# Multiplexing eMBB and URLLC in Wireless Powered Communication Networks: A Deep Reinforcement Learning-based Approach

Xiaotian Jiang, Kai Liang, Xiaoli Chu, *Senior Member*, Cheng Li and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—This paper investigates the multiplexing of enhanced mobile broadband (eMBB) and ultra-reliable low-latency communications (URLLC) services in a wireless powered communication network, where a hybrid access point coordinates the wireless energy transfer (WET) to users and receives information from them. The preemptive puncturing is adopted to multiplex URLLC traffic onto eMBB transmission. Apart from the energy used for wireless information transmission (WIT), the rest energy in user's battery is reserved to avoid insufficient energy for future WIT. The problem is formulated to jointly allocate subcarriers, time, and energy to maximize the uplink eMBB sum rate under the constraints of URLLC latency, radio frequency to direct current (RF/DC) sensitivity, user's battery capacity, and subcarriers availability. We propose a deep reinforcement learning-based approach named mixed deep deterministic policy gradient (Mixed-DDPG), which decomposes the optimization problem into a discrete subproblem for subcarriers allocation and a continuous subproblem for time and energy allocation, and solves them alternately. Numerical results show that the proposed algorithm achieves a higher eMBB sum rate than the existing schemes.**

***Index Terms*—eMBB, URLLC, wireless powered communication, preemptive puncturing, RF/DC sensitivity.**

## I. Introduction

WITH wireless devices becoming ubiquitous and carrying out various applications, wireless powered communication network (WPCN) has emerged to solve the energy supply problem of energy-limited devices [1]. The user association and time allocation in a WPCN were jointly optimized by adopting the $\alpha$-fair utility to maximize the sum, max-min, and proportional fairness rate in [2]. In [3], the total effective throughput was maximized by optimizing the trade-off between the transmission time and packet error rate of a WPCN while meeting the effective information requirements. A point-to-point energy harvesting system was considered in [4] with finite blocklength, where an achievable channel coding rate and a mean delay of the system were investigated.

In addition, how to efficiently multiplex enhanced mobile broadband (eMBB) and ultra-reliable low-latency communications (URLLC) on a shared channel has become a major challenge faced by 5G wireless networks [5]. Due to their different requirements, eMBB and URLLC transmit at different time scales [6]. Specifically, the time domain is divided into equal slots and each slot is further divided into multiple minislots, eMBB transmissions are performed on slots to achieve a high data rate and URLLC packets are transmitted on minislots to reduce latency. To solve the problem of optimal allocation of radio resources when eMBB and URLLC are multiplexed, [7] adopts a preemptive puncturing method, i.e. an arriving URLLC packet is scheduled to transmit in the next minislot by preempting subcarriers already allocated to eMBB users, which is shown to achieve higher expected rates than static or semi-static allocation of spectrum resources. The authors in [5] maximized the eMBB throughput under URLLC constraints by jointly optimizing the traffic scheduling for eMBB and the preemptive puncturing for URLLC. In [6], a simplified model-free deep reinforcement learning-based approach was proposed to minimize the loss of eMBB transmission rate due to URLLC packet puncturing under the assumption of advanced allocation of radio resources for eMBB users and each URLLC packet can preempt radio resources from multiple eMBB users.

However, the authors in [4] assumed an infinite battery capacity for the user, which is infeasible in practice. In [1]- [4], uplink wireless information transmission (WIT) in each slot relies on the energy harvested only in the current slot without any energy reservation, hence some slots may see the harvested energy insufficient for uplink WIT due to channel variations [8]. For example, a deep fading channel will result in reduced energy harvested by the user while requiring more energy for uplink WIT. Moreover, we note that the multiplexing of eMBB and URLLC services has not been studied for wireless energy transfer (WET) based WPCNs yet. As a result, the existing system models may not be readily applicable in WPCN scenarios where multiple services of different requirements, such as URLLC and eMBB, share the same spectrum.

In contrast to the above works, this paper investigates the multiplexing of eMBB and URLLC transmissions on a shared channel in a WPCN, where a hybrid access point (HAP) powers multiple users by WET with the consideration of energy reservation and the preemptive puncturing.

This work was supported in part by National Key R&D Program of China under Grant 2021YFE0205200, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23031, in part by National Natural Science Foundation of China under Grant 61901317, and in part by the European Commission's Horizon 2020 Research and Innovation Program under Grant Agreement No. 778305. Xiaotian Jiang, and Kai Liang are with the School of Telecommunications Engineering, Xidian University, Xi'an, 710071, China (email: x.jiang@stu.xidian.edu.cn, kliang@xidian.edu.cn). Xiaoli Chu is with the Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S1 3JD, U.K. (e-mail: x.chu@sheffield.ac.uk). Cheng Li is with Xi'an Aerospace Precision Electromechanical Institute, Xi'an, 710199, China (e-mail:xevillee@126.com). G. K. Karagiannidis is with Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece and also with Cyber Security Systems and Applied AI Research Center, Lebanese American University (LAU), Lebanon (email: geokarag@auth.gr).

(Corresponding author: Kai Liang)

The contributions of this paper are summarized as follows: (i) Unlike the existing works, we study the problem of how to multiplex eMBB and URLLC transmissions in the uplink of a WPCN while considering finite battery capacity at each user, radio frequency to direct current (RF/DC) sensitivity of the energy harvesting circuit, and energy reservation for each user's battery to ensure power supply for uplink transmission.

(ii) We formulate an optimization problem to maximize the uplink eMBB sum rate by jointly optimizing the allocation of subcarriers, time for WET, and energy reservation of each user's battery under the constraints of URLLC latency, RF/DC sensitivity, user's battery capacity, and subcarriers availability.

(iii) To solve this non-convex optimization problem that features mixed allocation of discrete subcarriers and continuous time and energy, we propose a deep reinforcement learning-based approach named mixed deep deterministic policy gradient (Mixed-DDPG) to decompose it into a discrete subproblem and a continuous subproblem and solve them alternately.

## II. System Model and Problem Formulation

We consider a WPCN that includes a HAP and a set $\mathcal{U}$ of $U$ users with eMBB and URLLC transmission requirements. Each user has a rechargeable battery. For analytical tractability, it is assumed that the HAP and each user are equipped with a single antenna [9]. Let $\mathcal{B} = \{1, 2, \ldots, B\}$ denote the set of available subcarriers each with a bandwidth of $f_b$ Hz. Thus, the total bandwidth is $\sum_{b\in\mathcal{B}} f_b$ Hz. A long time period is considered and is divided into $T$ equal slots, denoted by $\mathcal{T} = \{1, 2, \ldots, T\}$. Each slot has a duration $t_0$. On each subcarrier, we assume channel reciprocity and that the channel fading coefficient stays constant within a slot but may change across adjacent slots. The subcarriers are reclaimed and rescheduled to eMBB users at the beginning of each slot based on the channel state information (CSI) [5]. Due to the stringent latency requirement of URLLC transmissions, we adopt the "URLLC preemption" scheme [10], where each slot is further divided into minislots represented by $\mathcal{M} = \{1, 2, \ldots, M\}$, and an arriving URLLC packet is scheduled immediately for transmission in the next minislot by preempting the subcarriers already allocated to the same user for eMBB transmissions, without waiting for the eMBB transmissions on those subcarriers to finish [6]. Without loss of generality, we assume that each user has URLLC packets arriving at each minislot.

All users adopt the harvest-then-transmit protocol in each slot, where the users first harvest energy from the energy signal broadcast by the HAP and then transmit information to the HAP using the harvested energy. For instance, if the $u$th user is scheduled to transmit in the $t$th slot, then the $t$th slot is divided into a downlink WET phase of duration $\tau_{u,t}t_0$ and an uplink WIT phase of duration $(1-\tau_{u,t})t_0$, where $\tau_{u,t} \in (0, 1)$. The HAP's downlink transmission power is assumed to be the same on each subcarrier and is denoted by $P^{\mathrm{DL}}$. The received power $P^{\mathrm{r}}_{u,t}$ of the $u$th user in the $t$th slot is given by

$$P^{\mathrm{r}}_{u,t} = \eta_c P^{\mathrm{DL}} d_u^{-\alpha} \sum_{b\in\mathcal{B}} |h_{u,b,t}|^2 \sum_{m\in\mathcal{M}} x_{u,b,t,m}, \tag{1}$$

where $\eta_c$ is the energy conversion efficiency of the RF/DC circuit, $d_u$ is the distance between the $u$th user and the HAP, $\alpha$ is the path loss exponent, $h_{u,b,t} \in \mathcal{CN}(0, 1)$ denotes the Rayleigh fading coefficient between the HAP and the $u$th user on subcarrier $b$ in the $t$th slot, and $x_{u,b,t,m} \in \{0, 1\}$ is a binary indicator of subcarrier allocation, where $x_{u,b,t,m} = 1$ means that subcarrier $b$ is allocated to the $u$th user in minislot $m$ of slot $t$ for eMBB transmission, otherwise $x_{u,b,t,m} = 0$.

Since a user cannot harvest energy if its received power is less than the RF/DC circuit sensitivity $\phi$, the received energy at the $u$th user in the $t$th slot is given by

$$E_{u,t} = P^{\mathrm{r}}_{u,t}\tau_{u,t}t_0 \mathbb{1}\left(P^{\mathrm{r}}_{u,t} \geq \phi\right), \tag{2}$$

where $\mathbb{1}(\cdot)$ is the binary indicator function.

The uplink transmission power of the $u$th user during the WIT phase in slot $t$ is given by

$$P^{\mathrm{UL}}_{u,t} = \frac{(1-\rho_{u,t})\, Q_{u,t}}{(1-\tau_{u,t})\, t_0}, \tag{3}$$

where $\rho_{u,t} \in [0, 1]$ is the percentage of energy reserved by the $u$th user in the $t$th slot for the next WIT of slot $t+1$ and $Q_{u,t}$ is the battery energy level of the $u$th user at the end of WET in slot $t$, which is updated as

$$Q_{u,t} = \min\left\{\rho_{u,t-1}Q_{u,t-1} + E_{u,t}, Q_{\max}\right\}, \tag{4}$$

where $Q_{u,t-1}$ is the battery energy level of the $u$th user at the end of the WET in slot $t-1$, and $Q_{\max}$ is the battery capacity. The uplink received signal to noise ratio (SNR) of the $u$th user on subcarrier $b$ in slot $t$ is given as follows

$$\gamma_{u,b,t} = \frac{P^{\mathrm{UL}}_{u,t}|h_{u,b,t}|^2 d_u^{-\alpha}}{\sigma^2 f_b}, \tag{5}$$

where $\sigma^2$ is the power spectral density of additive noise.

Based on the Shannon capacity [9], the eMBB transmission rate of the $u$th user in minislot $m$ of slot $t$ is given by

$$R^{\mathrm{mbb}}_{u,t,m} = \sum_{b\in\mathcal{B}} f_b \left(x_{u,b,t,m} - y_{u,b,t,m}\right) \log_2\left(1 + \gamma_{u,b,t}\right), \tag{6}$$

where $y_{u,b,t,m} \in \{0, 1\}$ is the binary indicator of subcarrier preemption by URLLC packets. Specifically, $y_{u,b,t,m} = 1$ indicates that subcarrier $b$ is preemted by the $u$th user in minislot $m$ of slot $t$ for URLLC transmission, otherwise $y_{u,b,t,m} = 0$. To ensure that the URLLC packets of the $u$th user can only preempt the subcarriers that have been allocated to the $u$th user for eMBB transmissions, it is necessary to specify that $0 \leq x_{u,b,t,m} - y_{u,b,t,m} \leq 1, \forall u, b, t, m$.

Since the packet length of URLLC is typically much shorter than that of eMBB, using the Shannon capacity may significantly overestimate the delay of URLLC transmissions [7]. Instead, the URLLC transmission rate of the $u$th user in minislot $m$ of slot $t$ can be calculated based on the finite block length theorem [10]:

$$R^{\mathrm{llc}}_{u,t,m} = \sum_{b\in\mathcal{B}} y_{u,b,t,m} f_b \Bigg( \log_2\left(1 + \gamma_{u,b,t}\right) - \sqrt{\frac{C_{u,b,t}}{n_u}} Q^{-1}\left(\varepsilon\right) \Bigg), \tag{7}$$

where $n_u$ is the length (in symbols) of the codeword block for the $u$th user, $\varepsilon$ is the decoding error probability, $Q^{-1}(\varepsilon)$ is the

inverse of the Gaussian cumulative distribution function, and $C_{u,b,t} = 1 - \frac{1}{(1+\gamma_{u,b,t})^2}$ is the channel dispersion.

To maximize the eMBB sum rate of all the users, we formulate the following optimization problem:

$$(P): \max_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\rho}} \quad \sum_{t\in\mathcal{T}}\sum_{u\in\mathcal{U}}\sum_{m\in\mathcal{M}} R^{\text{mbb}}_{u,t,m} \tag{8}$$

$$\text{s.t.} \quad \sum_{u\in\mathcal{U}} x_{u,b,t,m} \le 1, \forall b,t,m, \tag{8a}$$

$$x_{u,b,t,m}, y_{u,b,t,m} \in \{0,1\}, \forall u,b,t,m, \tag{8b}$$

$$0 \le x_{u,b,t,m} - y_{u,b,t,m} \le 1, \forall u,b,t,m, \tag{8c}$$

$$\rho_{u,t-1}Q_{u,t-1} + E_{u,t} \le Q_{\max}, \forall u,t, \tag{8d}$$

$$R^{\text{mbb}}_{u,t,m} \ge \omega_u, \forall u,t,m, \tag{8e}$$

$$\frac{F_u}{R^{\text{llc}}_{u,t,m}} \le \psi, \forall u,t,m, \tag{8f}$$

$$0<\tau_{u,t}<1, \forall u,t, \tag{8g}$$

$$0 \le \rho_{u,t} \le 1, \forall u,t, \tag{8h}$$

where $\boldsymbol{X} = \{x_{u,b,t,m}\}_{u\in\mathcal{U},b\in\mathcal{B},t\in\mathcal{T},m\in\mathcal{M}}$, $\boldsymbol{Y} = \{y_{u,b,t,m}\}_{u\in\mathcal{U},b\in\mathcal{B},t\in\mathcal{T},m\in\mathcal{M}}$, $\boldsymbol{\tau} = \{\tau_{u,t}\}_{u\in\mathcal{U},t\in\mathcal{T}}$, $\boldsymbol{\rho} = \{\rho_{u,t}\}_{u\in\mathcal{U},t\in\mathcal{T}}$, $\omega_u$ is the minimum data rate requirement for eMBB transmission of the $u$th user, $F_u$ is the URLLC packet length (in bits) of the $u$th user, and $\psi$ is the maximum tolerable delay of URLLC packets. Constraints (8a) and (8b) ensure that each subcarrier is allocated to at most one user at any time, while (8c) is the subcarriers preemption availability constraint. Constraint (8d) is introduced to avoid energy overflow due to the battery capacity [11]. Constraint (8e) imposes the eMBB minimum data rate requirement. Constraint (8f) is the latency requirement for the URLLC transmission.

## III. Mixed-DDPG Based Resource Allocation

To tackle the non-convex optimization problem $(P)$ with mixed allocation of discrete subcarriers and continuous time and energy, we propose a novel alternate approach called Mixed-DDPG in this section. Specifically, we decompose the problem $(P)$ into discrete and continuous subproblems, where the discrete subproblem optimises the allocation of subcarriers while the continuous subproblem optimises the time allocation for WET and energy reservation for WIT. The two subproblems are then solved alternately until convergence.

### A. Discrete Subproblem

By fixing the continuous time allocation $\{\tau_{u,t}\}_{u\in\mathcal{U}}$ for WET and energy reservation $\{\rho_{u,t}\}_{u\in\mathcal{U}}$ for WIT of slot $t$ in problem $(P)$, we obtain a discrete subproblem that optimizes the binary indicators of subcarrier allocation to eMBB transmission at the beginning of slot $t+1$ and subcarrier preemption by URLLC packets in each minislot $m \in \mathcal{M}$ of slot $t+1$, $\forall t \in \mathcal{T}$. Moreover, based on (1)-(6), for fixed $\{\tau_{u,t}\}_{u\in\mathcal{U}}$ and $\{\rho_{u,t}\}_{u\in\mathcal{U}}$, $R^{\text{mbb}}_{u,t,m}$ becomes independent for different slot $t$, and the eMBB sum rate of all the users can be maximized separately in each slot. Hense, for slot $t$, under the given time allocation $\{\tau_{u,t}\}_{u\in\mathcal{U}}$ for WET and energy reservation $\{\rho_{u,}\}_{u\in\mathcal{U}}$ for WIT, problem $(P)$ reduces to the following discrete subproblem,

$$(P1): \max_{X_{t+1},Y_{t+1}} \quad \sum_{u\in\mathcal{U}}\sum_{m\in\mathcal{M}} R^{\text{mbb}}_{u,t,m}, \forall t \in \mathcal{T}, \tag{9}$$

$$\text{s.t.} \quad (8a),(8b),(8c),(8e),(8f),$$

where $X_t = \{x_{u,b,t,m}\}_{u\in\mathcal{U},b\in\mathcal{B},m\in\mathcal{M}}$ and $Y_t = \{y_{u,b,t,m}\}_{u\in\mathcal{U},b\in\mathcal{B},m\in\mathcal{M}}$. We can show that subproblem $(P1)$ is convex and can be solved by using existing convex optimization tools.

### B. Continuous Subproblem

For given binary indicators $\boldsymbol{X}$, $\boldsymbol{Y}$ of subcarrier allocation to eMBB transmission and subcarrier preemption by URLLC packets, problem $(P)$ reduces to the following continuous subproblem:

$$(P2): \max_{\boldsymbol{\tau},\boldsymbol{\rho}} \quad \sum_{t\in T}\sum_{u\in\mathcal{U}}\sum_{m\in\mathcal{M}} R^{\text{mbb}}_{u,t,m} \tag{10}$$

$$\text{s.t.} \quad (8d),(8e),(8f),(8g),(8h).$$

We note that $(P2)$ has large state spaces, including $\boldsymbol{X}$, $\boldsymbol{Y}$, CSI, and battery status of all users in different slots, and hence will be difficult to solve using conventional optimization methods, but can leverage deep reinforcement learning (DRL) [6]. Since the variables $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ are continuous, we adopt a model-free DRL, i.e., DDPG that has a continuous action space [10], to solve subproblem $(P2)$. The DDPG state, action, and reward are defined as follows.

- **State**: $s_t = \{H_t, Q_t, X_t, Y_t\}$, where $H_t = \{h_{u,b,t}\}_{u\in\mathcal{U},b\in\mathcal{B}}$ contains the CSI and $Q_t = \{Q_{u,t}\}_{u\in\mathcal{U}}$ denotes the battery status.
- **Action**: $a_t = \{\tau_t, \rho_t\}$, where $\tau_t = \{\tau_{u,t}\}_{u\in\mathcal{U}}$ denotes the time allocation for WET and $\rho_t = \{\rho_{u,t}\}_{u\in\mathcal{U}}$ denotes the energy reservation proportion.
- **Reward**: if action $a_t$ is chosen, the reward $r_t$ is given by

$$r_t = \sum_{u\in\mathcal{U}}\left(\sum_{m\in\mathcal{M}} R^{\text{mbb}}_{u,t,m} - \delta\sum_{m\in\mathcal{M}}\hat{R}^{\text{mbb}}_{u,t,m}\right), \tag{11}$$

  where $\delta>0$ is the penalty factor and $\hat{R}^{\text{mbb}}_{u,t,m} = \frac{1}{t-1}\sum_{i=1}^{t-1} R^{\text{mbb}}_{u,i,m}$. The penalty $\delta\sum_{m\in\mathcal{M}}\hat{R}^{\text{mbb}}_{u,t,m}$ will be imposed by the system on the agent when any constraint of $(P2)$ is violated, thereby avoiding overfitting.

DDPG consists of an actor network and a critic network for generating and evaluating policies, respectively [12]. Based on the input state $s_t$, the actor network $\mu(s_t|\theta^\mu)$ selects the deterministic action as follows [10]

$$a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t, \mathcal{N}_t \sim N(\mu_1, \sigma_1^2), \tag{12}$$

where $\theta^\mu$ is the actor network parameter and $\mathcal{N}_t$ is an additional noise that follows a normal distribution with a mean of $\mu_1$ and variance of $\sigma_1^2$ due to action exploration.

For given $s_t$, $a_t$ and reward $r_t$, after randomly selecting a minibatch of $N$ tuples $\{(s_j, a_j, r_j, s_{j+1})\}_{j=1,..,N}$ from the replay buffer $\mathcal{D}$, which is introduced to reduce the correlation among training samples, the critic network generates a Q value

$Q\left(s_j, a_j|\theta^Q\right)$ [6] to assess the selected action $a_t$ and updates its parameter $\theta^Q$ by minimizing the loss:

$$L = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - Q\left(s_j, a_j|\theta^Q\right)\right)^2, \tag{13}$$

where $y_j = r_j + \gamma Q^{'}\left(s_{j+1}, \mu^{'}\left(s_{j+1}|\theta^{\mu^{'}}\right)|\theta^{Q^{'}}\right)$, $\gamma$ is the discount rate, $\theta^{\mu^{'}}$ and $\theta^{Q^{'}}$ are the parameters of target actor network $\mu^{'}\left(s|\theta^{\mu^{'}}\right)$ and target critic network $Q^{'}\left(s, a|\theta^{Q^{'}}\right)$, respectively, which are introduced to ensure the stability of the DDPG learning process.

The actor policy is updated using the deterministic sampled gradient policy as follows,

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_{j=1}^{N}\nabla_a Q\left(s, a|\theta^Q\right)|_{s=s_j, a=\mu(s_j|\theta^\mu)} \times \nabla_{\theta^\mu}\mu\left(s|\theta^\mu\right)|_{s=s_j}. \tag{14}$$

The parameters $\theta^{\mu^{'}}$ of the target actor network and $\theta^{Q^{'}}$ of the target critic network are soft updated as follows,

$$\theta^{\mu^{'}} \leftarrow \zeta\theta^\mu + (1-\zeta)\,\theta^{\mu^{'}}, \tag{15}$$

$$\theta^{Q^{'}} \leftarrow \zeta\theta^Q + (1-\zeta)\,\theta^{Q^{'}}, \tag{16}$$

where $0<\zeta \ll 1$ is the soft updating rate [12].

### C. The Mixed-DDPG Algorithm

Based on the aforementioned solutions to subproblems $(P1)$ and $(P2)$, we propose a Mixed-DDPG algorithm to solve problem $(P)$ as shown in Algorithm 1. Specifically, we first initialize $Q_t$, $\mathcal{D}$, $X_t$ and $Y_t$ in line 2, then obtain the optimal $\tau_t$ and $\rho_t$ by solving subproblem $(P2)$ using the DDPG-based approach in lines 4-7 and 9-11. Next, based on the obtained $\tau_t$ and $\rho_t$, we get $X_{t+1}$ and $Y_{t+1}$ for the next slot by solving subproblem $(P1)$ in line 8. The above steps repeat alternately and iteratively until the maximum number of slots per episode and the maximum number of episodes are reached.

We analyze the computational complexity of the proposed Mixed-DDPG as follows. Since the convex subproblem $(P1)$ has $2UBTM$ variables and can be solved by existing convex optimization tools, the computational complexity of solving $(P1)$ is $\mathcal{O}\left((2UBTM)^3\right)$. For subproblem $(P2)$, let $j_i$ and $k_i$ denote the sizes of the input and the output of the layer $i$ of the DDPG, respectively, where $i \in \mathcal{I}$, then the complexity of solving $(P2)$ is $\mathcal{O}\left(\sum_{i\in\mathcal{I}} j_i k_i\right)$. Therefore, the total computational complexity of the proposed Mixed-DDPG is $\mathcal{O}\left(F_e F_s\left((2UBTM)^3 + \sum_{i\in\mathcal{I}} j_i k_i\right)\right)$, where $F_e$ and $F_s$ are the maximum number of episodes and the maximum number of slots per episode for Algorithm 1, respectively.

## IV. Numerical Results

The simulation scenario consists of $U = 10$ users and $B = 20$ subcarriers each with bandwidth $f_b = 1$ MHz. $\sigma^2$ is set as $-174$ dBm/Hz and $t_0$ is normalized to 1. For each user, we set $\eta_c = 0.5$, $F_u = 20$ bytes, and the disance $d_u$ is uniformly distributed within the range $(10m, 15m)$. Each battery has the maximum capacity of $Q_{\max} \in \{5, 10, 15, 20, 25, 30\}$ $\mu$J, and the circuit sensitivity and maximum tolerable delay of URLLC packets are set as $\phi \in \{3, 9, 15, 21, 27\}$ $\mu$W and $\psi \in \{0.01, 0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ ms, respectively. In addition, we assume $\alpha = 3$, $\varepsilon = 0.01$, $\omega_u = 10$ Mbps, $\gamma = 0.9$, $M = 8$ [5], $\mu_1 = 0.5$ and $\sigma_1 = 0.5$. The critic and actor networks have three hidden layers with 128 neurons for each layer. We use Adam optimizer training the network with an initial learning rate $\lambda_a = 0.001$ and $\lambda_c = 0.002$ for actor and critic networks, respectively, setting the batch size to 128, $F_e = 200$ and $F_s = 200$.

**Algorithm 1** Mixed-DDPG for eMBB sum-rate maximization

**Initialization:** $\mathcal{U}$, $\mathcal{B}$, $\{f_b\}_{b\in\mathcal{B}}$, $\eta_c$, $P^{\mathrm{DL}}$, $\{d_u\}_{u\in\mathcal{U}}$, $\sigma^2$, $\psi$, $\varepsilon$, $\phi$, $\{\omega_u\}_{u\in\mathcal{U}}$, $Q_{\max}$.

1: **for** all episodes **do**
2: Set $t = 0$. Randomly initialize $Q_t$, $X_t$, $Y_t$ and $\mathcal{D}$.
3: **for** all slots of an episode **do**
4: $t = t + 1$, observe CSI $H_t$ and obtain $Q_t$ based on (4), obtain state $s_t = \{H_t, Q_t, X_t, Y_t\}$.
5: Select action $a_t$ based on (12).
6: Get the reward $r_t$ based on (11).
7: Obtain new battery energy level $Q_{t+1}$ based on (4).
8: Observe new state $s_{t+1}$ by solving subproblem $(P1)$.
9: Store $(s_t, a_t, r_t, s_{t+1})$ in replay memory buffer $\mathcal{D}$.
10: Randomly sample $N$ tuples from $\mathcal{D}$ as training data.
11: Update $\theta^Q$, $\theta^\mu$, $\theta^{\mu^{'}}$ and $\theta^{Q^{'}}$ based on (13), (14), (15) and (16), respectively.
12: **end for**
13: **end for**

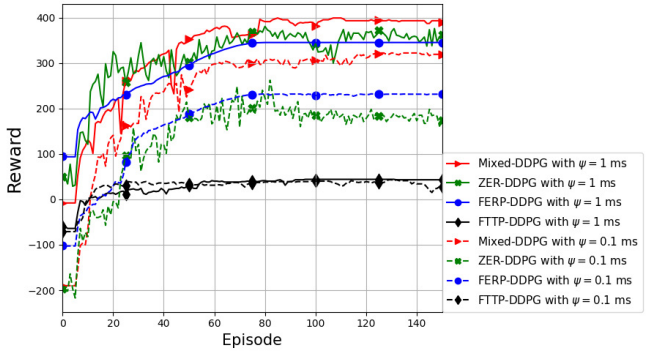


Fig. 1. DDPG reward versus the number of episodes.

For performance comparison with the proposed Mixed-DDPG algorithm, we include in the simulations the following three benchmark algorithms: zero energy reservation DDPG (ZER-DDPG), which differs from the Mixed-DDPG only in $\rho = 0$; fixed energy reservation proportional DDPG (FERP-DDPG), which differs from the Mixed-DDPG only in $\rho = 0.5$; and fixed transmission time proportional DDPG (FTTP-DDPG), which differs from the Mixed-DDPG only in $\tau = 0.5$.

Fig. 1 shows the rewards versus the number of episodes of the four algorithms for two different maximum tolerable URLLC delays, $\psi = 1$ ms and $\psi = 0.1$ ms, respectively, where $\phi = 20$ $\mu$W, $Q_{\max} = 25$ $\mu$J. The proposed algorithm

This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2023.3290040



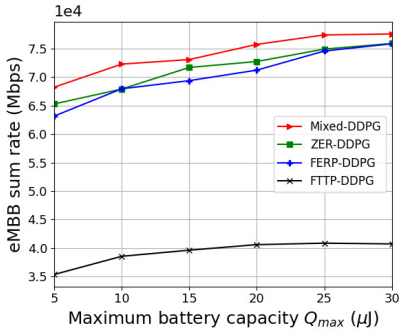


Fig. 2. eMBB sum rate versus the maximum battery capacity $Q_{\max}$.

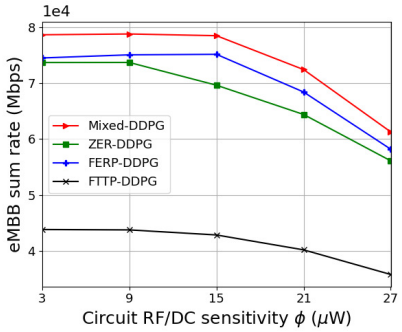


Fig. 3. eMBB sum rate versus RF/DC sensitivity $\phi$.

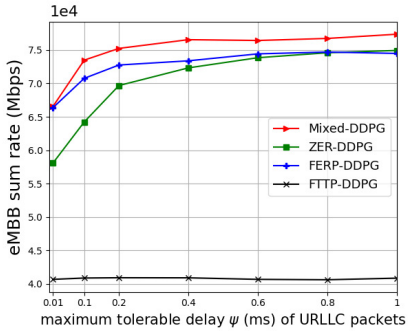


Fig. 4. eMBB sum rate versus the maximum tolerable delay $\psi$ of URLLC packets.

significantly outperforms the other three algorithms for both $\psi = 1$ ms and $\psi = 0.1$ ms because it can dynamically adjust the WET time allocation $\tau$ and the proportion $\rho$ of energy for reservation according to the users' battery level and CSI. Besides, each considered algorithm achieves a smaller reward for a smaller $\psi$, because the delay constraint will be violated more often during the algorithm execution.

In Fig. 2, we plot the eMBB sum rate versus the battery capacity $Q_{\max}$ of the four algorithms within an episode after convergence, where $\phi = 20$ $\mu$W and $\psi = 1$ ms. We can see that the performance of all these algorithms increases as $Q_{\max}$ grows and eventually stabilizes. This is because the larger battery capacity can store more energy for higher data rates, but when the battery capacity is larger than the energy received, the battery capacity no longer affects the data rate.

Fig. 3 depicts the eMBB sum rate versus RF/DC sensitivity $\phi$ of the four algorithms within an episode after convergence, where $Q_{\max} = 25$ $\mu$J and $\psi = 1$ ms. It shows that the proposed algorithm remains the highest eMBB sum rate and that all these four algorithms decrease as $\phi$ increases because a higher $\phi$ leads to less harvested energy and increases the change of energy shortage. We can also find that ZER-DDPG outperforms FERP-DDPG because flexibly changing $\tau$ affects both the amount of energy received in WET and the uplink transmission power in WIT, thereby affecting the system rate.

Fig. 4 depicts the eMBB sum rate versus the maximum tolerable delay $\psi$ of URLLC packets, where $\phi = 20$ $\mu$W and $Q_{\max} = 25$ $\mu$J. The figure shows that the proposed Mixed-DDPG algorithm outperforms the benchmark algorithms in terms of the eMBB sum rate. Moreover, for each considered algorithm apart from FTTP-DDPG, the eMBB sum rate increases as $\psi$ increases because a larger $\psi$ leads to fewer violations of the constraint in the DDPG-based algorithm, which will result in a larger reward and therefore a higher eMBB sum rate. The eMBB sum rate of FTTP-DDPG is limited by its fixed WET time allocation of $\tau = 0.5$, which leaves insufficient time for uplink WIT.

## V. Conclusion

This paper studies the multiplexing of eMBB and URLLC in the uplink WIT powered by downlink WET via preemption-based resource allocation in a WPCN, where the finite battery capacity and the RF/DC sensitivity of the energy-harvesting circuit are considered. The optimization of resource allocation is formulated as a problem that maximizes the eMBB sum rate of all users under all necessary constraints. To tackle this problem, we decompose it into two subproblems and propose a Mixed-DDPG algorithm to solve them alternately. The numerical results reveal that the proposed Mixed-DDPG algorithm can quickly converge to a stable state and achieve a higher eMBB sum rate than the existing schemes, but the performance is sensitive to the transmission time. In our future work, we will extend the proposed model and algorithm to more complex scenarios, such as multi-cell and reconfigurable intelligent surface-aided (RIS-aided) networks.

## References

[1] J. -M. Kang, "Reinforcement Learning Based Adaptive Resource Allocation for Wireless Powered Communication Systems," *IEEE Wireless Commun. Lett.*, vol. 24, no. 8, pp. 1752-1756, Aug. 2020.

[2] H. Lee *et al.*, "Message-Passing-Based Joint User Association and Time Allocation for Wireless Powered Communication Networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 34-47, Jan. 2022.

[3] J. Chen *et al.*, "Resource Allocation for Wireless-Powered IoT Networks With Short Packet Communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1447-1461, Feb. 2019.

[4] A. Guo *et al.*, "Performance Analysis of Energy Harvesting Wireless Communication System With Finite Blocklength," *IEEE Wireless Commun. Lett.*, vol. 20, no. 2, pp. 324-327, Feb. 2016.

[5] A. Anand *et al.*, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477-490, Apr. 2020.

[6] Yan Huang *et al.*, "A Deep-Reinforcement-Learning-Based Approach to Dynamic eMBB/URLLC Multiplexing in 5G NR," in *IEEE Internet of Things*, vol. 7, no. 7, pp. 6439-6456, Jul. 2020.

[7] A. K. Bairagi *et al.*, "Coexistence Mechanism Between eMBB and uRLLC in 5G Wireless Networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736-1749, Mar. 2021.

[8] T. A. Khan *et al.*, "Wirelessly Powered Communication Networks With Short Packets," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5529-5543, Dec. 2017.

[9] H. Ju *et al.*, "Throughput Maximization in Wireless Powered Communication Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 418-428, Jan. 2014.

[10] M. Alsenwi *et al.*, "Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585-4600, July 2021.

[11] K. Liang *et al.*, "Online Power and Time Allocation in MIMO Uplink Transmissions Powered by RF Wireless Energy Transfer," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 6819-6830, Aug. 2017.

[12] J. Li *et al.*, "Deep Reinforcement Learning-Based Joint Scheduling of eMBB and URLLC in 5G Networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1543-1546, Sept. 2020.