Natural Science Foundation of China (92367102, 92067101), The Key R&D Plan of Jiangsu Province (BE2021013-3)

提供广泛的通信服务覆盖，还可以支持多种不同类型的设备接入，有助于打破配电物联网场景中碎片化、差异化的通信壁垒，增强网络抗毁性，提高网络传输可靠性，构建智能配电物联网的“最后1公里”网络。

在面向新型电力系统的大规模多模Mesh网络中，需要灵活分配有限的网络资源，包括时域资源、频域资源和空间资源等，以提升网络性能，其中也涉及到路由选择和相关链路资源的调度。传统Mesh网络主要采用单路径路由算法[3]，可能导致某些链路因过度使用而造成网络拥塞。为解决此问题，一些多路径路由协议被提出[4–6]。然而，这些路由算法通常基于简化的数学模型，难以适应实际复杂动态的配电物联网环境和多样化业务需求，从而限制了它们在路由调度方面的高效性和可靠性。

近年来，强化学习算法已经被逐渐应用于各类Mesh网络的研究中。基于强化学习的方法可以根据实时网络环境和业务需求进行动态决策，为网络资源优化提供智能解决方案[7]。为应对水下多跳无线传感器网络信道的动态性，文献[8]提出了一种基于分布式强化学习框架的数据转发方案。文献[9]提出了一种自适应深度强化学习方法用于无线Mesh网络的通信流量控制，可以在模型训练中改变网络的数据传输拓扑和节点分簇方式。然而，其并未考虑到数据传输时调制方式选择的优化以及新型电力系统通信网络中通信方式多样化的特点。面向低功耗蓝牙Mesh网络场景，文献[10]提出了一种将邻近感知信息转发到远程服务器的Q-learning算法以最小化端到端数据包传输时延。但是，该方法仅从每个蓝牙信标自身出发考虑是否需要广播转发数据包，而未从全局视角进行网络多维度的优化。

当前已有的Mesh网络资源联合优化研究主要集中在路由选择与带宽分配联合优化、带宽预留与信道分配联合优化等方面[11,12]，其优化设计注重数据量大小，未充分考虑不同数据流的时延、可靠性等差异化需求。此外，目前的无线Mesh网络负载均衡研究[13–16]主要关注端到端时延、信噪比和节点能量等方面，通过优化路由选择或路由器位置来提升网络性能。然而，它们侧重于无线信道的容量等条件，忽略了链路多样化传输介质可能带来的影响。新型电力系统中的大规模多模Mesh网络不同于传统无线Mesh网络，呈现出复杂且融合的特性。与仅考虑RF通信的传统网络相比，新型电力系统通信网络需综合考虑RF和PLC的传输特性。同时，新型电力系统通信网络中节点类型更加多样，通信方式和能力差异明显，需灵活调整负载均衡策略以平衡网络性能。此外，传统Mesh网络负载均衡主要关注每个业务的数据量大小，而新型电力系统通信网络的业务需求差异显著，负载均衡需全面考虑各项业务指标。综上所述，目前大部分研究提出的优化方法难以满足新型电力系统通信网络的需要。本文主要的研究工作如下：

(1) 针对新型电力系统中的多模Mesh网络，同时关注RF链路和PLC链路资源，并充分考虑每个业务的数据量、传输时延和传输成功率等差异化需求，对每个业务传输选择的路径及路径上每一跳的传输介质和数据调制方式进行联合优化，形成一个以成本最小化为目标的大规模0-1整数线性规划问题。

(2) 针对新型电力系统通信网络中通信方式和数据传输速率多样化的特点，提出一种基于强化学习的联合路由选择及资源调度算法。本算法首先基于网络拓扑结构信息，通过改进的宽度优先搜索算法和多条最短路径输出算法，获取所有最短路径。

(3) 基于获得的最短路径集合，提出一种基于多臂老虎机(Multi-Armed Bandit, MAB)的资源调度算法。本算法从全局角度出发，将所有可能的资源调度方式视为MAB的摇臂集合，根据特定电力网络业务需求计算执行摇臂的回报。最终，本算法能够给出业务最优的路径选择及资源调度方式。

(4) 仿真评估了在不同节点数据包发送速率和网络规模下的算法性能，验证了所提算法能够有效利用新型电力系统通信网络的多模通信和多数据调制方式的优势来满足多样化业务需求。结果表明，所提算法在网络平均端到端时延和平均端到端传输成功率方面均表现良好。

2 系统模型与问题描述

2.1 网络模型

本文考虑的大规模多模Mesh网络模型如图1所示，其中节点分为中心主控节点、路由汇聚节点和末端感知节点3类。这3类节点之间的多样化通信方式体现了其独特的“多模”特性，它们包括双模通信(PLC链路l_{ij}^{PLC}与RF链路l_{ij}^{RF}共存)、单模通信(仅PLC链路l_{ij}^{PLC}或RF链路l_{ij}^{RF})与低功耗无线通信(如ZigBee、蓝牙和Wi-Fi等低功耗RF链路$l_{ij}^{\text{低功耗RF}}$)。具体来说，中心主控节点是一个边缘网关，提供连接到电力光纤、无线专网及公网的接口，负责本网络的生命周期管理、路由维护和资源分配等。路由汇聚节点由常供电的路由器构成，配置双模通信或单模通信方式，能够覆盖整个网络，负责本节点的数据传输及其子节点的数据转发。末端感知节点由各类低功耗设备构成，如用户电表、电缆感知设备

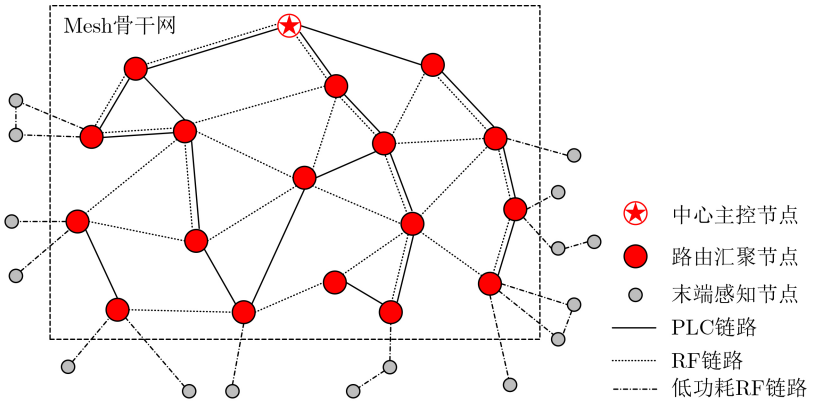


图 1 面向新型电力系统的大规模多模Mesh网络模型

和智能环境传感器等，可采集各类信息并以低功耗无线通信方式传输给路由汇聚节点。此外，唯一的中心主控节点和所有的路由汇聚节点共同组成了Mesh骨干网，本文将其建模为无向图$\mathcal{G}=(\mathcal{N},\mathcal{L})$。其中$\mathcal{N}=\{1,2,\cdots,N\}$和$\mathcal{L}=\{l_{ij}\},i,j\in\mathcal{N}$分别表示网络节点集合和物理链路集合。进一步地，分别用$\mathcal{L}^{\mathrm{RF}}=\{l_{ij}^{\mathrm{RF}}\}$和$\mathcal{L}^{\mathrm{PLC}}=\{l_{ij}^{\mathrm{PLC}}\}$表示RF链路集合和PLC链路集合，其中每个元素$l_{ij}^{\mathrm{RF}}$和$l_{ij}^{\mathrm{PLC}}$都是0-1变量，用来指示节点$i$和节点$j$之间是否存在RF链路和PLC链路，如果是则对应变量为1，否则为0。接下来，本文将专注于研究如何在Mesh骨干网中为每个业务进行合适的路由选择及资源调度，以提高整个网络的传输效率和可靠性。

2.2 通信模型

本文用$\mathcal{E}=\{1,\cdots,e,\cdots,E\}$表示业务集合，并用5元组$\{n_{\mathrm{s}}^e,n_{\mathrm{d}}^e,D^e,T^e,S^e\}$表示每个业务$e$的差异化需求。其中，$n_{\mathrm{s}}^e$和$n_{\mathrm{d}}^e$分别表示业务$e$的源节点和目的节点，$D^e,T^e$和$S^e$分别表示业务$e$的数据量大小、端到端时延要求和端到端传输成功率要求。为了指示业务e是否使用链路l_{ij}进行传输，定义了一个向量$\boldsymbol{y}_{ij}^e=[y_{ij,\mathrm{RF}}^e,y_{ij,\mathrm{PLC}}^e]$。其中，0-1变量$y_{ij,\mathrm{RF}}^e$和$y_{ij,\mathrm{PLC}}^e$分别用来指示业务$e$是否使用RF链路$l_{ij}^{\mathrm{RF}}$和PLC链路$l_{ij}^{\mathrm{PLC}}$进行传输，如果是则对应变量为1，否则为0。与此同时，还定义了一个向量$\boldsymbol{M}_{ij}^e=[M_{ij,2}^e,M_{ij,4}^e,M_{ij,16}^e]$用来指示业务$e$使用链路$l_{ij}$进行传输时的数据调制方式选择。其中，0-1变量$M_{ij,2}^e,M_{ij,4}^e$和$M_{ij,16}^e$分别用来指示业务$e$是否使用二进制相移键控(Binary Phase Shift Keying, BPSK)、正交相移键控(Quadrature Phase Shift Keying, QPSK)和16进制正交幅度调制(16-ary Quadrature Amplitude Modulation, 16QAM)调制方式，如果是则对应变量为1，否则为0。进一步地，业务e在链路l_{ij}上的传输时延$T_{ij}^{e,\mathrm{trans}}$可以表示为

$$\begin{aligned}T_{ij}^{e,\mathrm{trans}}&=\frac{D^e}{\Lambda_{ij}^e}\\&=\frac{D^e}{R_{\mathrm{B}}\cdot M_{ij,2}^e+2R_{\mathrm{B}}\cdot M_{ij,4}^e+4R_{\mathrm{B}}\cdot M_{ij,16}^e}\end{aligned}\tag{1}$$

其中，Λ_{ij}^e为传信率，R_{B}为波特率。此外，业务e到达父节点j等待处理时的排队时延$T_j^{e,\mathrm{wait}}$可以表示为

$$T_j^{e,\mathrm{wait}}=\frac{X_j}{V_j}\tag{2}$$

其中，X_j表示当前父节点j的缓存队列中等待处理的业务量，V_j表示父节点j的业务处理速率。而与传输时延和排队时延相比，业务e在不同链路上转换传输介质与数据调制方式的时延很短，故可以忽略不计。

通常，现实中的误码分布接近泊松分布，因此业务e在传输中出现f个误码的概率可以表示为

$$P_{ij}^e(f)=\frac{(D^e\cdot P_M^e/\log_2 M)^f}{f!}\exp\left(-D^e\cdot\frac{P_M^e}{\log_2 M}\right)\tag{3}$$

其中，P_M^e表示业务e在对应调制阶数M下的误码率。在已知链路信噪比SNR_{ij}的情况下，BPSK, QPSK和16QAM 3种调制方式下的误码率可分别按式(4)—式(6)进行计算

$$\begin{aligned}P_{\mathrm{BPSK}(M=2)}^e&=\frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\mathrm{SNR}_{ij}}{2}}\right)\\&=\frac{1}{\sqrt{\pi}}\int_0^{\sqrt{\frac{\mathrm{SNR}_{ij}}{2}}}e^{-\theta^2}\mathrm{d}\theta\end{aligned}\tag{4}$$

$$\begin{aligned}P_{\mathrm{QPSK}(M=4)}^e&=\frac{1}{2}\mathrm{erfc}\left(\sqrt{\mathrm{SNR}_{ij}}\right)\\&=\frac{1}{\sqrt{\pi}}\int_0^{\sqrt{\mathrm{SNR}_{ij}}}e^{-\theta^2}\mathrm{d}\theta\end{aligned}\tag{5}$$

$$P^e_{16\mathrm{QAM}(M=16)} = \frac{3}{8}\mathrm{erfc}\left(\sqrt{\frac{\mathrm{SNR}_{ij}}{10}}\right) = \frac{3}{4\sqrt{\pi}}\int_0^{\sqrt{\frac{\mathrm{SNR}_{ij}}{10}}} e^{-\theta^2}\mathrm{d}\theta \tag{6}$$

从而，可以推导得到业务e在链路l_{ij}上传输的丢包率PER^e_{ij}和传输成功率S^e_{ij}

$$\mathrm{PER}^e_{ij} = 1 - P^e_{ij}(f=0) = 1 - \exp\left(-D^e \cdot \frac{P^e_M}{\log_2 M}\right) = \begin{cases} \mathrm{PER}^e_{ij,\mathrm{RF}}, \boldsymbol{y}^e_{ij} = [1,0] \\ \mathrm{PER}^e_{ij,\mathrm{PLC}}, \boldsymbol{y}^e_{ij} = [0,1] \end{cases} \tag{7}$$

$$S^e_{ij} = \left[1 - \mathrm{PER}^e_{ij,\mathrm{RF}}\right] \cdot y^e_{ij,\mathrm{RF}} + \left[1 - \mathrm{PER}^e_{ij,\mathrm{PLC}}\right] \cdot y^e_{ij,\mathrm{PLC}} \tag{8}$$

最后，综合考虑传输时延、排队时延和传输成功率，业务e使用链路l_{ij}传输的成本c^e_{ij}可以表示为

$$c^e_{ij} = w_1 \cdot c(T^{e,\mathrm{wait}}_j) + w_2 \cdot c(T^{e,\mathrm{trans}}_{ij}) + w_3 \cdot c(S^e_{ij}) \tag{9}$$

其中，w_1，w_2和w_3是可按需调整的权重系数，且$w_1, w_2, w_3 \in [0,1]$，$w_1 + w_2 + w_3 = 1$。当业务e(如配网保护业务)对时延更加敏感时，可以增加w_1和w_2的比重；而当业务e(如分布式能源调控业务)对可靠性要求更高时，可以增加w_3的比重。此外，归一化成本函数$c(x)$定义为

$$c(x) = \frac{x - x^{\min}}{x^{\max} - x^{\min}} \tag{10}$$

其中，x表示某项指标值，$x^{\max}$和$x^{\min}$分别表示该项指标的最大值和最小值。

2.3 问题形成

本文目标是通过联合优化所有业务在Mesh骨干网中的传输路径、传输介质和数据调制方式来最小化系统成本，此优化问题可以表示为

$$\begin{aligned} &\min_{\{\boldsymbol{y}^e_{ij}, \boldsymbol{M}^e_{ij}\}} \sum_{e\in\mathcal{E}}\sum_{l_{ij}\in\mathcal{L}} c^e_{ij} \\ &\text{s.t. C1}: \sum_{e=1}^{E} y^e_{ij,\mathrm{RF}} \cdot \Lambda^e_{ij} \le C^{\max}_{ij,\mathrm{RF}}, \sum_{e=1}^{E} y^e_{ij,\mathrm{PLC}} \cdot \Lambda^e_{ij} \le C^{\max}_{ij,\mathrm{PLC}}, \forall i,j \in \mathcal{N} \\ &\quad \mathrm{C2}: \ y^e_{ij,\mathrm{RF}} \le l^{\mathrm{RF}}_{ij}, y^e_{ij,\mathrm{PLC}} \le l^{\mathrm{PLC}}_{ij}, \forall i,j \in \mathcal{N}, e \in \mathcal{E} \\ &\quad \mathrm{C3}: y^e_{ij,\mathrm{RF}} + y^e_{ij,\mathrm{PLC}} \le 1, \forall i,j \in \mathcal{N}, e \in \mathcal{E} \\ &\quad \mathrm{C4}: M^e_{ij,2} + M^e_{ij,4} + M^e_{ij,16} \le 1, \forall i,j \in \mathcal{N}, e \in \mathcal{E} \\ &\quad \mathrm{C5}: \sum_{l_{oj}\in\mathcal{L}} \left(y^e_{oj,\mathrm{RF}} + y^e_{oj,\mathrm{PLC}}\right) - \sum_{l_{io}\in\mathcal{L}} \left(y^e_{io,\mathrm{RF}} + y^e_{io,\mathrm{PLC}}\right) = \begin{cases} 1, o = n^e_{\mathrm{s}} \\ -1, o = n^e_{\mathrm{d}} \\ 0, \text{其他} \end{cases}, \forall i,j,o \in \mathcal{N}, e \in \mathcal{E} \\ &\quad \mathrm{C6}: \sum_{l_{ij}\in\mathcal{L}} \left(T^{e,\mathrm{trans}}_{ij} + T^{e,\mathrm{wait}}_j\right) \le T^e, \prod_{l_{ij}\in\mathcal{L}} S^e_{ij} \ge S^e, \forall e \in \mathcal{E} \end{aligned} \tag{11}$$

具体来说，C1用于确保在每条RF链路l^{RF}_{ij}和PLC链路l^{PLC}_{ij}上传输的业务流不会超过各自的最大容量$C^{\max}_{ij,\mathrm{RF}}$和$C^{\max}_{ij,\mathrm{PLC}}$，C2表明业务使用链路$l_{ij}$传输的前提是存在RF链路$l^{\mathrm{RF}}_{ij}$或PLC链路$l^{\mathrm{PLC}}_{ij}$。C3和C4用于确保业务在链路$l_{ij}$上传输时传输介质和数据调制方式选择的唯一性。C5表明除源节点和目的节点外，业务不能在任何节点上产生或终止。C6则是对业务差异化服务质量需求的约束。分析可知，在优化问题中，目标函数以及约束条件C2～C5都是线性的。而对于非线性约束条件C1和C6而言，虽然存在变量相乘的形式，但因为都是0-1变量，所以可以按一定的方式将它们线性化，因而所提优化问题可以视为一个大规模0-1型整数线性规划问题。通常这类问题可以采用穷举法或隐枚举法求解，然而对于大规模Mesh网络来说，用这类传统数学方法求解复杂度较高、运算时间较长，因此需要设计一种更加行之有效的算法。

3 路由选择及资源调度算法

3.1 算法架构

在实际传输中，端到端路径上每增加一跳都会显著增加业务传输时延。因此，必须找到从业务源节点到目的节点的最短路径，以从路由角度减少传输时延。此外，为实现最优传输，考虑到网络中不同链路和节点状况的动态性，需找到的是所有最短路径。在此基础上，进一步考虑选择哪条最短路径及该路径上每一跳的传输介质和数据调制方式。为减少人工设计、提高泛化能力和求解速度，鉴于优化问题中的所有决策变量均为离散值，本文提出了一种基于MAB的联合路由选择及资源调度算法，流程如图2所示。

3.2 多条最短路径路由算法

宽度优先搜索(Breadth First Search, BFS)是求解最短路径的经典算法之一[17]，其缺陷在于只能得到1条最短路径，这可能会造成网络拥塞。为

此，本文采用了一种多条最短路径路由算法，分为改进的宽度优先搜索和多条最短路径输出两部分。

首先需要建立网络拓扑邻接矩阵$\boldsymbol{A}(a_{ij})$，其中每个元素a_{ij}的值定义为

$$a_{ij}=\begin{cases}0,\text{节点}i\text{和}j\text{不相连}\\1,\text{节点}i\text{和}j\text{之间存在RF链路}\\2,\text{节点}i\text{和}j\text{之间存在PLC链路}\\3,\text{节点}i\text{和}j\text{之间存在RF链路和PLC链路}\end{cases},i,j\in\mathcal{N}\tag{12}$$

基于$\boldsymbol{A}$，通过改进的宽度优先搜索算法可以得到最短路径邻接矩阵$\boldsymbol{A}^{\mathrm{s}}$，其具体执行过程如算法1所示。

分析可知，算法1中步骤(2)～(18)需循环$N\times Z$次，步骤(19)需循环N^2次。由于$N^2\geq N\times Z$，因此其时间复杂度为$O(N^2)$。基于$\boldsymbol{A}^{\mathrm{s}}$，接下来通过算法2可以输出所有最短路径。

分析可知，算法2的时间复杂度为$O(v_2)$，其中v_2表示算法2中步骤(2)～(11)的循环次数，其大小与输入的A^{s}，n_{s}^e和n_{d}^e都相关，会随着网络拓扑结构的不同而变化。

3.3 基于MAB的资源调度算法

MAB问题是一种环境状态保持恒定不变的强化学习[18]。结合Path^e和实际网络拓扑结构，本文从第1条最短路径开始依次选择，然后依次遍历所

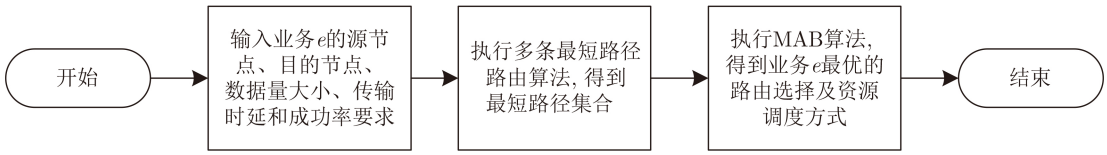


图 2 基于MAB的联合路由选择及资源调度算法流程

算法 1　改进的宽度优先搜索算法

输入：网络拓扑邻接矩阵$\boldsymbol{A}$，源节点n_{s}^e

(1) 将n_{s}^e放入队列Queue中；将n_{s}^e的Visited值设置为True，而其余节点的该值设置为False；将n_{s}^e到源节点的最短跳数$\mathrm{HopCount}(n_{\mathrm{s}}^e)$值设置为0，而其余节点的该值设置为无穷大；将所有节点的Searched值设置为False；将每个节点n的前置节点数量$\mathrm{FrontCount}(n)$值设置为0，表示所有网络节点的前置节点集合$\mathrm{FrontPoint}(n)$中前置节点数量为0

(2) **WHILE** Queue不是空集 **DO**

(3)　获取Queue的队首节点Head，遍历$\boldsymbol{A}$，找出Head的所有邻居节点$\mathrm{Neighbor}_z,z=1,2,\cdots,Z$

(4)　**FOR** $\mathrm{Neighbor}_z,z=1,2,\cdots,Z$ **DO**

(5)　　**IF** $\mathrm{Neighbor}_z$的Searched值为True **THEN** 跳出本次循环

(6)　　**ELSE IF** $\mathrm{Neighbor}_z$的Visited值为False **THEN**

(7)　　令$\mathrm{HopCount}(\mathrm{Neighbor}_z)=\mathrm{HopCount}(\mathrm{Head})+1$；将$\mathrm{FrontCount}(\mathrm{Neighbor}_z)$值加1
　　把Head存入$\mathrm{FrontPoint}(\mathrm{Neighbor}_z)$中；把$\mathrm{Neighbor}_z$存入Queue中，并将其Visited值设置为True

(8)　　**ELSE** $\mathrm{Neighbor}_z$的Visited值为True **THEN**

(9)　　**IF** $\mathrm{HopCount}(\mathrm{Head})+1<\mathrm{HopCount}(\mathrm{Neighbor}_z)$ **THEN**

(10)　　令$\mathrm{HopCount}(\mathrm{Neighbor}_z)=\mathrm{HopCount}(\mathrm{Head})+1$；$\mathrm{FrontCount}(\mathrm{Neighbor}_z)$值保持不变
　　把$\mathrm{FrontPoint}(\mathrm{Neighbor}_z)$中最近一个存入的元素替换为Head

(11)　　**ELSE IF** $\mathrm{HopCount}(\mathrm{Head})+1=\mathrm{HopCount}(\mathrm{Neighbor}_z)$ **THEN**

(12)　　将$\mathrm{FrontCount}(\mathrm{Neighbor}_z)$值加1；把Head存入$\mathrm{FrontPoint}(\mathrm{Neighbor}_z)$中

(13)　　**ELSE** 跳出本次循环

(14)　　**END IF**

(15)　**END IF**

(16)　**END FOR**

(17) 将当前Head的Searched值设置为True，并将其移出Queue

(18) **END WHILE**

(19) 按如下方式遍历设置$\boldsymbol{A}^{\mathrm{s}}(a_{ij}^{\mathrm{s}})$中每个元素$a_{ij}^{\mathrm{s}}$的值：$a_{ij}^{\mathrm{s}}=\begin{cases}1,j\in\mathrm{FrontPoint}(i)\\0,\text{其他}\end{cases},i,j\in\mathcal{N}$

输出：最短路径邻接矩阵$\boldsymbol{A}^{\mathrm{s}}$

选路径中每一跳可以选择的传输介质和数据调制方式来构造摇臂，形成了一个MAB问题。摇臂集合Arm^e和其中每一个摇臂$\mathbf{arm}_k^e$可以表示为

$$\begin{aligned}\mathrm{Arm}^e &= \{\mathbf{arm}_1^e,\cdots,\mathbf{arm}_k^e,\cdots,\mathbf{arm}_K^e\},\mathbf{arm}_k^e \\ &= [p_u^e,(a_u^1,b_u^1),\cdots,(a_u^h,b_u^h),\cdots,(a_u^H,b_u^H)]\end{aligned} \tag{13}$$

其中，$p_u^e,u=1,2,\cdots,U$表示$\mathbf{arm}_k^e$选择了第u条最短路径；a_u^h和$b_u^h,h=1,2,\cdots,H$是$\mathbf{arm}_k^e$为第u条路径上第h跳选择的传输介质和数据调制方式。其中，$a_u^h=1$或2表示选择的是RF链路或PLC链路，$b_u^h=0,1$或2表示选择的是BPSK, QPSK或16-QAM调制。随着n_{s}^e和n_{d}^e的不同，最短路径数量U和最短路径跳数H会变化，结合实际网络拓扑结构，所构造的Arm^e也会相应不同。在此举例说明：假设某个业务e的$n_{\mathrm{s}}^e=141,n_{\mathrm{d}}^e=1$，基于给定的网络拓扑结构，通过算法1和算法2得到$\mathrm{Path}^e=\{p_1^e,p_2^e,p_3^e,p_4^e\}$，其中每条最短路径如图3所示，图中实线和虚线分别表示PLC链路和RF链路。分析可知，此时$U=4,H=4$。对于p_1^e,p_2^e,p_3^e和p_4^e来说，可以构造的摇臂数量分别为$3\times3\times3\times3=81$, $3\times(2\times3)\times3\times3=162$, $3\times3\times3\times3=81$ 和$3\times(2\times3)\times(2\times3)\times3=324$，从而$\mathrm{Arm}^e$的大小$K=81+162+81+324=648$。

在每次迭代中，需要从K个摇臂中选择1个$\mathbf{arm}_k^e$执行。利用第2节中定义的传输时延、排队时延和传输成功率等公式，可以计算出$\mathbf{arm}_k^e$对应路径的端到端时延和传输成功率，再与给定的业务端到端时延要求以及传输成功率要求作差，即可得到回报$R(\mathbf{arm}_k^e)$

$$\begin{aligned}R(\mathbf{arm}_k^e) = &-\alpha\cdot\left|T^e-\left(\sum_{j\in p_u^e}T_j^{e,\mathrm{wait}}+\sum_{l_{ij}\in p_u^e}T_{ij}^{e,\mathrm{trans}}\right)\right| \\ &-\beta\cdot\left|\prod_{l_{ij}\in p_u^e}S_{ij}^e-S^e\right|\end{aligned} \tag{14}$$

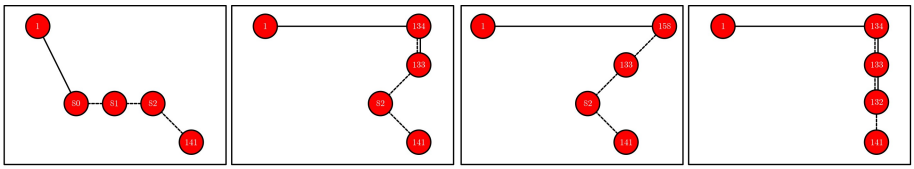
其中，α和β分别表示时延和成功率的重要程度，可按需调整，且$\alpha,\ \beta\in[0,1],\alpha+\beta=1$。此外，本文采用如式(15)的增量形式对$\mathbf{arm}_k^e$在当前迭代次数$t$时的回报期望估值$Q_t(\mathbf{arm}_k^e)$进行高效地更新

$$\begin{aligned}Q_t(\mathbf{arm}_k^e) = &Q_{t-1}(\mathbf{arm}_k^e)+\frac{1}{\mathrm{count}(\mathbf{arm}_k^e)} \\ &\cdot(R_t(\mathbf{arm}_k^e)-Q_{t-1}(\mathbf{arm}_k^e))\end{aligned} \tag{15}$$

其中，$\mathrm{count}(\mathrm{arm}_k^e)$表示$\mathrm{arm}_k^e$的执行次数。

最后，本文采用衰减的ϵ-贪心策略选取摇臂$\mathbf{arm}_k^e$，探索概率$\epsilon(t)$参考文献[19]中的方式进行更新

$$\epsilon(t)=\epsilon_{\mathrm{init}}\times(1-\epsilon_{\mathrm{init}})^{\frac{t}{\chi\times K}} \tag{16}$$



(a) p_1^e (b) p_2^e (c) p_3^e (d) p_4^e

图 3 最短路径集合$\mathrm{Path}^e=\{p_1^e,p_2^e,p_3^e,p_4^e\}$示例

算法 2 多条最短路径输出算法

输入：最短路径邻接矩阵$\boldsymbol{A}^s$，源节点n_{s}^e和目的节点n_{d}^e

(1) 将n_{s}^e压入主栈$\mathrm{Stack}_{\mathrm{main}}$；遍历$\boldsymbol{A}^s$，将$n_{\mathrm{s}}^e$的邻居节点存入邻居节点列表Array，然后将Array作为栈顶压入辅栈$\mathrm{Stack}_{\mathrm{assist}}$

(2) **WHILE** $\mathrm{Stack}_{\mathrm{main}}$不是空集 **DO**

(3) 获取$\mathrm{Stack}_{\mathrm{assist}}$栈顶，作为新的Array

(4) **IF** Array非空 **THEN**

(5) 获取Array中的首个元素，将其压入$\mathrm{Stack}_{\mathrm{main}}$，并将剩余元素构成的列表重新压入$\mathrm{Stack}_{\mathrm{assist}}$

(6) 查询栈顶元素的Array，将$\mathrm{Stack}_{\mathrm{main}}$中包含的元素从其中剔除，再将其压入$\mathrm{Stack}_{\mathrm{assist}}$

(7) **ELSE** 将$\mathrm{Stack}_{\mathrm{main}}$和$\mathrm{Stack}_{\mathrm{assist}}$的栈顶元素弹出

(8) **END IF**

(9) **IF** $\mathrm{Stack}_{\mathrm{main}}$的栈顶元素与$n_{\mathrm{d}}^e$相等 **THEN** 将最短路径$p_u^e=\mathrm{Stack}_{\mathrm{main}}$存入$\mathrm{Path}^e$；将$\mathrm{Stack}_{\mathrm{main}}$和$\mathrm{Stack}_{\mathrm{assist}}$的栈顶元素弹出

(10) **END IF**

(11) **END WHILE**

输出：最短路径集合$\mathrm{Path}^e=\{p_1^e,p_2^e,\cdots,p_U^e\}$

其中，ϵ_{init}表示$\epsilon(t)$的初始值，χ表示衰减系数。综上，具体流程见算法3。

在算法3中，步骤2～6需循环$T_{\max}$次，因此其时间复杂度为$O(T_{\max})$。分析可知，每次迭代选择执行的摇臂与$\epsilon(t)$密切相关。根据式(16)，随着迭代次数t的增加，$\epsilon(t)$逐渐减小，减小速率逐渐放缓。算法3由最初的偏向探索(选择新的摇臂执行)逐渐过渡为偏向利用(选择已知的最优摇臂执行)，但仍会以较小的概率尝试新的摇臂。因此，当$T_{\max}$远大于K时，基本上能够遍历所有摇臂并找到最优摇臂。

4　仿真结果与分析

本文基于NetLogo 6.3和Python 3.6进行仿真，考虑了一个基于真实网络环境获得的150 m × 150 m的仿真网络场景，包含$N=186$个节点，具体网络拓扑结构如图4所示。由于本文主要研究Mesh骨干网的性能，所以未考虑末端感知节点的分布情况，其他仿真参数设置如表1所示。此外，本文考虑了新型电力系统中的一种控制类业务e'，其需求如下：$n_{\text{s}}^{e'}=0$(中心主控节点)，$n_{\text{d}}^{e'}=181$(181号路由汇聚节点)，$T^{e'}=20$ ms, $S^{e'}=99.999\%$。执行完算法1和算法2后，可以得到$U=7$, $H=4$的$\text{Path}^{e'}$，相应可以构建得到$K=972$种资源调度方式。基于构造的摇臂集合，MAB算法的执行过程如图5所示。图5(a)和图5(b)分别展示了随着迭代次数的增加资源调度方式选择和相应平均回报值的变化，其中平均回报值是在每10次迭代后求算数平均的结果。可以看出，在大约1 300次迭代之前，由于算法倾向于探索，其主要随机选择动作或选择某种次优的动作执行，导致平均回报值在较低的范围内波动且波动幅度较大。而在大约1 300次迭代之后，随着

算法 3　基于MAB的资源调度选择算法

输入：摇臂集合Arm^e，最大迭代次数$T_{\max}$

(1) 令$Q(\mathbf{arm}_k^e)=0$,$\text{count}(\mathbf{arm}_k^e)=0$

(2) **FOR** $t=1,2,\cdots,T_{max}$ **DO**

(3)　根据式(16)更新$\epsilon(t)$，然后按如下方式选择摇臂arm_k^e：

$$\text{arm}_k^e=\begin{cases}\text{从}\mathbf{arm}_1^e,\mathbf{arm}_2^e,\cdots,\mathbf{arm}_K^e\text{中以均匀分布随机选取，以}\epsilon(t)\text{的概率}\\ \underset{\mathbf{arm}_k^e}{\mathbf{argmax}}Q_t(\mathbf{arm}_k^e)\text{，以}1-\epsilon(t)\text{的概率}\end{cases}$$

(4)　令$\text{count}(\mathbf{arm}_k^e)=\text{count}(\mathbf{arm}_k^e)+1$

(5)　根据式(14)计算$R_t(\mathbf{arm}_k^e)$，然后根据式(15)更新$Q_t(\mathbf{arm}_k^e)$

(6) **END FOR**

输出：最佳摇臂$\mathbf{arm}_{\text{best}}^e=\underset{\mathbf{arm}_k^e}{\text{argmax}}\,Q(\mathbf{arm}_k^e)$

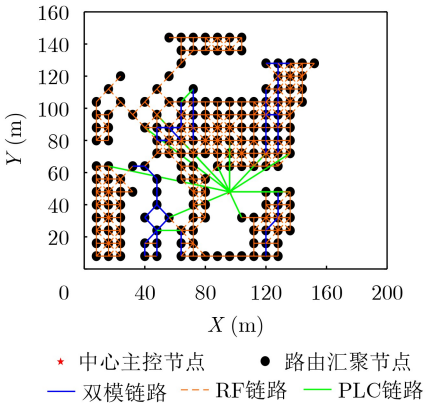
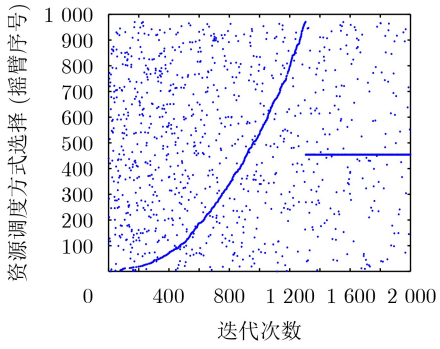
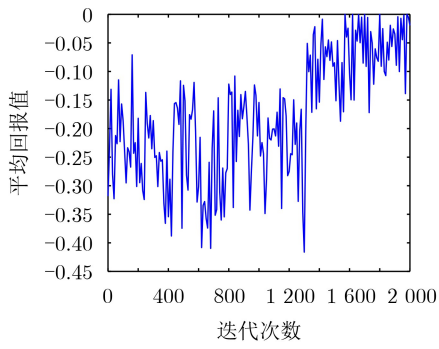


图 4 仿真网络拓扑结构



(a) 资源调度方式选择散点图



(b) 平均回报变化曲线

图 5 MAB算法执行过程

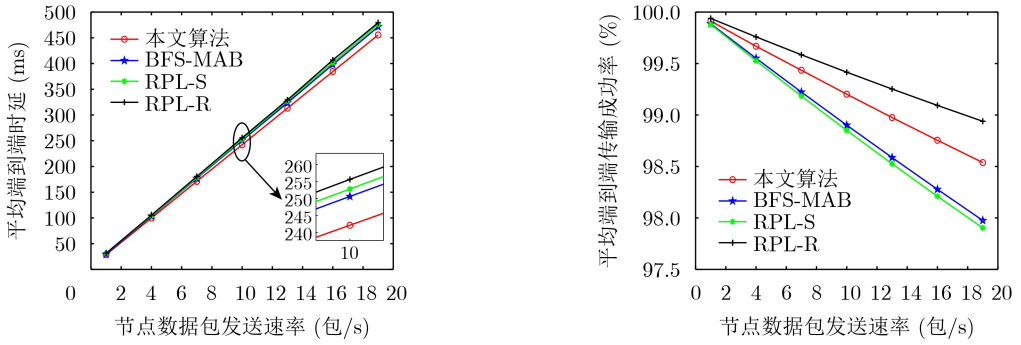
表 1　相关仿真参数设置

参数	最大迭代次数$T_{\max}$	探索概率$\epsilon(t)$初始值ϵ_{init}	衰减系数χ	业务包数据量D^e	传码率R_{B}	路由汇聚节点排队时延$T_j^{e,\text{wait}}$	时延重要程度α，成功率重要程度β
数值	$2\times K$	0.99	6	600 bit	115 200	1～3 ms的随机值	0.5

$\epsilon(t)$逐渐减小，算法倾向于利用，在每次迭代中高概率执行最优的动作，使得平均回报值明显上升，并保持在较高的范围内以较小的幅度波动。

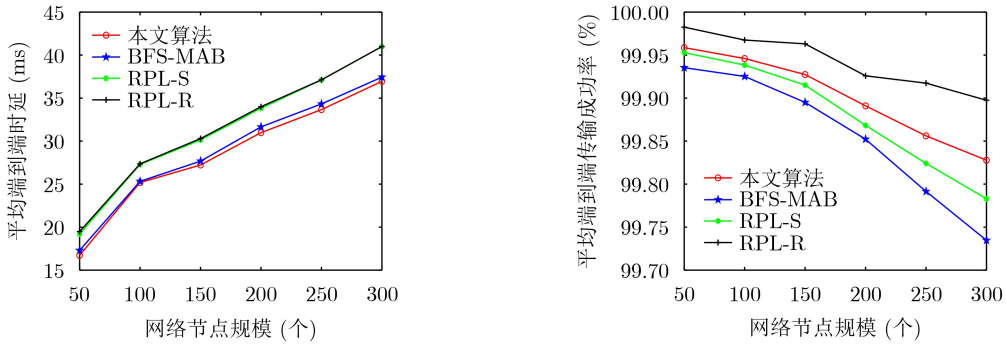
接下来，选用3种算法作为基准与本文算法进行性能比较。其中两种算法基于目前Mesh网络广泛使用的低功耗和有损网络路由协议(Routing Protocol for Low power and lossy networks, RPL)[20]。根据指标不同，分为最短跳数RPL算法(RPL-Shortest hop count, RPL-S)和最高可靠性RPL算法(RPL-highest Reliability, RPL-R)，并假设业务在PLC链路和RF链路传输时分别使用QPSK调制和BPSK调制。第3种宽度优先搜索-多臂老虎机算法(Breadth First Search-Multi-Armed Bandit, BFS-MAB)利用BFS算法仅寻找1条最短路径，然后基于本文提出的MAB算法进行资源调度。在图6中，对比了4种算法在不同传输业务数量下的端到端路径的平均时延和成功率。可以看出，随着网络中路由汇聚节点所需传输业务数量的增加，4种算法的平均端到端时延和平均端到端传输成功率分别呈上升和下降趋势。主要原因在于，随着节点传输业务数量的增加，业务因网络逐渐出现拥塞需排队等待传输，且它们在每一段点到点链路上传输产生误码或传输冲突的可能性也会提升。此外，尽管本文算法的平均端到端传输成功率低于RPL-R算法，但本文算法综合考虑了业务流量和网络链路状况进行灵活的路由选择及资源调度，通过牺牲一小部分传输成功率换取了比其他3种算法更低的平均端到端时延，因此是值得的。由于4种算法的平均端到端时延随着节点传输业务数量的增加从28～32 ms上升至455～480 ms，时延范围跨度较大，因此在图6(a)中不同算法的时延差异看起来并不明显，特别是在数据包发送速率较低时。为了更清晰地呈现差异，本文以10个业务包/s下的4种算法的平均端到端时延为例，在图中做局部放大展示。可以看到，此时RPL-R算法的平均端到端时延最大，为255.645 2 ms，而本文算法的平均端到端时延最小，为242.136 9 ms。

此外，本文还基于随机网络模型，在6种不同的网络节点规模(网络中路由汇聚节点数量)下对比了四种算法的性能，如图7所示。可以看出，随着网络中路由汇聚节点数量的增加，4种算法的平均端到端时延均呈上升趋势，而平均端到端传输成功率逐渐降低。这主要是由于随着网络规模的扩大，业务从源节点到目的节点需要经历的跳数逐渐增加，从而在端到端传输过程中出现误码的概率提升。此外，基于MAB的两种算法的平均端到端时



(a) 平均端到端时延变化曲线　(b) 平均端到端传输成功率变化曲线

图 6 随路由汇聚节点传输业务数量不同变化曲线



(a) 平均端到端时延变化曲线　(b) 平均端到端传输成功率变化曲线

图 7 随网络节点规模不同变化曲线

延明显小于另外两种算法，且本文算法与RPL-R算法的平均传输成功率最接近，可见本文算法能够在不同的网络节点规模下同时保持较高的平均传输成功率和最短的平均时延。

5　结论

本文研究了新型电力系统中大规模多模Mesh网络的传输效率和可靠性均衡问题，提出了一种基于强化学习的大规模多模Mesh网络联合路由选择及资源调度算法。具体而言，本文所提算法由多条最短路径路由算法和基于MAB的资源调度算法组成，能够在找到给定业务的所有最短路径的基础上，构建MAB相应的摇臂集合，并为特定业务匹配到最适合的路由选择及资源调度方式。仿真结果表明，在复杂动态的大规模多模Mesh网络中，本文所提算法能够实现端到端路径的平均时延和平均传输成功率的高效平衡，性能优于多种传统算法，能够在保证网络可靠性的前提下，最大化网络的传输效率。

参 考 文 献

[1] BEDI G, VENAYAGAMOORTHY G K, SINGH R, *et al*. Review of Internet of Things (IoT) in electric power and energy systems[J]. *IEEE Internet of Things Journal*, 2018, 5(2): 847–870. doi: 10.1109/JIOT.2018.2802704.

[2] TALEB S M, MERAIHI Y, GABIS A B, *et al*. Nodes placement in wireless mesh networks using optimization approaches: A survey[J]. *Neural Computing and Applications*, 2022, 34(7): 5283–5319. doi: 10.1007/s00521–022-06941-y.

[3] ALOTAIBI E and MUKHERJEE B. A survey on routing algorithms for wireless ad-hoc and mesh networks[J]. *Computer Networks*, 2012, 56(2): 940–965. doi: 10.1016/j.comnet.2011.10.011.

[4] WANG Lei, ZHANG Lianfang, SHU Yantai, *et al*. Multipath source routing in wireless ad hoc networks[C]. 2000 Canadian Conference on Electrical and Computer Engineering. Conference Proceedings. Navigating to a New Era (Cat. No. 00TH8492), Halifax, Canada, 2000: 479–483. doi: 10.1109/CCECE.2000.849755.

[5] GUO Xiaoyuan, WANG Feng, LIU Jiangchuan, *et al*. Path diversified multi-QoS optimization in multi-channel wireless mesh networks[J]. *Wireless Networks*, 2014, 20(6): 1583–1596. doi: 10.1007/s11276-014-0698-x.

[6] JIA Dongyao, ZOU Shengxiong, LI Meng, *et al*. Adaptive multi-path routing based on an improved leapfrog algorithm[J]. *Information Sciences*, 2016, 367/368: 615–629. doi: 10.1016/j.ins.2016.07.021.

[7] SUN Yaohua, PENG Mugen, ZHOU Yangcheng, *et al*. Application of machine learning in wireless networks: Key techniques and open issues[J]. *IEEE Communications Surveys & Tutorials*, 2019, 21(4): 3072–3108. doi: 10.1109/COMST.2019.2924243.

[8] DI VALERIO V, LO PRESTI F, PETRIOLI C, *et al*. CARMA: Channel-aware reinforcement learning-based multi-path adaptive routing for underwater wireless sensor networks[J]. *IEEE Journal on Selected Areas in Communications*, 2019, 37(11): 2634–2647. doi: 10.1109/JSAC.2019.2933968.

[9] LIU Qingzhi, CHENG Long, JIA A L, *et al*. Deep reinforcement learning for communication flow control in wireless mesh networks[J]. *IEEE Network*, 2021, 35(2): 112–119. doi: 10.1109/MNET.011.2000303.

[10] NG P C and SHE J. Remote proximity sensing with a novel Q-learning in Bluetooth low energy network[J]. *IEEE Transactions on Wireless Communications*, 2022, 21(8): 6156–6166. doi: 10.1109/TWC.2022.3147411.

[11] WANG Jinxin, ZHANG Fan, XIE Zhonglin, *et al*. Joint bandwidth allocation and path selection in WANs with path cardinality constraints[J]. *Journal of Communications and Information Networks*, 2021, 6(3): 237–250. doi: 10.23919/JCIN.2021.9549120.

[12] APPINI N R and REDDY A R. Joint channel assignment and bandwidth reservation using Improved FireFly Algorithm (IFA) in Wireless Mesh Networks (WMN)[J]. *Wireless Personal Communications*, 2023, 131(1): 455–470. doi: 10.1007/s11277-023-10439-8.

[13] BINH L H and DUONG T V T. Load balancing routing under constraints of quality of transmission in mesh wireless network based on software defined networking[J]. *Journal of Communications and Networks*, 2021, 23(1): 12–22. doi: 10.23919/JCN.2021.000004.

[14] KUMAR R, VENKANNA U, and TIWARI V. Opt-ACM: An optimized load balancing based admission control mechanism for software defined hybrid wireless based IoT (SDHW-IoT) network[J]. *Computer Networks*, 2021, 188: 107888. doi: 10.1016/j.comnet.2021.107888.

[15] ALHARBI N, MACKENZIE L, and PEZAROS D. Enhancing graph routing algorithm of industrial wireless sensor networks using the covariance-matrix adaptation evolution strategy[J]. *Sensors*, 2022, 22(19): 7462. doi: 10.3390/s22197462.

[16] BAROLLI A, BYLYKBASHI K, QAFZEZI E, *et al*. A comparison study of Weibull, normal and Boulevard distributions for wireless mesh networks considering different router replacement methods by a hybrid intelligent simulation system[J]. *Journal of Ambient Intelligence and Humanized Computing*, 2023, 14(8): 10181–10194. doi: 10.

1007/s12652-021-03680-1.

[17] ROZHOŇ V, HAEUPLER B, MARTINSSON A, *et al*. Parallel breadth-first search and exact shortest paths and stronger notions for approximate distances[C]. Proceedings of the 55th Annual ACM Symposium on Theory of Computing, Orlando, USA, 2023: 321–334. doi: 10.1145/3564246.3585235.

[18] SILVA N, WERNECK H, SILVA T, *et al*. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions[J]. *Expert Systems with Applications*, 2022, 197: 116669. doi: 10.1016/j.eswa.2022.116669.

[19] LEE S, YU H, and LEE H. Multiagent Q-learning-based multi-UAV wireless networks for maximizing energy efficiency: Deployment and power control strategy design[J]. *IEEE Internet of Things Journal*, 2022, 9(9): 6434–6442. doi: 10.1109/JIOT.2021.3113128.

[20] ZAATOURI I, ALYAOUI N, GUILOUFI A B, *et al*. Design and performance analysis of objective functions for RPL routing protocol[J]. *Wireless Personal Communications*, 2022, 124(3): 2677–2697. doi: 10.1007/s11277-022-09484-6.

朱晓荣：女，博士，教授，研究方向为5G/6G通信系统、物联网、区块链等关键技术及系统研发.

贺楚闳：男，硕士生，研究方向为无线通信、5G/6G网络、多维资源调度等.

责任编辑：余 蓉