异构云无线接入网架构下面向混合能源供应的动态资源分配及能源管理算法

陈前斌* 谭 颀 魏延南 贺兰钦 唐 伦

(重庆邮电大学通信与信息工程学院 重庆 400065)

(重庆邮电大学移动通信技术重点实验室 重庆 400065)

摘 要：针对面向混合能源供应的 5G 异构云无线接入网(H-CRANs)网络架构下的动态资源分配和能源管理问题，该文提出一种基于深度强化学习的动态网络资源分配及能源管理算法。首先，由于可再生能源到达的波动性及用户数据业务到达的随机性，同时考虑到系统的稳定性、能源的可持续性以及用户的服务质量(QoS)需求，将H-CRANs网络下的资源分配以及能源管理问题建立为一个以最大化服务提供商平均净收益为目标的受限无穷时间马尔科夫决策过程(CMDP)。然后，使用拉格朗日乘子法将所提CMDP问题转换为一个非受限的马尔科夫决策过程(MDP)问题。最后，因为行为空间与状态空间都是连续值集合，因此该文利用深度强化学习解决上述MDP问题。仿真结果表明，该文所提算法可有效保证用户QoS及能量可持续性的同时，提升了服务提供商的平均净收益，降低了能耗。

关键词：异构云无线接入网；混合能源；资源分配；能源管理；深度强化学习

中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2020)06-1428-08

DOI: 10.11999/JEIT190499

Dynamic Resource Allocation and Energy Management Algorithm for Hybrid Energy Supply in Heterogeneous Cloud Radio Access Networks

CHEN Qianbin TAN Qi WEI Yannan HE Lanqin TANG Lun

(*School of Communication and Information Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

(*Key Laboratory of Mobile Communications Technology*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

Abstract: Considering the dynamic resource allocation and energy management problem in the 5G Heterogeneous Cloud Radio Access Networks(H-CRANs) architecture for hybrid energy supply, a dynamic network resource allocation and energy management algorithm based on deep reinforcement learning is proposed. Firstly, due to the volatility of renewable energy and the randomness of user data service arrival, taking into account the stability of the system, the sustainability of energy and the Quality of Service(QoS) requirements of users, the resource allocation and energy management issues in the H-CRANs network as a Constrained infinite time Markov Decision Process (CMDP) are modeled with the goal of maximizing the average net profit of service providers. Then, the Lagrange multiplier method is used to transform the proposed CMDP problem into an unconstrained Markov Decision Process (MDP) problem. Finally, because the action space and the state space are both continuous value sets, the deep reinforcement learning is used to solve the above MDP problem. The simulation results show that the proposed algorithm can effectively guarantee the QoS and energy sustainability of the system, while improving the average net income of the service provider and reducing energy consumption.

Key words: Heterogeneous Cloud Radio Access Networks (H-CRANs); Hybrid energy; Resource allocation; Energy management; Deep reinforcement learning

收稿日期：2019-07-04；改回日期：2020-01-29；网络出版：2020-02-20

*通信作者： 陈前斌 cqb@cqupt.edu.cn

基金项目：国家自然科学基金(6157073)，重庆市教委科学技术研究项目(KJZD-M201800601)

Foundation Items: The National Natural Science Foundation of China (61571073), The Science and Technology Research Program of Chongqing Municipal Education Commission (KJZD-M201800601)

1　引言

为了支持5G提出的性能需求，当前网络需要一个新型的5G无线网络架构来满足传输速率及能耗等性能需求。值得注意的是，异构云无线接入架构(Heterogeneous Cloud Radio Access Network, H-CRAN)是一个有效的解决方案[1]。H-CRAN把传统的基站分离为宏基站(Macro Base Station, MBS)、无线远端射频单元(Remote Radio Head, RRH)和由基带处理单元(Base Band processing Unit, BBU)集中形成的BBU池[2]。相比较于传统的云无线接入网，H-CRAN通过运用3G和4G时代的蜂窝网络中的MBS来支持无缝覆盖以及控制平面和业务平面的功能分离。随着业务量的增加，运营商亟需进行基站扩建，这意味着更大的能源消耗。基于上述背景，人们提出了“绿色通信”的概念[3]。“绿色通信”技术是指在基站端配置能量收集设备用来收集可再生的能源。然而，由于可再生能源的波动性，人们提出使用混合能源供能技术。由上述背景，可以容易地看出，将H-CRAN网络架构与混合能源供应技术结合起来，可以保证服务质量并保证网络稳定性的同时降低成本支出。

目前，已有大量工作深入研究了混合能源供应技术与H-CRAN架构[4–10]。然而，现有的研究工作大多是将混合能源供应技术与H-CRAN架构分别讨论的，很少有研究将两种技术结合起来讨论，另外，现存的大部分研究都是基于环境状态是完全已知的，没有考虑到时变的无线网络环境。因此，本文针对基于混合能源供应的5G H-CRAN网络架构下的动态资源分配和能源管理问题，通过联合考虑了跨层干扰、用户业务的随机到达以及可再生能源的波动性，在保证系统稳定性、能源可持续性以及用户服务质量(Quality of Service, QoS)需求的前提下，对子载波和功率进行联合分配，并对每个基站进行能源管理，实现服务提供商的平均净收益最大化。其次，考虑到所提问题为一个随机优化问题，将上述问题看作是一个有约束的无穷时间马尔科夫决策过程(Constrained Markov Decision Process, CMDP)。接着，本文利用Lagrange乘子法将所提的CMDP问题转化为一个无约束的MDP问题，然后基于深度强化学习算法在每个离散的时隙中，系统根据当前时隙的系统环境状态为各个用户设备分配合适的功率和子载波，同时各个基站根据当前的能量队列状态进行能量管理，从而实现服务提供商的平均净收益最大化。

2　系统模型

如图1所示，在本文中，宏基站由从电力公司购买的电能供能，RRHs则考虑使用可再生能源和购买的电能组成的混合能源供能。本文考虑一个集中式的资源管理系统，其部署在BBU池。如上所述，假设整个H-CRAN架构网络中有一个宏基站和S个RRHs进行数据传输，下文将用$s, s \in \{0, 1, 2, \cdots, S\}$表示各个基站，其中$s = 0$时表示宏基站。在下文中，$\boldsymbol{U}$表示用户设备集合。另外，MUEs表示连接到宏基站的用户设备，用m表示；RUEs表示连接

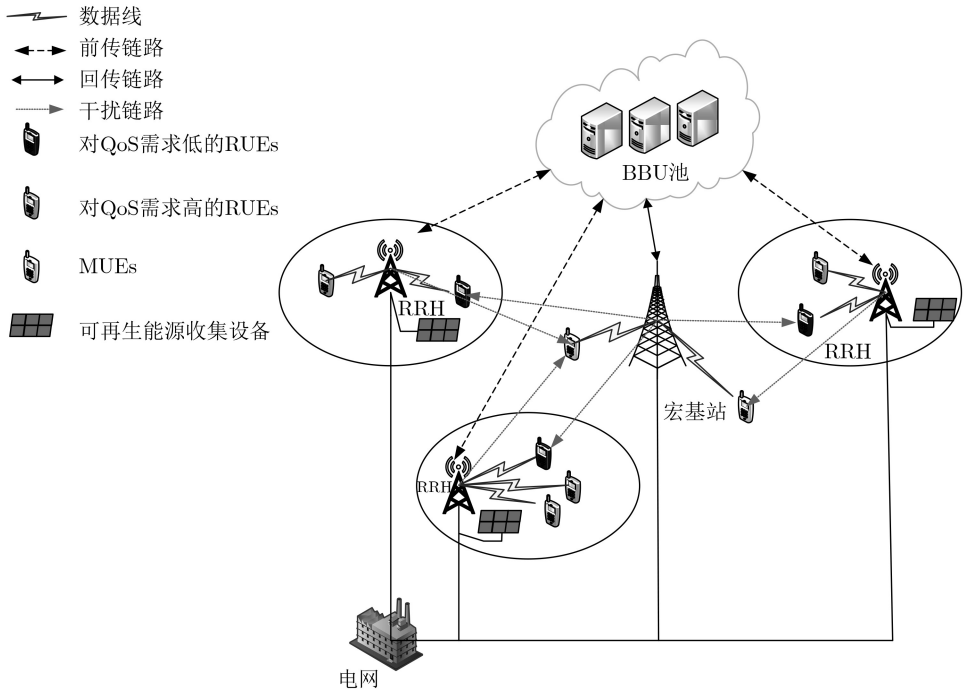

图 1 混合能源供能的RRHs和电能供能的MBS下行传输场景

到RRHs的用户设备，用n表示。在本文中，考虑采用文献[11]提出的频率复用方案将可用频谱资源分为两类：第1类频谱资源分配给对QoS要求高的RUEs，将此类频谱资源记为$\boldsymbol{\Gamma}_1$，此类RUEs集合记为$\boldsymbol{U}_{\boldsymbol{\Gamma}_1}$；第2类频谱资源分配给对QoS要求低的RUEs和MUEs，将此类频谱资源记为$\boldsymbol{\Gamma}_2$，此类RUEs集合记为$\boldsymbol{U}_{\boldsymbol{\Gamma}_2}$。本文使用$\boldsymbol{K}$表示子载波集合，每个子载波的带宽定义为$B$，本文假定基站和用户间的信道模型采用加性高斯白噪声信道。定义变量$a_{u,k}(t)$表示如果子载波k分配给了用户u，则$a_{u,k}(t)=1$，否则为0。定义$p_u^k(t)$表示时隙t分配给在子载波k上用户设备u的发射功率。对于连接到宏基站分配了子载波k的MUE m的SINR可以表示为

$$\gamma_{s,m}^k(t)=\frac{p_m^k(t)d_m^0h_{m,k}^0(t)}{\sum\limits_{u\in\boldsymbol{U}_{\boldsymbol{\Gamma}_2}}\sum\limits_{s\in\{1,2,\cdots,S\}}a_{u,k}(t)p_u^k(t)d_m^sh_{s,m}^k(t)+BN_0},k\in\boldsymbol{\Gamma}_2 \tag{1}$$

其中，$h_{s,n}^k(t)$表示基站s到RUEn在子载波k上的信道增益，d_n^s表示基站s到RUEn的路径损失，N_0表示噪声功率谱密度。类似地，对于连接到基站s分配了子载波k的RUEn的SINR表达式为

$$\gamma_{s,n}^k(t)=\begin{cases}\dfrac{p_n^k(t)d_n^sh_{s,n}^k(t)}{BN_0},k\in\boldsymbol{\Gamma}_1,n\in\boldsymbol{U}_{\boldsymbol{\Gamma}_1}\\ \dfrac{p_n^k(t)d_n^sh_{s,n}^k(t)}{\sum\limits_{u\in\boldsymbol{U}_0}a_{u,k}(t)p_u^k(t)d_n^0h_{0,\mathrm{n}}^k(t)+BN_0},k\in\boldsymbol{\Gamma}_2,n\in\boldsymbol{U}_{\boldsymbol{\Gamma}_2}\end{cases} \tag{2}$$

因此，可以得到对于连接到基站s上的用户设备u的传输速率可以表达为

$$r_{s,u}(t)=\sum_{k=1}^{K}a_u^k(t)B\log_2(1+\gamma_{s,u}^k(t)),\forall u\in\boldsymbol{U},\forall s\in\boldsymbol{S} \tag{3}$$

在本文中，定义系统中所有基站的能量到达过程相互独立。定义$e_s(t)$为在时隙t内基站s收集到的可再生能量，定义$\psi_s(t)$表示RRH s在时隙t内可以收集到的能量，其中$\psi_0(t)=0$且$e_0(t)=0$。定义$o_s(t)$表示基站s在时隙t时购买的电能，进一步地，定义$\alpha(t)$为时隙t时电能的单价。定义RRHs的功率消耗与MBS的功率消耗的表达式分别为

$$P_s(t)=\sum_{k\in\boldsymbol{K}}\sum_{u\in\boldsymbol{U}_s}a_{u,k}(t)p_u^k(t)+p_{c,s}(t)+p_{\mathrm{fh}},s(t),s\in\{1,2,\cdots,S\} \tag{4}$$

$$P_0(t)=\sum_{k\in\boldsymbol{K}}\sum_{u\in\boldsymbol{U}_{\mathrm{MBS}}}a_{u,k}(t)p_u^k(t)+p_{c,0}(t)+p_{\mathrm{bh},0}(t) \tag{5}$$

其中，$p_{c,s}(t),s\in\boldsymbol{S}$表示在时隙$t$内的基站$s$的静态功率消耗，$p_{\mathrm{fh},s}(t)$表示RRH s在时隙t内的去程链路功率消耗，$p_{\mathrm{bh},0}(t)$则表示MBS在时隙t内的回传链路功率消耗，$E_s(t)$表示时隙t开始时，在基站s上的可用的能量大小。定义$B_{\max}$为基站的电池容量最大值。因此，基站s的可用能量更新过程可以表示为

$$E_s(t+1)=\min\{B_{\max},E_s(t)-P_s(t)+o_s(t)+e_s(t)\},\forall s\in\boldsymbol{S} \tag{6}$$

本文假设每个用户设备都拥有一个缓存队列用于存储新到达的数据包。$Q_u(t)$表示用户u在时隙t开始时的队列长度，$A_u(t)$表示在时隙t内的新到达的数据包的个数，假设包的到达过程相互独立，且服从参数为λ_u的泊松分布。定义ΔT为时隙t的时长，用户u的队列更新过程可以表示为

$$Q_u(t+1)=[Q_u(t)-r_{s,u}(t)\Delta T+A_u(t)]^+,\forall s\in\boldsymbol{S},\forall u\in\boldsymbol{U} \tag{7}$$

根据Little's Law[6]定理，定义用户设备u的数据队列时间平均长度表达式为

$$\bar{Q}_u=\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T-1}\mathrm{E}[Q_u(t)],\forall u\in\boldsymbol{U} \tag{8}$$

2.1 CMDP模型建立

本文基于服务提供商的角度，将目标定义为最大化服务提供商的经济收益，具体可以表示为

$$G(t)=U\left(\sum_{s,u,k}\omega(t)r_{s,u}(t)\right)-\sum_{s=0}^{S}\alpha(t)o_s(t) \tag{9}$$

式(9)右边第1部分代表服务提供商在时隙t内向用户提供服务而得到的收入，参考文献[7]，将$U(\cdot)$定义为$U(\cdot)=\lg(1+\cdot)$。$\omega(t)$表示在时隙t时服务提供商向用户提供服务收取的单位费用。第2部分表示服务提供商向电力公司购买电能的成本。本文的资源调度及能源管理问题建立成以下数学模型，其中，约束C1和C2分别保证了RUEs和MUEs的用户数据队列时延。C3代表$a_{u,k}(t)$是一个二进制变量。C4表示在时隙t内1个用户只能由1个子载波进行服务。C5和C6分别保证对QoS需求高的RUEs和对QoS需求低的RUEs的速率要求。C7保证了MUEs的速率需求。C8表示对每个用户的功率限制。C9则表示对于在时隙t时基站s收集的能量不能大于环境中可以收集到的能量大小，C10表示在时隙t时基站s从电力公司购买电能大小的约束，C11表示在时隙t时基站s内的能量队列长度要提供足够的功率进行数据传输，C12表示在时隙t时基站s内的能量队列长度不能超过电池容量大小，C13则表示的队列的动态变化。

问题转化　CMDP 问题可以通过拉格朗日乘子转化为不受限的MDP问题，定义式(10)的广义拉格朗日表达式为

$$\begin{aligned}L(\xi,\upsilon,X,A)=&\bar{G}(X,A)+\sum_{u\in\boldsymbol{U}_{\mathrm{RRH}}}\xi_u(Q_{\max}-\bar{Q}_u)\\&+\sum_{u\in\boldsymbol{U}_{\mathrm{MBS}}}\upsilon_u(Q'_{\max}-\bar{Q}_u)\\=&\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{E}[G(t)-\sum_{u\in\boldsymbol{U}_{\mathrm{RRH}}}\xi_uQ_u(t)\\&-\sum_{u\in\boldsymbol{U}_{\mathrm{MBS}}}\upsilon_uQ_u(t)]+\sum_{u\in\boldsymbol{U}_{\mathrm{RRH}}}\xi_uQ_{\max}\\&+\sum_{u\in\boldsymbol{U}_{\mathrm{MBS}}}\upsilon_uQ'_{\max}\end{aligned}\tag{10}$$

其中，$\xi_u\geq 0,\forall u\in\boldsymbol{U}_{\mathrm{RRH}}$, $\upsilon_u\geq 0,\forall u\in\boldsymbol{U}_{\mathrm{MBS}}$分别是式(10)中约束条件C1和约束条件C2引入的拉格朗日乘子矢量，X,A则分别表示状态空间和行为空间。进一步的，定义时隙t时的拉格朗日回报函数为

$$r(t)=G(t)-\sum_{u\in\boldsymbol{U}_{\mathrm{RRH}}}\xi_uQ_u(t)-\sum_{u\in\boldsymbol{U}_{\mathrm{MBS}}}\upsilon_uQ_u(t)\tag{11}$$

据拉格朗日理论，则可以将原优化问题式(10)转化为$\max\limits_{X,A}\min\limits_{\xi,\upsilon}L(\xi,\upsilon,X,A)$。

(1) 状态空间X：定义$x(t)$表示网络在时隙t时的状态，其表达式为

$$\begin{aligned}x(t)=&(\gamma^k_{s,u}(t),\psi_s(t),E_s(t),Q_u(t),\alpha(t),\omega(t)),\\&\forall s\in\boldsymbol{S},\forall u\in\boldsymbol{U},\forall k\in\boldsymbol{K}\end{aligned}\tag{12}$$

(2) 行为空间A：定义$a(t)$表示网络在时隙t时的行为，另外，行为空间需要满足约束C3～C13。其表达式为

$$\begin{aligned}a(t)=&(a_{u,k}(t),p^k_u(t),e_s(t),o_s(t)),\\&\forall u\in\boldsymbol{U},\forall k\in\boldsymbol{K},\forall s\in\boldsymbol{S}\end{aligned}\tag{13}$$

(3) 状态转移概率 $P_a(x'|x)$：定义$f(\cdot)$为状态转换概率密度函数，则状态转移概率表达式为

$$P_a(x'\in X(t+1)|x(t),a(t))=\int_{x(t+1)}f(x(t),a(t),x')\mathrm{d}x'\tag{14}$$

(4) 即时奖励$r_a(x,x')$：本文定义即时奖励函数为拉格朗日回报，表达式为 $r_a(x,x')=r(t)$。令$\pi{:}X\to A$表示一个确定性策略，给定初始状态$x(0)$与策略π，定义行为值函数为

$$Q^{\pi}(x(t),a(t))=\mathrm{E}\left\{\sum_{t=0}^{\infty}\gamma^tr(t)|x(0)=x,a(0)=a,\pi\right\}\tag{15}$$

其中，$\gamma\in(0,1)$表示折扣因子，优化问题可以转化为如式(16)所示问题

$$\max_{X,A}\min_{\xi,\upsilon}\left(Q^{\pi}(x,a)+\sum_{u\in\boldsymbol{U}_{\mathrm{RRH}}}\xi_uQ_{\max}+\sum_{u\in\boldsymbol{U}_{\mathrm{MBS}}}\upsilon_uQ'_{\max}\right)\tag{16}$$

则对于给定的$\xi_u\geq 0,\forall u\in\boldsymbol{U}_{\mathrm{RRH}},\upsilon_u\geq 0,\forall u\in\boldsymbol{U}_{\mathrm{MBS}}$，最优策略$\pi^{*,\xi,\nu}$为$\pi^{*,\xi,\nu}=\arg\max\limits_{A}Q^{\pi}(x,a)$。

2.2 基于DDPG算法的资源分配及能源管理算法

对于有限状态的MDP，系统通常将值函数存储在一个查找表中[12,13]，由于本文的状态空间与行为空间都是连续值变量，因此对每个状态-行为对进行存储是不现实的。为了解决上述问题，本文基于深度确定性策略迭代(Deep Deterministic Policy Gradient, DDPG) 算法提出集中式的动态资源分配及能量管理算法。DDPG算法由4个神经网络构成：2个结构相同的行动者策略网络，分别为行动者网络$\mu(\theta^{\mu})$和行动者目标网络$\mu'(\theta^{\mu'})$; 2个结构相同的评判家评价网络，分别为评判家网络$Q(x_t,a_t|\theta^Q)$和评判家目标网络$Q'(x_t,a_t|\theta^{Q'})$ [14]。

(1) 评判家网络：评判家网络主要是为了解决连续值状态空间及行为空间的维度灾难问题，使用函数逼近法来估计值函数。在DDPG算法中，采用一个参数为θ^Q的神经网络来近似值函数$Q^{\pi}(x_t,a_t)$，则近似值函数可以表示为

$$\tilde{Q}(x_t,a_t|\theta^Q)\approx Q^{\pi}(x_t,a_t)\tag{17}$$

其中，a_t通过行动者网络输出得到。评判家网络的参数θ^Q可以通过最小化损失函数来更新，即

$$\theta^Q=\arg\min_{X,A}l(\theta^Q)\tag{18}$$

其中，$l(\theta^Q)$为损失函数，表达式为$l(\theta^Q)=\mathrm{E}[y_t-\tilde{Q}(x_t,a_t|\theta^Q)]^2$。其中$y_t$为

$$y_t=r(x_t,a_t)+\gamma Q(x_{t+1},a_{t+1}|\theta^Q)\tag{19}$$

其中，a_{t+1}由行动者目标网络输出得到。不失一般性，在从经验回放池D中随机采样N_{D}个样本后，损失函数通过式(20)计算

$$l(\theta^Q)=\frac{1}{N_{\mathrm{D}}}\sum_{i=1}^{N_{\mathrm{D}}}\left(y_i-Q(x_i,a_i|\theta^Q)\right)^2\tag{20}$$

(2) 行动者网络：在行动者网络中，通过优化行动者网络的参数生成最优策略。在DDPG算法中，参数θ^{μ}根据策略最优函数梯度进行更新，策略最优函数梯度的表达式为

$$\begin{aligned}\nabla_{\theta^{\mu}}J\approx&\mathrm{E}[\nabla_{\theta^{\mu}}\tilde{Q}(x_t,a_t|\theta^Q)|_{x=x_t,a=\mu(x_t|\theta^{\mu})}]\\=&\mathrm{E}[\nabla_a\tilde{Q}(x_t,a_t|\theta^Q)|_{x=x_t,a=\mu(x_t|\theta^{\mu})}\\&\cdot\nabla_{\theta^{\mu}}\pi(x_t|\theta^{\mu})|_{x=x_t}]\end{aligned}\tag{21}$$

不失一般性，在从经验回放池D中随机采样N_{D}个样本后，策略最优函数梯度表达式为

$$\nabla_{\theta^\mu} J = \frac{1}{N_{\rm D}} \sum_{i=1}^{N_{\rm D}} \nabla_a \tilde{Q}(x_i, a_i|\theta^Q)|_{x=x_i, a=\mu(x_i|\theta^\mu)} \cdot \nabla_{\theta^\mu} \pi(x_i|\theta^\mu)|_{x=x_i} \tag{22}$$

通过参数更新得到$\mu(\theta^\mu)$，为了对行为进行探索，通常需要给行为空间加一个探索噪声，定义$\boldsymbol{N}$为随机过程，则有$a_t = \mu(x_t|\theta_t^\mu) + \boldsymbol{N}$。

(3) 目标网络参数更新：在DDPG中，目标网络参数更新采用“软更新”方法，定义ς为软更新因子，则参数更新表达式为

$$\theta^{Q'} \leftarrow \varsigma\theta^Q + (1-\varsigma)\theta^{Q'},\ \theta^{\mu'} \leftarrow \varsigma\theta^\mu + (1-\varsigma)\theta^{\mu'} \tag{23}$$

完整的算法流程如表1所示。

特别地，式(22)和式(23)中，α_ξ与α_ν是学习步长。另外，本文所采用的DDPG算法，假设行动者网络有$N_{\rm actor}$个全连接层，评判家网络有$N_{\rm critic}$个全连接层，则时间复杂度为$O\left(\sum_{j=1}^{N_{\rm actor}} u_j u_{j+1} + \sum_{k=1}^{N_{\rm critic}} u_k u_{k+1}\right)$，$u_j$，$u_k$则分别表示对应层的神经元数量。

3 仿真结果与分析

为了评估本文所提基于DDPG的网络资源分配以及能源管理算法的有效性，本节将通过大量的仿真对本文所提出的算法进行了数值分析。另外，本节将通过与DQN算法以及文献[7]提出的GRA算法进行对比，并根据仿真结果进行详尽的分析。本文采用来自电力运营商California ISO[16]的可再生能源到达的数据文件。另外，仿真实验中$\alpha(t) - N(3,3)$，并参考文献[15]设置服务单价$\omega(t) = 0.9\alpha(t)$。本文参考文献[17]设置各仿真参数，通过python平台进行仿真。具体参数如表2所示。

由于在深度强化学习算法的代码实现时，分为学习阶段和实验阶段，学习阶段通过对神经网络进

表 1 算法流程表

算法1：基于DDPG算法的资源分配与能源管理算法

(1)初始化

随机初始化参数θ^Q和θ^μ；初始化目标网络参数：$\theta^{Q'} \leftarrow \theta^Q$，$\theta^{\mu'} \leftarrow \theta^\mu$；初始化拉格朗日乘子$\xi_u \geq 0, \forall u \in \boldsymbol{U}_{\rm RRH}, \upsilon_u \geq 0, \forall u \in \boldsymbol{U}_{\rm MBS}$；初始化经验回放池D

(2)学习阶段

For episode=1 to M do

初始化一个随机过程作为行为噪声$\boldsymbol{N}$，并观察初始状态x_0

For t=1 to T do

根据$a_t = \mu(x_t|\theta_t^\mu) + \boldsymbol{N}$选择一个行为

if 约束C3-C13满足：

执行行动a_t，并得到回报值r_t与下一状态x_{t+1}

将状态转换组$< x_t, a_t, r_t, x_{t+1} >$存入经验回放池D

从经验回放池D中随机采样$N_{\rm D}$个样本，每个样本用i表示

(a) 更新评判家网络

从行动者目标网络得到$\mu'(x_{i+1}|\theta^{\mu'})$

从评判家目标网络中得到$Q(s_{i+1}, \mu'(x_{i+1}|\theta^{\mu'})|\theta^{Q'})$

根据式(20)得到y_i，从评判家网络得到$Q(x_i, a_i|\theta^Q)$

根据式(21)计算损失函数，并根据式(19)更新评判家网络参数θ^Q

(b) 更新行动者网络

从评判家网络得到$Q(x_i, a_i|\theta^Q)$，并根据式(22)计算策略梯度

根据策略梯度更新行动者网络参数θ^μ

(c) 更新行动者目标网络和评判家目标网络

根据式(23)更新行动者目标网络和评判家网络参数

(d)基于标准次梯度法[15]更新拉格朗日乘子

$$\xi_{u,t+1} \leftarrow \left[\xi_{u,t} - \alpha_\xi(Q_{\max} - \bar{Q}_u)\right]^+, \forall u \in \boldsymbol{U}_{\rm RRH} \tag{24}$$

$$\nu_{u,t+1} \leftarrow \left[\nu_{u,t} - \alpha_\nu(Q_{\max} - \bar{Q}_u)\right]^+, \forall u \in \boldsymbol{U}_{\rm MBS} \tag{25}$$

End for

End for

表 2　仿真参数

仿真参数	值	仿真参数	值
RRH最大发射功率	3 W	数据包大小L	4 kbit/packet
MBS最大发射功率	10 W	MUEs路径损耗模型	31.5+35lg(d) (d[km])
热噪声功率谱密度	–102 dBm/Hz	RUEs路径损耗模型	31.5+40lg(d) (d[km])
子载波个数N	12	折扣因子γ	0.99
单个资源块带宽	180 kHz	$r_{\boldsymbol{\Gamma}_1}$	4 Mbps
软更新因子ς	0.01	$r_{\boldsymbol{\Gamma}_2}$	4.5 Mbps
时隙长度τ	10 ms	r_{MBS}	512 kbps

行学习，最终收敛得到最优策略。在实验阶段，可以直接通过已经学习好的参数得到当前状态下的最优行为。本文首先研究算法的收敛性能，如图2所示。在图2中，参考文献[10]固定行动者网络的学习率为0.02，并将评判家网络的学习速率分别设置为[0.010 0.015 0.020]，得到系统的平均净收益的变化曲线。从图2可以得到，当学习速率为0.015时，虽然收敛速度略慢于学习速率为0.020时的学习速率，但是可以得到一个更优的平均回报值，因此本文将评判家部分的学习速率定义为0.015。进一步地，如图3所示，本文对实验阶段进行了仿真。在实验阶段，本文购买电能的单价每30 min更新一次，从而得到系统净收益的性能。从图3中可以发现，在每次价格更新时，系统可以根据当前电力公司提供的价格进行合适的资源分配及能源管理，并达到当前价格下的系统净收益最大化。值得注意的是，在每30 min内，由于系统内用户位置的变动以及信道质量的动态变化等原因，造成了数据的波动。

图4将采取深度Q学习(Deep Q-Learning, DQL)算法与本文所提基于DDPG的资源分配与能源管理算法在不同的量化级数下进行对比。在此次实验中，将本文所定义的状态空间进行量化级数为L的均匀量化。如图4所示，基于DQN算法的平均净收益会随着量化级数增大而增长，这是因为随着量化级数增加，量化噪声随之降低，使得基于DQN算法的平均净收益会越来越逼近最优值。图5将本文所提的算法与文献[7]中GRA算法在不同的用户数下进行对比，用户数的设置分别为12与20。可以明显得到，随着用户数增加，基于GRA算法的平均净收益会减少，相反，本文所提出的资源分配及能源管理算法会使平均净收益随着用户数增加而增加。这是由于GRA算法更倾向于为用户提供更多的数据速率，这导致购买的能量也逐渐增加，最终使得为用户提供服务时所获得的净收益减少。

图6、图7分别比较了不同用户数下，分别使用混合能源与传统能源供能在系统和速率以及购买的电能方面的性能。在此次实验中，用户数分别设定为[12 16 20 24 28 30]。如图6所示，使用传统电能供能时的系统和速率在不同用户数下会略大于使用混合能源供能时的系统，这是因为使用传统电能供能时，服务提供商可以通过购买电能，为用户提供更稳定的数据传输服务。然而，从图7可得，由

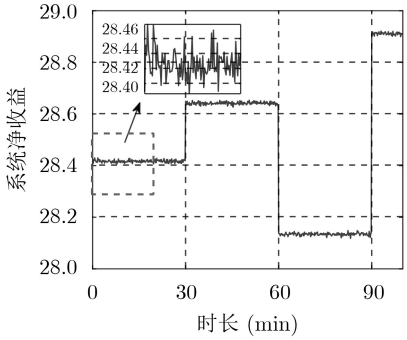

图 3 每30 min更新1次价格得到的系统净收益

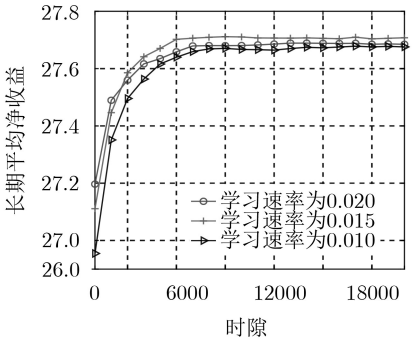

图 2 不同评判家学习速率下的平均净收益

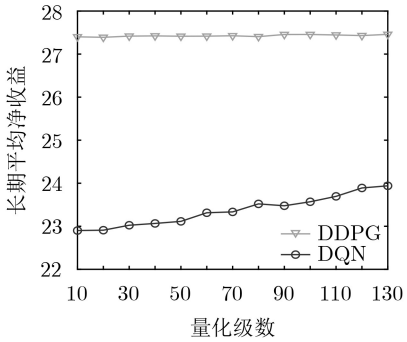

图 4 不同量化级数下的平均网络净收益对比

于服务提供商需要不断购买电能为用户提供数据传输服务，在传统电能供能下系统购买的电能多于使用混合能源供能时购买的电能。因此，通过图6、图7可以得到，使用传统电能供能虽然可以使得系统和速率略大于使用混合能源供能时的系统和速率，但是却要购买大量的电能，会造成大量的成本支出。

进一步地，本文分别对使用混合能源以及使用传统能源时的能效性能进行了验证。在此次实验中，设定用户数为12。本文参考文献[10]仅考虑传统电能部分的能耗，将能效的计算表达式定义为

$$\mathrm{EE}=\frac{\dfrac{1}{T}\lim\limits_{T\to\infty}\sum\limits_{t=0}^{T-1}\sum\limits_{s,u}r_{s,u}(t)}{\dfrac{1}{T}\lim\limits_{T\to\infty}\sum\limits_{t=0}^{T-1}\sum\limits_{s}o_s(t)} \tag{26}$$

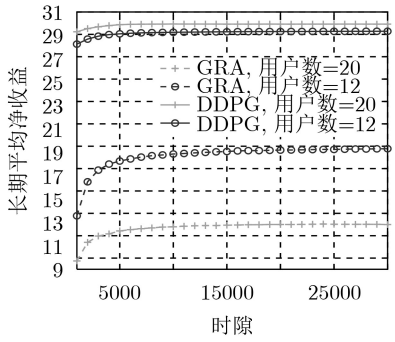

图 5 不同算法下的平均净收益对比

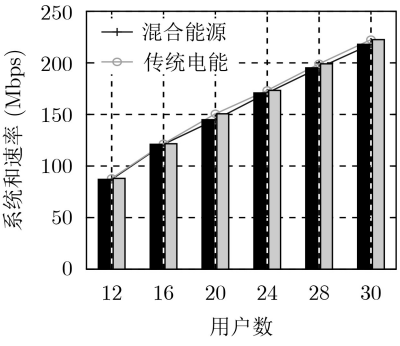

图 6 不同用户数下的系统和速率

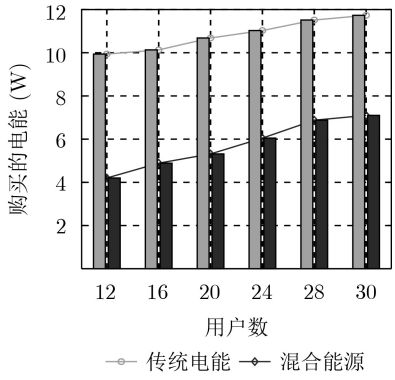

图 7 不同用户数下系统购买的电能

由图8可得，本文所采用的DDPG算法会随着可再生能源的波动而变化。由此可以看出本文所采用的深度强化学习算法可以根据当前到达的可再生能源动态的进行能量管理及资源分配。

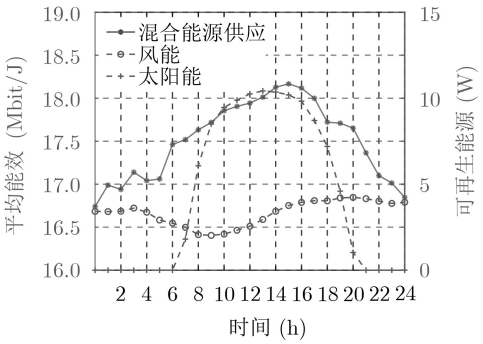

图 8 一天内的平均能效与可再生能源到达

4 结束语

为了满足5G时代的移动数据流量的增长需求，H-CRAN架构及混合能源技术可以有效解决在保证用户QoS及系统稳定性的同时，提升能效，降低运营商的成本支出。本文考虑在RRHs使用混合能源供能，MBS由传统电能供能的场景下，最大化服务提供商的平均净收益。本文将H-CRANs网络架构下的资源分配和能源管理问题描述为一个CMDP问题，建立一个在系统稳定性、用户QoS需求以及能量队列可持续性的约束条件下的随机优化模型。进一步地，因为状态空间及行为空间都是连续值变量，本文设计了一个基于DDPG算法的动态资源分配和能源管理方案，在每个离散的时隙中，集中式管理设备根据当前时隙的环境状态为各个用户设备分配合适的子载波与功率，并为每个基站进行能量管理，最终实现平均净收益最大化。仿真结果显示，本文所提的算法可以有效地收敛，通过使用可再生能源提高了能效，同时满足用户QoS并保证系统稳定性及能源的可持续性。

参 考 文 献

[1] 彭木根, 艾元. 异构云无线接入网络: 原理、架构、技术和挑战[J]. 电信科学, 2015, 31(5): 41–45.
PENG Mugen and AI Yuan. Heterogeneous cloud radio access networks: Principle, architecture, techniques and challenges[J]. *Telecommunications Science*, 2015, 31(5): 41–45.

[2] ALNOMAN A, CARVALHO G H S, ANPALAGAN A, *et al*. Energy efficiency on fully cloudified mobile networks: Survey, challenges, and open issues[J]. *IEEE Communications Surveys & Tutorials*, 2018, 20(2): 1271–1291. doi: 10.1109/COMST.2017.2780238.

[3] AKTAR M R, JAHID A, AL-HASAN M, *et al*. User

association for efficient utilization of green energy in cloud radio access network[C]. 2019 International Conference on Electrical, Computer and Communication Engineering, Cox' sBazar, Bangladesh, 2019: 1–5. doi: 10.1109/ECACE.2019.8679128.

[4] ALQERM I and SHIHADA B. Sophisticated online learning scheme for green resource allocation in 5G heterogeneous cloud radio access networks[J]. *IEEE Transactions on Mobile Computing*, 2018, 17(10): 2423–2437. doi: 10.1109/TMC.2018.2797166.

[5] LIU Qiang, HAN Tao, ANSARI N, *et al*. On designing energy-efficient heterogeneous cloud radio access networks[J]. *IEEE Transactions on Green Communications and Networking*, 2018, 2(3): 721–734. doi: 10.1109/TGCN.2018.2835451.

[6] 吴晓民. 能量捕获驱动的异构网络资源调度与优化研究[D]. [博士论文], 中国科学技术大学, 2016.
WU Xiaomin. Resources optimization and control in the energy harvesting heterogeneous network[D]. [Ph.D. dissertation], University of Science and Technology of China, 2016.

[7] ZHANG Deyu, CHEN Zhigang, CAI L X, *et al*. Resource allocation for green cloud radio access networks with hybrid energy supplies[J]. *IEEE Transactions on Vehicular Technology*, 2018, 67(2): 1684–1697. doi: 10.1109/TVT.2017.2754273.

[8] 孔巧. 混合能源供能的异构蜂窝网络中能源成本最小化问题的研究[D]. [硕士论文], 华中科技大学, 2016.
KONG Qiao. Research on energy cost minimization problem in heterogeneous cellular networks with hybrid energy supplies[D]. [Master dissertation], Huazhong University of Science and Technology, 2016.

[9] YANG Jian, YANG Qinghai, SHEN Zhong, *et al*. Suboptimal online resource allocation in hybrid energy supplied OFDMA cellular networks[J]. *IEEE Communications Letters*, 2016, 20(8): 1639–1642. doi: 10.1109/LCOMM.2016.2575834.

[10] WEI Yifei, YU F R, SONG Mei, *et al*. User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach[J]. *IEEE Transactions on Wireless Communications*, 2018, 17(1): 680–692. doi: 10.1109/TWC.2017.2769644.

[11] PENG Mugen, ZHANG Kecheng, JIANG Jiamo, *et al*. Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks[J]. *IEEE Transactions on Vehicular Technology*, 2015, 64(11): 5275–5287. doi: 10.1109/TVT.2014.2379922.

[12] 陈前斌, 杨友超, 周钰, 等. 基于随机学习的接入网服务功能链部署算法[J]. 电子与信息学报, 2019, 41(2): 417–423. doi: 10.11999/JEIT180310.
CHEN Qianbin, YANG Youchao, ZHOU Yu, *et al*. Deployment algorithm of service function chain of access network based on stochastic learning[J]. *Journal of Electronics & Information Technology*, 2019, 41(2): 417–423. doi: 10.11999/JEIT180310.

[13] 深度强化学习-DDPG算法原理和实现[EB/OL]. https://www.jianshu.com/p/6fe18d0d8822, 2018.

[14] 齐岳, 黄硕华. 基于深度强化学习DDPG算法的投资组合管理[J]. 计算机与现代化, 2018(5): 93–99. doi: 10.3969/j.issn.1006-2475.2018.05.019.
QI Yue and HUANG Shuohua. Portfolio management based on DDPG algorithm of deep reinforcement learning[J]. *Computer and Modernization*, 2018(5): 93–99. doi: 10.3969/j.issn.1006-2475.2018.05.019.

[15] California ISO[EB/OL]. http://www.caiso.com, 2019.

[16] WANG Xin, ZHANG Yu, CHEN Tianyi, *et al*. Dynamic energy management for smart-grid-powered coordinated multipoint systems[J]. *IEEE Journal on Selected Areas in Communications*, 2016, 34(5): 1348–1359. doi: 10.1109/JSAC.2016.2520220.

[17] LI Jian, PENG Mugen, YU Yuling, *et al*. Energy-efficient joint congestion control and resource optimization in heterogeneous cloud radio access networks[J]. *IEEE Transactions on Vehicular Technology*, 2016, 65(12): 9873–9887. doi: 10.1109/TVT.2016.2531184.

陈前斌：男，1967年生，教授，博士生导师，研究方向为个人通信、多媒体信息处理与传输、异构蜂窝网络等.

谭　颀：女，1995年生，硕士生，研究方向为5G网络切片、资源分配、随机优化理论.

魏延南：男，1995年生，硕士生，研究方向为5G网络切片、虚拟资源分配、随机优化理论.

贺兰钦：男，1995年生，硕士生，研究方向为5G网络切片，机器学习算法.

唐　伦：男，1973年生，教授，博士，研究方向为下一代无线通信网络、异构蜂窝网络、软件定义无线网络等.