Online Policies for Throughput Maximization of Energy-Constrained Wireless-Powered Communication Systems

Xian Li, Xiangyun Zhou, Changyin Sun, and Derrick Wing Kwan Ng

***Abstract*—In this paper, we consider the design of online transmission policies in a single-user wireless-powered communication system over an infinite horizon, aiming at maximizing the long-term system throughput for the user equipment (UE) subject to a given energy budget. The problem is formulated as a constrained Markov decision process (CMDP) problem, which is subsequently converted into an equivalent Markov decision process (MDP) problem via the Lagrangian approach. The corresponding optimal resource allocation policy is obtained through jointly solving the corresponding MDP problem and updating the Lagrangian multiplier. To reduce the complexity, a sub-optimal policy named "quasi-best-effort" is proposed, where the transmit power of the UE is structurally designed so that in each block the UE either exhausts its entire battery energy for transmission or suspends its transmission. To validate the effectiveness of our proposed policy, extensive numerical simulations are conducted with various system parameters. The results show that the proposed quasi-best-effort policy requires far less computation time but achieves a similar long-term throughput performance as the optimal policy.**

***Index Terms*—Wireless-powered communication, long-term throughput, energy budget, constrained Markov decision process.**

I. Introduction

Radio-frequency (RF) energy harvesting (EH) has recently been shown to be a potential technology to provide perpetual energy supply in energy-constrained wireless networks [2]–[4]. As one of the most attractive applications of RF-based EH technology, wireless powered communication network (WPCN) has attracted a lot of attention. To realize an efficient WPCN, in practice, a hybrid access point (H-AP) [5], [6] or a dedicated power beacon (PB) [7], [8] is usually deployed to broadcast wireless energy. Then, based on the "harvest-then-transmit" protocol [6], energy-constrained user equipments (UEs) first harvest energy from the H-AP or PB in downlink during the wireless energy transfer (WET) period and then perform wireless information transfer (WIT) in uplink to transmit their data to the hybrid or separate AP.

The system performance of WPCNs in terms of throughput, outage, and energy efficiency were thoroughly studied in the recent literature [5]–[11]. In [9], the optimal time and power allocation was jointly designed to maximize the throughput of a generalized WPCN, where users are equipped with constant energy supplies along with RF energy harvesting circuits. In order to tackle the "doubly-near-far" problem, user cooperation in WPCNs was proposed in [10], where the user closer to the H-AP was designed to devote part of its time and energy resources to help relay the information of the far user to the H-AP, so as to achieve a more balanced system throughput. In [11], the authors extended the study into a full-duplex (FD) scenario, where the WPCN consists of a FD H-AP and several UEs. By jointly optimizing the time allocation for the WET and WIT as well as the transmit power allocation at the H-AP, the weighted sum-throughput of the system was maximized. However, most of these works assumed that all the harvested RF energy is exhausted immediately within a transmission block without exploiting the possibility of energy storage and considering long-term system performance. In practice, each transceiver in a WPCN is usually equipped with certain energy storage, e.g., battery or capacitor. When the communication channel suffers from deep fading, it is more reasonable to store part of or even all the harvested energy in the battery rather than deplete it instantly. Thus in this paper, we emphasize on the long-term system performance of battery-powered communication networks, in which the harvested energy can be stored and exploited for future operations.

Some research efforts have been devoted to improving the system performance in the long run. In [12], a WPCN in which each wireless node carrying an energy queue and a data queue was studied. Based on the states of the energy and data queues, the Lyapunov optimization technique was applied to design an online stochastic control algorithm, so as to minimize the expected energy consumption while stabilizing the data queues of all nodes. Via jointly utilizing the theory of stochastic geometry (Geo) and dynamic programming (DP), a Geo-DP-based online policy was proposed in [13] to cope with the fluctuations of the on-grid power prices, the amount of the harvested energy, as well as the network traffic loads and thereby minimizing the long-term on-grid energy cost. With the assumption of a simple transmission policy, whereby UE performs data transmission only if the battery level is no smaller than a required value, the spatial throughput maxi-

X. Li and C. Sun are with the School of Automation, Southeast University, Nanjing 210096, China. (email: seulixian@gmail.com, cysun@seu.edu.cn). This work is supported by the National Natural Science Foundation of China (U1713209, 61520106009, 61533008, 61573103).

X. Zhou is with the Research School of Engineering, The Australian National University, Canberra, Australia. (e-mail: xiangyun.zhou@anu.edu.au). The work of X. Zhou is supported by the Australian Research Councils Discovery Project Funding Scheme under Project DP170100939.

D. W. K. Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney N.S.W., Australia (e-mail: w.k.ng@unsw.edu.au). D. W. K. Ng is supported by the Australian Research Council's Discovery Early Career Researcher Award funding scheme (DE170100137).

Part of this work has been submitted to ICC 2019 [1].

mization problem of a large-scale WPCN with multiple UEs and multiple H-APs was investigated in [14]. In [15], two simple online transmission policies for a single-user WPCN were investigated. Considering Nakagami fading channels, the limiting distribution of the stored energy at the EH node as well as the analytical expressions for outage probability were obtained for both policies. By assuming the availability of non-causal channel state information (CSI) and causal CSI at the central controller, an offline algorithm and an online algorithm were proposed to improve the achievable rate over multiple slots for an orthogonal frequency division multiplexing (OFDM)-based WPCN in [16], respectively. In [17], the long-term throughput maximization problem of a two-user WPCN was addressed, where the H-AP was assumed to have an unlimited energy supply. Based on the theory of Markov decision process, the optimal policy as well as an approximate strategy were derived to maximize the system throughput over an infinite time horizon. Moreover, this work was extended to a FD scenario in [18] where the H-AP is able to transfer energy and receive information data simultaneously. Besides, the corresponding optimal policy for the FD case was obtained and the long-term performance gap between the full-duplex WPCN and the half-duplex WPCN was discussed. However, these works do not consider the temporal correlation of the time-varying channels, which can be exploited to improve the system performance. Also, the energy consumption of the H-AP in the previous studies, e.g., [14]–[18], has not been considered in their design.

In this paper, we consider a single-user WPCN where the UE is equipped with a finite-capacity battery. The system model is most closely related to that in [17]. However, there are three fundamental differences. First, the finite state Markov channel (FSMC) model is utilized in our work to capture the temporal-correlation of the channel. Second, the circuit power of the H-AP and the UE, both of which are closely related to the long-term system performance, are considered for the accurate estimation of the system energy consumption. Finally, we devote our effort to the performance optimization in an energy-constrained WPCN, i.e., to maximize the long-term system throughput with a limited energy budget. Our main contributions are summarized as follow.

1) With the consideration of limited system energy budget, we formulate the long-term throughput optimization problem as a constrained Markov decision process (CMDP) problem. The problem of finding the optimal online policy is solved through the Lagrangian approach. Besides, by carefully exploiting the structure of the problem, the number of variables to be numerically optimized is reduced, such that the computational complexity is cut down without loss of system performance.

2) The offline version of the CMDP problem is considered and the property of the corresponding optimal solution is analyzed. Through mimicking the structure of the offline optimal solution, we propose a sub-optimal online policy named "quasi-best-effort", which either exhausts the battery at the UE or keeps the UE silent in each time block.

3) Numerical simulations in terms of the long-term throughput and the computation time performance are carried out in different practical scenarios for the comparison between the proposed quasi-best-effort policy and the optimal policy. The results demonstrate that the proposed sub-optimal policy shows comparable performance to the optimal policy but requires considerable less computation time.

II. System model

In this paper, we consider a WPCN consisting of a H-AP and a single-antenna UE equipped with a rechargeable battery. The system model is shown in Fig. 1. The H-AP is equipped with a directional antenna. The channel between the H-AP and the UE is assumed to be block fading and time-correlated, e.g., the channel remains constant in each time block but varies from block to block. The channel power gain in block t is $H_t = \theta_t d^{-\alpha}$, where θ_t captures the multipath fading, d is the distance between the H-AP and the UE, and α is the path loss exponent. In each block, a WET period is followed by a WIT period. The UE first harvests energy from the H-AP and store it into the battery during WET. Then during the following WIT period, the UE transmits its data to the H-AP using the stored energy in the battery. We assume that in the current block, perfect CSI as well as the information of the UE's battery state is available at the H-AP for resource allocation. In practice, this information can be acquired in the training phase at the beginning of each block. In particular, the battery state of the UE and the CSI can be fed back to the H-AP. Compared to the durations of WET and WIT, the time and energy spent in training phase are very small and thus are neglected in our work. Moreover, to model the imperfect operation of the UE, similar to [19], we assume that the UE has the probability of $\lambda \in [0, 1)$ to survive the hardware (or software) failure and continue to operate normally in a block.

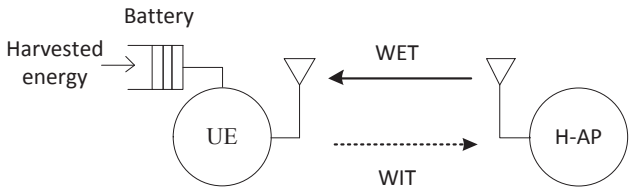


Fig. 1: System model of a WPCN.

In practical applications, the communication system is required to be optimized under some given constraints [14], [20]. In this paper, we investigate the problem of long-term system throughput optimization under a given energy constraint and focus on the design of online policies. The considered problem can be formulated as a CMDP problem (see details in Sec. III) through the tuple, $\{\mathcal{S}, \mathcal{A}, \boldsymbol{P}, r(\cdot), e(\cdot)\}$, where $\mathcal{S}$ is the system state space, $\mathcal{A}$ is the action space, $\boldsymbol{P}$ is the probability transition matrix, and $r(\cdot)$ and $e(\cdot)$ are the reward function and the cost function, respectively. In the following, detailed descriptions of these five elements are provided.

A. System States

For the considered system, in block t, the system state $\boldsymbol{s}_t \in \mathcal{S}$ contains the channel and battery information, i.e.,

$\boldsymbol{s}_t = [h_t, b_t]$, where h_t is the channel state in block t and b_t is the battery state in block t, respectively. Quantized channel power gain and battery energy storage are considered in this paper. Thus the number of system state is finite and countable. Specifically, the system state space $\mathcal{S}$ can be expressed as $\mathcal{S} = \mathcal{H} \times \mathcal{B}$, where $\mathcal{H} \triangleq \{1, 2, ..., k, ..., K\}$ and $\mathcal{B} \triangleq \{0, 1, 2, ..., l, ..., L\}$ define the set of channel state and battery state, respectively. The battery is at state 0 when it is exhausted.

For channel state $h_t \in \mathcal{H}$, a finite-state Markov channel model is used to describe the time-varying behaviour of the fading channel [21]–[27]. Denote by $\mathbf{\Gamma} = [\Theta_1, \Theta_2, ..., \Theta_k, ...\Theta_{K+1}] \times d^{-\alpha}$ the channel gain boundaries in the increasing order with $\Theta_1 = 0$ and $\Theta_{K+1} = \infty$. The channel state in the t-th block is said to be at state k (i.e., $h_t = k$), if $\Theta_k \leq \theta_t < \Theta_{k+1}$. That is to say, a range of channel power gain values belong to the same channel state.

We assume that there is only a one-step channel state transition from block to block. Denoting π_k as the steady state probability of the channel being in state k and assuming equiprobable partition[1] of the channel gain, i.e. $\pi_k = \frac{1}{K}, \forall k \in \{1, 2, ..., K\}$, the channel gain boundaries can be found by solving the following equations:

$$\pi_k = \int_{\Theta_k}^{\Theta_{k+1}} \rho(\theta_t) d\theta_t = \frac{1}{K}, \forall k \in \{1, 2, ..., K\}, \tag{1}$$

where $\rho(\theta_t)$ is the probability density function of the variable θ_t. Then for the channel state k, i.e., $h_t = k$, the quantized value of the channel gain is

$$\bar{H}_t = \frac{\int_{\Theta_k}^{\Theta_{k+1}} H_t \rho(\theta_t) \mathrm{d}\theta_t}{\int_{\Theta_k}^{\Theta_{k+1}} \rho(\theta_t) \mathrm{d}\theta_\mathrm{t}} = \frac{\int_{\Theta_k}^{\Theta_{k+1}} \theta_t d^{-\alpha} \rho(\theta_t) \mathrm{d}\theta_t}{\pi_k}. \tag{2}$$

Similarly, the available energy in the battery of the UE is discretized into $L + 1$ states. Denote $B_{\max} = LQ$ as the maximum capacity of the battery, where Q represents an energy quantum level. In the t-th block, the battery state is at state l (i.e., $b_t = l$), if $\lfloor \frac{B_t}{Q} \rfloor = l$, where B_t is the available battery energy at the beginning of block t. As a special case, the battery state is $b_{\max} = L$ when the battery is full.

B. Actions, Reward, and Cost Functions

At the beginning of each block, the H-AP makes a decision according to the current system state. Since both WET and WIT are performed in each block, the time duration of each block T is divided into two orthogonal time slots: τ_t^E for WET and τ_t^I for WIT with $\tau_t^\mathrm{E} + \tau_t^\mathrm{I} \leq T$. Correspondingly, the action adopted at the beginning of block t (denoted by $\boldsymbol{a}_t$) contains four elements, i.e., $\boldsymbol{a}_t = \{\tau_t^\mathrm{E}, \tau_t^\mathrm{I}, P_t^\mathrm{E}, P_t^\mathrm{I}\}$, where P_t^E and P_t^I are the transmit power of the H-AP and the transmit power of the UE, respectively.

When an action is performed in block t, the system receives an immediate reward but also incurs an immediate cost. Specifically, in our work, the immediate reward and immediate cost refer to the throughput per block (defined as the data conveyed in one block) and the energy consumption per block, respectively. Thus for a given state $\boldsymbol{s}_t$ and an action $\boldsymbol{a}_t \in \mathcal{A}(\boldsymbol{s}_t)$, where $\mathcal{A}(\boldsymbol{s}_t)$ is the feasible action set at state $\boldsymbol{s}_t$, the immediate reward, i.e., $r(\boldsymbol{s}_t, \boldsymbol{a}_t) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is given as

$$r(\boldsymbol{s}_t, \boldsymbol{a}_t) = \frac{\int_{\Theta_k}^{\Theta_{k+1}} \tau_t^\mathrm{I} W \log_2 \left(1 + \frac{P_t^\mathrm{I} \theta_t d^{-\alpha}}{\zeta \sigma^2}\right) \rho(\theta_t) \mathrm{d}\theta_\mathrm{t}}{\pi_k}, \tag{3}$$

where W is the bandwidth of the considered system, $\sigma^2 = N_0 W$ is the thermal noise power (where N_0 represents the noise power spectral density), and ζ is a correction factor characterizing the gap between the achievable rate and the channel capacity due to the use of practical modulation and coding schemes [8].

For a WPCN operating in an infinite horizon, the UE can store part of or all the harvested energy in block t in the battery for future use, or utilize the energy stored in the previous blocks to transmit data (even there is no WET in the current block). In this block, the immediate cost of the system, which ties up with the energy cost and the energy accumulated at the UE, is not always identical to the energy cost of the H-AP. As a result, the immediate cost of the system at block t, i.e., $e(\boldsymbol{s}_t, \boldsymbol{a}_t) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is expressed as

$$e(\boldsymbol{s}_t, \boldsymbol{a}_t) = \frac{P_t^\mathrm{E} \tau_t^\mathrm{E}}{\vartheta_\mathrm{AP}} + P_{\mathrm{C_{AP}}} \tau_t^\mathrm{E} + e_t^\mathrm{IT} - e_t^\mathrm{AC}, \tag{4}$$

where the first two terms are the energy consumption at the H-AP and the last two terms give the battery consumption (which can be either positive or negative) at the UE. Specifically, $0 < \vartheta_\mathrm{AP} < 1$ is the power amplifier efficiency of the H-AP. Hence the first term in (4) stands for the energy consumption of the power amplifier due to WET. $P_{\mathrm{C_{AP}}}$ is the circuit power at the H-AP. Hence the second term in (4) stands for the energy consumption of the circuit during WET. For the energy consumption of the UE,

$$e_t^\mathrm{IT} = \frac{P_t^\mathrm{I} \tau_t^\mathrm{I}}{\vartheta_\mathrm{U}} + P_{\mathrm{C_U}} \tau_t^\mathrm{I} \tag{5}$$

is the energy consumption of the UE during WIT, where $0 < \vartheta_\mathrm{U} < 1$ denotes the power amplifier efficiency of the UE and $P_{\mathrm{C_U}}$ denotes the circuit power at the UE, respectively. Finally,

$$e_t^\mathrm{AC} = \min(B_t + e_t^\mathrm{EH}, B_{\max}) - B_t \tag{6}$$

is the energy accumulated at the UE during WET in block t. Since the stored energy at the UE is limited by its maximum battery capacity, the first part of (6) represents the battery level after WET in block t, where

$$e_t^\mathrm{EH} = \eta G_\mathrm{A} \bar{H}_t P_t^\mathrm{E} \tau_t^\mathrm{E} \tag{7}$$

is the corresponding harvested energy during WET at the UE, in which η is the energy conversion efficiency, and G_A is the antenna gain at the H-AP.

Apparently, the maximum value of e_t^AC is e_t^EH. By conservation of energy, the energy harvested at the UE is always no more than the energy cost at the H-AP, i.e., $\frac{P_t^\mathrm{E} \tau_t^\mathrm{E}}{\vartheta_\mathrm{AP}} + P_{\mathrm{C_{AP}}} \tau_t^\mathrm{E} \geq e_t^\mathrm{EH} \geq e_t^\mathrm{AC}$. Therefore, both $e(\boldsymbol{s}_t, \boldsymbol{a}_t)$ and $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$

[1] Although this is a reasonable and commonly adopted assumption (e.g., [24]–[27]) in a FSMC model, imposing this assumption is not necessary in our work. For any other partition methods, the corresponding optimal and suboptimal policies can be easily established by following a similar approach as in this paper.

are non-negative. Denote $P_{\max}^{\mathrm{E}}$ as the maximum transmit power of the H-AP. The maximum available energy of the UE in slot t cannot exceed its current available energy, i.e., $e_t^{\mathrm{IT}} \leq \min(B_t + e_t^{\mathrm{EH}}, B_{\max})$. The feasible action set at state $\boldsymbol{s}_t$ is:

$$\begin{aligned}\mathcal{A}(\boldsymbol{s}_t) = \{\boldsymbol{a}_t | & \tau_t^{\mathrm{E}} + \tau_t^{\mathrm{I}} \leq T, \tau_t^{\mathrm{E}} \geq 0, \tau_t^{\mathrm{I}} \geq 0, \\ & 0 \leq P_t^{\mathrm{E}} \leq P_{\max}^{\mathrm{E}}, P_t^{\mathrm{I}} \geq 0, e_t^{\mathrm{IT}} \leq \min(B_t + e_t^{\mathrm{EH}}, B_{\max})\}.\end{aligned} \tag{8}$$

C. Transition Probabilities

By denoting $\boldsymbol{s}_t$ and $\boldsymbol{s}_{t+1}$ as the state in block t and $t+1$, respectively, the probability of transiting from state $\boldsymbol{s}_t$ to state $\boldsymbol{s}_{t+1}$ given an action $\boldsymbol{a}_t$ can be expressed as

$$\begin{aligned}\mathcal{P}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) &\overset{(a)}{=} \mathcal{P}(h_{t+1}, b_{t+1}|h_t, b_t, \boldsymbol{a}_t) \\ &\overset{(b)}{=} \mathcal{P}(h_{t+1}|h_t)\mathcal{P}(b_{t+1}|h_t, b_t, \boldsymbol{a}_t),\end{aligned} \tag{9}$$

where (a) holds by definition and (b) holds since the channel state evolution is independent of the battery state and the action.

As stated in [21], the channel state transition probability from state h_t to h_{t+1} can be described by the level crossing rate $\Lambda(\Theta_{\mathrm{b}})$, which is defined as the average number of times that the instantaneous value of θ_t crosses a given level Θ_{b}. Specifically, the transition probabilities can be approximated by the ratio of $\Lambda(\Theta_{\mathrm{b}})$ divided by the average number of blocks the value of θ_t falls in the interval associated with the state $h_t = k$. In our work, we assume that the channel state transits between its adjacent state only. This assumption, which is verified in [21], is commonly adopted in the related existing works, e.g., [22]–[26], to simplify the model in capturing the transition between channel states. Subsequently, the transition probabilities can be approximated as

$$\mathcal{P}(h_{t+1} = k+1|h_t = k) \approx \frac{\Lambda(\Theta_{k+1})T}{\pi_k}, \tag{10}$$

$$\mathcal{P}(h_{t+1} = k-1|h_t = k) \approx \frac{\Lambda(\Theta_{k-1})T}{\pi_k}, \tag{11}$$

$$\mathcal{P}(h_{t+1} = k|h_t = k) \approx 1 - \frac{\Lambda(\Theta_{k+1})T}{\pi_k} - \frac{\Lambda(\Theta_{k-1})T}{\pi_k}. \tag{12}$$

On the other hand, the battery state transition probability can be determined as follows. If $b_{t+1} < b_{\max}$,

$$\mathcal{P}(b_{t+1}|h_t, b_t, \boldsymbol{a}_t) = \chi\left\{b_t + \left\lfloor \frac{e_t^{\mathrm{AC}} b_{\max}}{B_{\max}} - \frac{e_t^{\mathrm{IT}} b_{\max}}{B_{\max}} \right\rfloor = b_{t+1}\right\}, \tag{13}$$

otherwise,

$$\mathcal{P}(b_{\max}|h_t, b_t, \boldsymbol{a}_t) = \chi\left\{b_t + \left\lfloor \frac{e_t^{\mathrm{AC}} b_{\max}}{B_{\max}} - \frac{e_t^{\mathrm{IT}} b_{\max}}{B_{\max}} \right\rfloor \geq b_{\max}\right\}, \tag{14}$$

where $\chi(\cdot)$ is the indicator function.

III. CMDP Formulation and The Optimal Policy

The problem formulation and its corresponding optimal solution is shown in this section.

A. Problem Formulation

Define the decision rule in block t as a function mapping from the system state $\boldsymbol{s}$ to the action to be taken, i.e., $\mu_t : \mathcal{S} \to \mathcal{A}$. A policy $\boldsymbol{\mu} = \{\mu_1, \mu_2, ...\}$ is a sequence of decision rules. If the decision rule in a policy does not depend on time, i.e., $\mu_1 = \mu_2 = ...$, then the policy is stationary. It is known that a policy is a pure policy if it is stationary and deterministic. For an admissible stationary policy $\boldsymbol{\mu}$, the long-term throughput is defined as

$$R(\boldsymbol{s}_0, \boldsymbol{\mu}) = (1-\lambda)\sum_{t=1}^{\infty} \lambda^{t-1} \mathbb{E}_{\boldsymbol{s}_0}^{\boldsymbol{\mu}}\{r(\boldsymbol{s}_t, \boldsymbol{a}_t)\}, \tag{15}$$

and the long-term energy cost is defined as

$$E(\boldsymbol{s}_0, \boldsymbol{\mu}) = (1-\lambda)\sum_{t=1}^{\infty} \lambda^{t-1} \mathbb{E}_{\boldsymbol{s}_0}^{\boldsymbol{\mu}}\{e(\boldsymbol{s}_t, \boldsymbol{a}_t)\}. \tag{16}$$

The normalization factor $(1-\lambda)$ is introduced here to avoid the situation where, for fixed immediate reward r and cost d, the values of these two functions become very large if λ is close to one. As shown in [28], when the discount factor λ approaches 1, the infinite horizon discounted costs defined in (15) and (16) converge to their corresponding infinite horizon expected average costs, which are in the form of $\overline{\lim}_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}_{\boldsymbol{s}_0}^{\boldsymbol{\mu}}\{X(\boldsymbol{s}_t, \boldsymbol{a}_t)\}$, where N is the number of blocks and $X \in \{r, e\}$. Accordingly, (15) and (16) can be interpreted as the expected average throughput and expected average energy cost per block, respectively.

Our purpose is to find an optimal policy $\boldsymbol{\mu}^*$ such that the long-term throughput is maximized satisfying the maximum allowed energy cost E_{th}. This can be formulated as a CMDP problem as below:

$$\max_{\boldsymbol{\mu}} \quad R(\boldsymbol{s}_0, \boldsymbol{\mu}) \tag{17a}$$

$$\text{s. t.} \quad E(\boldsymbol{s}_0, \boldsymbol{\mu}) \leq E_{\mathrm{th}}. \tag{17b}$$

B. The Optimal Policy

As shown in [28], a CMDP problem can be transferred into an equivalent unconstrained MDP problem using the Lagrangian approach. By introducing the Lagrangian multiplier β with $\beta > 0$ for the CMDP problem in (17), a new reward function $\widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^+ \to \mathbb{R}$, can be defined as

$$\widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) = r(\boldsymbol{s}, \boldsymbol{a}) - \beta e(\boldsymbol{s}, \boldsymbol{a}) \tag{18}$$

and the corresponding Bellman's optimality equation is:

$$J_\beta(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\mathcal{A}(\boldsymbol{s})}\left\{(1-\lambda)\widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) + \lambda \sum_{\boldsymbol{s}'\in\mathcal{S}} \mathcal{P}(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) J_\beta(\boldsymbol{s}')\right\}, \tag{19}$$

which can be solved via the Value Iteration Algorithm (VIA) [29] for any fixed β. The corresponding optimal policy, i.e., $\boldsymbol{\mu}_\beta = \{\mu_\beta(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$, can be computed by:

$$\mu_\beta(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}\in\mathcal{A}(\boldsymbol{s})}\left\{(1-\lambda)\widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) + \lambda \sum_{\boldsymbol{s}'\in\mathcal{S}} \mathcal{P}(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) J_\beta(\boldsymbol{s}')\right\}. \tag{20}$$

As described in [28], the optimal policy of a CMDP problem with a single constraint is a randomized mixture of two pure policies (i.e., $\boldsymbol{\mu}_{\beta^-} = \{\mu_{\beta^-}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$ and $\boldsymbol{\mu}_{\beta^+} = \{\mu_{\beta^+}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$, with β^- and β^+ as their associated multipliers, respectively). The policy $\boldsymbol{\mu}_{\beta^-}$ satisfies the energy constraint as much as possible from below, while the policy $\boldsymbol{\mu}_{\beta^+}$ breaks the energy constraint but is as close to it as possible from above. Since $J_\beta(\boldsymbol{s})$ is monotonically non-increasing in β [30], the values of β^- and β^+ can be found through the bisection search method. Define $q \in [0, 1]$ as the mixing weight parameter, $E_{\beta^-} \leq E_{\text{th}}$, and $E_{\beta^+} \geq E_{\text{th}}$ as the costs of the policies $\boldsymbol{\mu}_{\beta^-}$ and $\boldsymbol{\mu}_{\beta^+}$, respectively. Then the optimal policy, i.e., $\boldsymbol{\mu}^* = \{\mu^*(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$, is given by:

$$\mu^*(\boldsymbol{s}) = \begin{cases} \mu_{\beta^-}(\boldsymbol{s}), & \text{w.p. } q \quad (21) \\ \mu_{\beta^+}(\boldsymbol{s}), & \text{w.p. } 1-q \quad (22) \end{cases}$$

where q can be obtained through solving equation $E_{\text{th}} = qE_{\beta^-} + (1-q)E_{\beta^+}$.

C. Complexity Reduction

Although the method proposed in the last section can achieve a globally optimal solution, for every step of VIA, it is required to perform an exhaustive search over all feasible actions for every state. Hence, reducing the number of variables in actions and narrowing down the value range of variables are desirable so as to alleviate the computation burden.

Lemma 1. The H-AP always transmits with its maximum available power in all the time blocks, i.e., $P_t^{\text{E}} = P_{\max}^{\text{E}}, \forall t = 1, 2, ...$

Proof. We prove this by contradiction. Assume that the maximum throughput $\widetilde{R}^*$ is obtained when an action $\widetilde{\boldsymbol{a}}_t^*$ is adopted in block t, where $\widetilde{\boldsymbol{a}}_t^* = \{\widetilde{\tau}_t^{\text{E}}, \widetilde{\tau}_t^{\text{I}}, \widetilde{P}_t^{\text{E}}, \widetilde{P}_t^{\text{I}}\}$ with $\widetilde{P}_t^{\text{E}} < P_{\max}^{\text{E}}$. Let $\boldsymbol{a}_t^* = \{\tau_t^{\text{E}*}, \tau_t^{\text{I}*}, P_{\max}^{\text{E}}, P_t^{\text{I}*}\}$ be another feasible action in block t with the same total energy consumption, i.e., $e(\boldsymbol{s}_t, \boldsymbol{a}_t^*) = e(\boldsymbol{s}_t, \widetilde{\boldsymbol{a}}_t^*)$, where $\tau_t^{\text{E}*} P_{\max}^{\text{E}} = \widetilde{\tau}_t^{\text{E}} \widetilde{P}_t^{\text{E}}$. Denote the throughput corresponding to a_t^* by R^*. Since $\widetilde{P}_t^{\text{E}} < P_{\max}^{\text{E}}$, we always have $\tau_t^{\text{E}*} < \widetilde{\tau}_t^{\text{E}}$ and $P_{\text{C}_{\text{AP}}} \tau_t^{\text{E}*} < P_{\text{C}_{\text{AP}}} \widetilde{\tau}_t^{\text{E}}$. Then according to (4), the energy used in WIT stage by taking action $\boldsymbol{a}_t^*$ must be higher than that by taking action $\widetilde{a}_t^*$, i.e., $e_t^{\text{IT}}(a_t^*) > e_t^{\text{IT}}(\widetilde{a}_t^*)$. In other words, we can always have $R^* > \widetilde{R}^*$, which contradicts the assumption. Thus H-AP always transmits with its maximum available power in all the time blocks. □

Lemma 2. The optimal time duration for energy transfer in block t, denoted as $\tau_t^{\text{E}*}$, is limited by $0 \leq \tau_t^{\text{E}*} \leq \frac{B_{\max} - B_t}{\eta G_{\text{A}} P_{\max}^{\text{E}} \bar{H}_t}$.

Proof. The proof is intuitive. Remind that the transmit power of the H-AP is fixed at $P_{\max}^{\text{E}}$ and the battery capacity at the UE is limited by $B_{\max}$. Let $\tau_t^{\text{E}\prime} = \frac{B_{\max} - B_t}{\eta G_{\text{A}} P_{\max}^{\text{E}} \bar{H}_t}$. Any $\tau_t^{\text{E}} > \tau_t^{\text{E}\prime}$ yields the same available energy at the UE (i.e., $B_{\max}$), but incurs higher system cost, which spends more energy budget but makes no contribution to the system throughput performance. Thus the optimal value of τ_t^{E} should be in the range of $0 \leq \tau_t^{\text{E}*} \leq \frac{B_{\max} - B_t}{\eta G_{\text{A}} P_{\max}^{\text{E}} \bar{H}_t}$. □

Since the optimal policy of the CMDP problem (17) consists of two pure policies, both of which are independent of time sequence. Here, we drop the subscript 't' for convenience. By exploiting *Lemma* 1 and *Lemma* 2, the action $\boldsymbol{a} = \{\tau^{\text{E}}, \tau^{\text{I}}, P^{\text{E}}, P^{\text{I}}\}$ in the CMDP problem (17) is simplified as $\boldsymbol{a}' = \{\tau^{\text{E}}, \tau^{\text{I}}, P^{\text{I}}\}$. Subsequently, the feasible action set at state $\boldsymbol{s} = [h, b]$ shrinks to $\mathcal{A}'(\boldsymbol{s}) = \{\{\tau^{\text{E}}, \tau^{\text{I}}, P^{\text{I}}\} | \tau^{\text{E}} + \tau^{\text{I}} \leq T, 0 \leq \tau^{\text{E}} \leq \frac{B_{\max} - bQ}{\eta G_{\text{A}} P_{\max}^{\text{E}} \bar{H}(h)}, \tau^{\text{I}} \geq 0, P^E = P_{\max}^{\text{E}}, P^{\text{I}} \geq 0, e^{\text{IT}} \leq \min(bQ + e^{\text{EH}}, B_{\max})\}$, where $\bar{H}(h)$ is the expectation of H_t with channel state h and is computed by (2). The corresponding algorithm solving (17) is described in Algorithm 1. Specifically, initializations are performed in lines 1-5, where n and ε_β are the iteration sequence and the error bound for updating β, m and ε_J are the iteration sequence and the error bound for the operation of the VIA, respectively. In lines 6-7, the VIA is conducted to solve the corresponding unconstrained MDP problem, where $\boldsymbol{J}_\beta^m = (J_\beta^m(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S})$, $\boldsymbol{J}_\beta^{m+1} = (J_\beta^{m+1}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S})$, and the norm function is defined as $\|\boldsymbol{J}_\beta\| = \max |J_\beta(\boldsymbol{s})|$ for $\boldsymbol{s} \in \mathcal{S}$. Then the optimal policy of a given β is selected in line 8, where $\mu_\beta(\boldsymbol{s}) = (\tau^{\text{E}}(\boldsymbol{s}; \beta), \tau^{\text{I}}(\boldsymbol{s}; \beta), P^{\text{I}}(\boldsymbol{s}; \beta))$ represents the corresponding optimal decision rule at system state $\boldsymbol{s}$. That is, for the given β, at state $\boldsymbol{s}$, the H-AP performs WET for duration of $\tau^{\text{E}}(\boldsymbol{s}; \beta)$ with transmit power $P_{\max}^{\text{E}}$ and the UE transmits data for duration of $\tau^{\text{I}}(\boldsymbol{s}; \beta)$ with transmit power $P^{\text{I}}(\boldsymbol{s}; \beta)$. For updating the value of β, the bisection search is carried out in lines 9-16 and is stopped until the updating tolerance satisfies that $|\beta^{n+1} - \beta^n| < \varepsilon_\beta$ (see line 17). After that, with the obtained policy $\mu_{\beta^-}(\boldsymbol{s})$ and $\mu_{\beta^+}(\boldsymbol{s})$, the mixing weight q is computed in line 21. Finally, the optimal reward and the optimal policy are obtained in line 22.

Algorithm 1 The Optimal Policy for the CMDP (17)

1: Set $n = 0$, $\beta^- = 0$, β^+, $\beta^0 = \beta^-$, specify $\varepsilon_\beta > 0$ and $\varepsilon_J > 0$.
2: **repeat**
3: Set $\beta = \beta^n$.
4: Set $n = n + 1$ and $m = 0$.
5: Initialize $J_\beta^0(\boldsymbol{s}) = 0$ for each $\boldsymbol{s} \in \mathcal{S}$.
6: For each $\boldsymbol{s} \in \mathcal{S}$, compute $J_\beta^{m+1}(\boldsymbol{s})$ by

$$J_\beta^{m+1}(\boldsymbol{s}) := \max_{\boldsymbol{a} \in \mathcal{A}'(\boldsymbol{s})} \left\{ (1-\lambda) \widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) + \lambda \sum_{\boldsymbol{s}' \in \mathcal{S}} \mathcal{P}(\boldsymbol{s}' | \boldsymbol{s}, \boldsymbol{a}) J_\beta^m(\boldsymbol{s}') \right\}. \quad (23)$$

7: If $\|\boldsymbol{J}_\beta^m - \boldsymbol{J}_\beta^{m+1}\| < \varepsilon_J (1-\lambda)/2\lambda$, go to line 8. Otherwise, set $m = m + 1$ and go to line 6.
8: For each $\boldsymbol{s} \in \mathcal{S}$, select the policy

$$\mu_\beta(\boldsymbol{s}) := \arg\max_{\boldsymbol{a} \in \mathcal{A}'(\boldsymbol{s})} \left\{ (1-\lambda) \widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) + \lambda \sum_{\boldsymbol{s}' \in \mathcal{S}} \mathcal{P}(\boldsymbol{s}' | \boldsymbol{s}, \boldsymbol{a}) J_\beta^{m+1}(\boldsymbol{s}') \right\}. \quad (24)$$

9: Compute the stationary distribution $\Psi(\boldsymbol{s})$ induced by $\boldsymbol{\mu}_\beta(\boldsymbol{s}) = \{\mu_\beta(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$.
10: **if** $\sum_{\boldsymbol{s} \in \mathcal{S}} \Psi(\boldsymbol{s}) e(\boldsymbol{s}, \mu_\beta(\boldsymbol{s})) > E_{\text{th}}$ **then**
11: $\beta^{n+1} = \frac{\beta^+ + \beta^n}{2}$.
12: $\beta^- = \beta^n$.
13: **else**
14: $\beta^{n+1} = \frac{\beta^- + \beta^n}{2}$.

15: $\beta^+ = \beta^n$.
16: **end if**
17: **until** $|\beta^{n+1} - \beta^n| < \varepsilon_\beta$.
18: Select the policies $\mu_{\beta^-}(\boldsymbol{s})$ and $\mu_{\beta^+}(\boldsymbol{s})$ based on (24) with obtained β^- and β^+, respectively.
19: Compute the stationary distribution $\Psi_{\beta^-}(\boldsymbol{s})$ and $\Psi_{\beta^+}(\boldsymbol{s})$ induced by $\boldsymbol{\mu}_{\beta^-}(\boldsymbol{s})$ and $\boldsymbol{\mu}_{\beta^+}(\boldsymbol{s})$, respectively.
20: Compute

$$R_{\beta^-} = \sum_{\boldsymbol{s}\in\mathcal{S}} \Psi_{\beta^-}(\boldsymbol{s}) r(\boldsymbol{s}, \mu_{\beta^-}(\boldsymbol{s})), \tag{25}$$

$$R_{\beta^+} = \sum_{\boldsymbol{s}\in\mathcal{S}} \Psi_{\beta^+}(\boldsymbol{s}) r(\boldsymbol{s}, \mu_{\beta^+}(\boldsymbol{s})), \tag{26}$$

$$E_{\beta^-} = \sum_{\boldsymbol{s}\in\mathcal{S}} \Psi_{\beta^-}(\boldsymbol{s}) e(\boldsymbol{s}, \mu_{\beta^-}(\boldsymbol{s})), \tag{27}$$

$$E_{\beta^+} = \sum_{\boldsymbol{s}\in\mathcal{S}} \Psi_{\beta^+}(\boldsymbol{s}) e(\boldsymbol{s}, \mu_{\beta^+}(\boldsymbol{s})). \tag{28}$$

21: Compute q by solving $E_{\text{th}} = qE_{\beta^-} + (1-q)E_{\beta^+}$.
22: Obtian the optimal reward $R = qR_{\beta^-} + (1-q)R_{\beta^+}$ and the optimal policy for each $\boldsymbol{s} \in \mathcal{S}$

$$\mu^*(\boldsymbol{s}) = \begin{cases} \mu_{\beta^-}(\boldsymbol{s}), & \text{w.p. } q \quad (29) \\ \mu_{\beta^+}(\boldsymbol{s}), & \text{w.p. } 1-q \quad (30) \end{cases}$$

According to [31], the running time for solving a MDP using VIA is $O(\frac{1}{1-\lambda}\log(\frac{1}{1-\lambda})|\mathcal{A}||\mathcal{S}|^2)$. For the implementation of VIA, the variables in actions are quantized. Specifically, τ^{E}, τ^{I}, P^{E}, P^{I} are discretized into levels of V_τ^{E}, V_τ^{I}, V_P^{E} and V_P^{I}, respectively, which indicates that there are at most $V_\tau^{\text{E}} V_\tau^{\text{I}} V_P^{\text{E}} V_P^{\text{I}}$ candidate actions in a state before applying the reduction on the action space. Through using *Lemma* 1 and *Lemma* 2, the cardinality of the action space is reduced from $|\mathcal{A}| = V_\tau^{\text{E}} V_\tau^{\text{I}} V_P^{\text{E}} V_P^{\text{I}} |\mathcal{S}|$ to $|\mathcal{A}'| = V_\tau^{\text{E}} V_\tau^{\text{I}} V_P^{\text{I}} |\mathcal{S}|$, thus resulting in less computation time. Also, since the update of β is independent from the action space and the channel state space, the computational complexity of Algorithm 1 is $O(\frac{1}{1-\lambda}\log(\frac{1}{1-\lambda}) V_\tau^{\text{E}} V_\tau^{\text{I}} V_P^{\text{I}} |\mathcal{S}|^3)$.

IV. The Sub-optimal Policy

Although the complexity of the optimal policy is reduced through *Lemma* 1 and *Lemma* 2, it is still time consuming when the number of system states is large. Thus it is natural to design less complex policies even at the cost of some performance degradation. In this section, we propose a sub-optimal policy named the quasi-best-effort policy whose performance is close to the optimal policy but with lower complexity. We derive this sub-optimal policy by using the insights obtained from the optimal offline policy, which will be discussed next.

Let us consider the long-term throughput maximization problem from an offline point of view, in which the system has a priori knowledge of the channel condition and battery condition in each block. Let $N \to \infty$, $E_t^{\text{I}} = P_t^{\text{I}} \tau_t^{\text{I}}$ and $\gamma_t = \frac{H_t}{\zeta\sigma^2}$, the offline optimization problem corresponding to problem (17) can be written as

$$\max_{\tau_t^{\text{I}}, \tau_t^{\text{E}}, E_t^{\text{I}}, \forall t} \ (1-\lambda)\sum_{t=1}^{N} \lambda^{t-1} \tau_t^{\text{I}} W \log_2(1 + \frac{E_t^{\text{I}}}{\tau_t^{\text{I}}}\gamma_t) \tag{31a}$$

$$\text{s.t.} \quad (1-\lambda)\sum_{t=1}^{N} \lambda^{t-1} \left[\frac{E_t^{\text{I}}}{\vartheta_{\text{U}}} + P_{\text{C}_\text{U}} \tau_t^{\text{I}} + \frac{P_{\max}^{\text{E}} \tau_t^{\text{E}}}{\vartheta_{\text{AP}}} + P_{\text{C}_\text{AP}} \tau_t^{\text{E}} - \eta G_{\text{A}} H_t P_{\max}^{\text{E}} \tau_t^{\text{E}} \right] \le E_{\text{th}}, \tag{31b}$$

$$\sum_{i=1}^{t} (\frac{E_i^{\text{I}}}{\vartheta_{\text{U}}} + P_{\text{C}_\text{U}} \tau_i^{\text{I}}) \le B_0 + \sum_{i=1}^{t} \eta G_{\text{A}} H_i P_{\max}^{\text{E}} \tau_i^{\text{E}}, \quad \forall t = 1, 2, ..., N, \tag{31c}$$

$$B_0 + \sum_{i=1}^{t} \eta G_{\text{A}} H_i P_{\max}^{\text{E}} \tau_i^{\text{E}} - \sum_{i=1}^{t-1} (\frac{E_i^{\text{I}}}{\vartheta_{\text{U}}} + P_{\text{C}_\text{U}} \tau_i^{\text{I}}) \le B_{\max}, \quad \forall t = 1, 2, ..., N, \tag{31d}$$

$$\tau_t^{\text{I}} + \tau_t^{\text{E}} \le T, \forall t = 1, 2, ..., N, \tag{31e}$$

$$E_t^{\text{I}}, \tau_t^{\text{I}}, \tau_t^{\text{E}} \ge 0, \forall t = 1, 2, ..., N, \tag{31f}$$

where (31b) is the energy budget constraint corresponding to (17b). (31c)-(31e) describe the constraints on the available energy at the UE, the battery capacity and the time duration, respectively. Note that the result in *Lemma* 1 is used here, i.e., $P_t^{\text{E}} = P_{\max}^{\text{E}}, \forall t = 1, ..., N$. In the case of infinite horizon (i.e., $N \to \infty$), the total number of variables tends to be infinite and it is intractable to obtain the optimal offline solution. Nevertheless, it can provide us some useful insights on the design of sub-optimal online policy.

Since (31a) is a concave function and all the constraints (31b)-(31f) are affine, problem (31) is a typical convex optimization problem. Its Lagrangian function can be written as

$$\begin{aligned} &\mathcal{L}_{\text{QBE}}(\boldsymbol{\tau}^{\mathbf{E}}, \boldsymbol{\tau}^{\mathbf{I}}, \boldsymbol{E}^{\mathbf{I}}, \delta, \boldsymbol{\phi}, \boldsymbol{\upsilon}, \boldsymbol{\psi}) \\ &= (1-\lambda)\sum_{t=1}^{N} \lambda^{t-1} \tau_t^{\text{I}} W \log_2(1 + \frac{E_t^{\text{I}}}{\tau_t^{\text{I}}}\gamma_t) \\ &+ \delta \left\{ E_{\text{th}} - (1-\lambda)\sum_{t=1}^{N} \lambda^{t-1} \left[\frac{E_t^{\text{I}}}{\vartheta_{\text{U}}} + P_{\text{C}_\text{U}} \tau_t^{\text{I}} + \frac{P_{\max}^{\text{E}} \tau_t^{\text{E}}}{\vartheta_{\text{AP}}} \right. \right. \\ &\left. \left. + P_{\text{C}_\text{AP}} \tau_t^{\text{E}} - \eta G_{\text{A}} H_t P_{\max}^{\text{E}} \tau_t^{\text{E}} \right] \right\} + \sum_{t=1}^{N} \phi_t (T - \tau_t^{\text{I}} - \tau_t^{\text{E}}) \\ &+ \sum_{t=1}^{N} \upsilon_t \left[B_0 + \sum_{i=1}^{t} \eta G_{\text{A}} H_i P_{\max}^{\text{E}} \tau_i^{\text{E}} - \sum_{i=1}^{t} \left(\frac{E_i^{\text{I}}}{\vartheta_{\text{U}}} + P_{\text{C}_\text{U}} \tau_i^{\text{I}} \right) \right] \\ &+ \sum_{t=1}^{N} \psi_t \left\{ B_{\max} - \left[B_0 + \sum_{i=1}^{t} \eta G_{\text{A}} H_i P_{\max}^{\text{E}} \tau_i^{\text{E}} - \sum_{i=1}^{t-1} \left(\frac{E_i^{\text{I}}}{\vartheta_{\text{U}}} + P_{\text{C}_\text{U}} \tau_i^{\text{I}} \right) \right] \right\}, \end{aligned} \tag{32}$$

where $\boldsymbol{\tau}^{\mathbf{E}}, \boldsymbol{\tau}^{\mathbf{I}}, \boldsymbol{E}^{\mathbf{I}}$ are primal variable vectors consisting of τ_t^{E}, τ_t^{I} and E_t^{I}, $\forall t = 1, 2, ..., N$, respectively. δ, $\boldsymbol{\phi}$, $\boldsymbol{\upsilon}$ and $\boldsymbol{\psi}$ are non-negative Lagrange multipliers, in which $\boldsymbol{\phi} = [\phi_1, \phi_2, ..., \phi_N]$, $\boldsymbol{\upsilon} = [\upsilon_1, \upsilon_2, ..., \upsilon_N]$, and $\boldsymbol{\psi} = [\psi_1, \psi_2, ..., \psi_N]$. The complementary slackness condition considered in this paper is given

by

$$\upsilon_t \left[B_0 + \sum_{i=1}^{t} \eta G_{\mathrm{A}} H_i P_{\max}^{\mathrm{E}} \tau_i^{\mathrm{E}} - \sum_{i=1}^{t} \left(\frac{E_i^{\mathrm{I}}}{\vartheta_{\mathrm{U}}} + P_{\mathrm{C_U}} \tau_i^{\mathrm{I}} \right) \right] = 0, \forall t. \tag{33}$$

The other complementary slackness conditions, which provide no further useful information for the design of the sub-optimal online policy, are omitted due to the limited space.

Lemma 3. For given δ, ϕ, υ and ψ, the optimal solution of problem (31) could have following characteristics:

1) The optimal transmit power at the UE in block t is in the structure of

$$P_t^{\mathrm{I}*} = \frac{E_t^{\mathrm{I}}}{\tau_t^{\mathrm{I}}} = \left[\frac{W\vartheta_{\mathrm{U}}}{\ln 2} A_1^{-1} - \frac{1}{\gamma_t} \right]^+, \tag{34}$$

where $A_1 = \delta + \sum_{i=t}^{N} \frac{\upsilon_i}{(1-\lambda)\lambda^{t-1}} - \sum_{i=t+1}^{N} \frac{\psi_i}{(1-\lambda)\lambda^{t-1}}$.

2) If $\gamma_t < \gamma_t^*$, $\tau_t^{\mathrm{I}} = 0$. Otherwise, if $\gamma_t > \gamma_t^*$, $B_0 + \sum_{i=1}^{t} \eta G_{\mathrm{A}} H_i P_{\max}^{\mathrm{E}} \tau_i^{\mathrm{E}} - \sum_{i=1}^{t} \left(\frac{E_i^{\mathrm{I}}}{\vartheta_{\mathrm{U}}} + P_{\mathrm{C_U}} \tau_i^{\mathrm{I}} \right) = 0$. Here γ_t^* is the solution of

$$\begin{aligned} &(1-\lambda)\lambda^{t-1} W \left[\log_2\left(\frac{W\vartheta_{\mathrm{U}}}{A_1' \ln 2} \gamma_t \right) - \frac{1}{\ln 2} \frac{\frac{W\vartheta_{\mathrm{U}}}{A_1' \ln 2} \gamma_t - 1}{\frac{W\vartheta_{\mathrm{U}}}{A_1' \ln 2} \gamma_t} \right] \\ &- \phi_t - P_{\mathrm{C_U}} \left[\delta(1-\lambda)\lambda^{t-1} + \sum_{i=t+1}^{N} \upsilon_i - \sum_{i=t+1}^{N} \psi_i \right] = 0, \end{aligned} \tag{35}$$

where $A_1' = \delta + \sum_{i=t+1}^{N} \frac{\upsilon_i}{(1-\lambda)\lambda^{t-1}} - \sum_{i=t+1}^{N} \frac{\psi_i}{(1-\lambda)\lambda^{t-1}}$.

Proof. Please refer to Appendix A. □

Lemma 3 shows us two potential properties of the optimal offline solution:

1) The optimal transmit power $P_t^{\mathrm{I}*}$ is increasing with γ_t. This property intrigues a stair-case structure design in the sub-optimal online policy, i.e., a higher channel gain comes with a larger transmit power at the UE.

2) If γ_t exceeds a threshold γ_t^*, then the energy stored in the battery should be exhausted, i.e., $B_0 + \sum_{i=1}^{t} \eta G_{\mathrm{A}} H_i P_{\max}^{\mathrm{E}} \tau_i^{\mathrm{E}} - \sum_{i=1}^{t} \left(\frac{E_i^{\mathrm{I}}}{\vartheta_{\mathrm{U}}} + P_{\mathrm{C_U}} \tau_i^{\mathrm{I}} \right) = 0$. Otherwise, no energy is allocated at the UE in block t (i.e., $\tau_t^{\mathrm{I}} = 0 \Rightarrow E_t^{\mathrm{I}} = 0$). As shown in (35), the threshold γ_t^* is closely related to the Lagrange multipliers υ_i and ϕ_i, where $i = t+1, ..., N$. It is intractable to obtain the exact value of γ_t^* when N goes to infinity. Nevertheless, this property inspires us that the energy consumption of the battery at the UE could be managed in an on-off structure. Specifically, in a block, the UE either exhausts the battery, or keeps silent and sends no data to the H-AP.

Remark 1. In problem (31), the time duration and the energy cost for WIT as well as the time duration for WET (i.e., τ_t^{I}, E_t^{I}, and τ_t^{E}, respectively) are optimally designed over time sequence so that the system throughput in N blocks is maximized. Different from works in [32] and [33] where the energy arrivals are predetermined and known at the transmitter, the energy arrival at the UE in block t is unknown and to be determined by τ_t^{E} in problem (31). In other words, the energy arrivals in problem (31) are variables to be designed under the constraints of (31b)-(31f). As shown in *Lemma* 3, this controllability of the energy arrivals leads to a different solution structure compared to those in [32] and [33]. Consequently, the directional water-filling algorithm [32] and the directional glue pouring algorithm [33], both of which require the knowledge of the energy arrivals, are not applicable to our case. Nevertheless, if $E_{\mathrm{th}} \to \infty$ and a sequence of $\tau_t^{\mathrm{E}}, \forall t$, is already given, problem (31) with non-zero $P_{\mathrm{C_u}}$ turns out to be in the same form with that in [33] and is equivalent to that in [32] if zero $P_{\mathrm{C_u}}$ is considered.

TABLE I: The Complexity Performance Comparison

Policies	Computational complexity	Optimality
Optimal (without complexity reduction)	$O(\frac{1}{1-\lambda}\log(\frac{1}{1-\lambda})V_\tau^E V_\tau^I V_P^E V_P^I \lvert\mathcal{S}\rvert^3)$	Global Optimal
Optimal (with complexity reduction)	$O(\frac{1}{1-\lambda}\log(\frac{1}{1-\lambda})V_\tau^E V_\tau^I V_P^I \lvert\mathcal{S}\rvert^3)$	Global Optimal
Quasi-best-effort	$O(\frac{1}{1-\lambda}\log(\frac{1}{1-\lambda})V_\tau^E V_\tau^I \lvert\mathcal{S}\rvert^3)$	Sub-optimal

Through exploiting the two properties from *Lemma* 3, a sub-optimal quasi-best-effort policy[2] is proposed and described in Algorithm 2. To mimic the first property, we design a stair-case structure for the UE's transmit power P^{I}, where higher P^{I} is allocated for the higher channel state. For example, if $P_{\mathrm{th}}^{\mathrm{I}}$ is allocated at system state $\boldsymbol{s} = [h = k, b = l]$, then $P^{\mathrm{I}} \geq P_{\mathrm{th}}^{\mathrm{I}}$ should be allocated at the system state $\boldsymbol{s} = [h > k, b = l]$, where $k = 1, 2, ..., K$ and $l = 0, 1, ..., L$. Particularly, $P_{\mathrm{th}}^{\mathrm{I}}$ is initialized as 0 for $h = 1$ and is updated iteratively during the process of VIA (e.g., lines 7-14 in Algorithm 2). Furthermore, we design an on-off structure for the energy consumption at the UE to mimic the second property. Specifically, for a given system state $\boldsymbol{s} = [h, b]$, the actions at the UE, i.e., τ^{I} and P^{I}, comply with one of the following rules: 1) $\tau^{\mathrm{I}} = 0$ so that no energy is consumed at the UE; 2) $P^{\mathrm{I}}\tau^{\mathrm{I}}/\vartheta_{\mathrm{U}} + P_{\mathrm{C_U}}\tau^{\mathrm{I}} - \min(bQ + \eta G_{\mathrm{A}} \bar{H}(h) P_{\max}^{\mathrm{E}} \tau^{\mathrm{E}}, B_{\max}) = 0$ so that the energy stored at the UE is exhausted. Correspondingly, in Algorithm 2, the action space at the system state $\boldsymbol{s} = [h, b]$ (denoted as $\mathcal{A}''(h, b)$) can be described as an intersection of three action sets, i.e., $\mathcal{A}''(h,b) = \{\{\tau^{\mathrm{E}}, \tau^{\mathrm{I}}, P^{\mathrm{I}}\} | \mathcal{A}'(h,b) \cap \{P^{\mathrm{I}} \geq P_{\mathrm{th}}^{\mathrm{I}}\} \cap \{P^{\mathrm{I}}\tau^{\mathrm{I}}/\vartheta_{\mathrm{U}} + P_{\mathrm{C_U}}\tau^{\mathrm{I}} - \min(bQ + \eta G_{\mathrm{A}} \bar{H}(h) P_{\max}^{\mathrm{E}} \tau^{\mathrm{E}}, B_{\max}) = 0\} \cup \{\tau^I = 0\}\}$. Obviously, compared to the optimal online policy, the action space of the quasi-best-effort policy is smaller and the maximum number of actions in a state shrinks to $V_\tau^{\mathrm{E}} V_\tau^{\mathrm{I}}$. Correspondingly, the computational complexity of the quasi-best-effort policy is $O(\frac{1}{1-\lambda}\log(\frac{1}{1-\lambda})V_\tau^{\mathrm{E}} V_\tau^{\mathrm{I}} |\mathcal{S}|^3)$, which is substantially lower compared to the optimal online policy. The computational complexity of the optimal policy and that of the quasi-best-effort policy are listed in the Table I.

Algorithm 2 The Quasi-best-effort Policy

1: Set $n = 0$, $\beta^- = 0$, β^+, $\beta^0 = \beta^-$.
2: Specify $\varepsilon_\beta > 0$ and $\varepsilon_J > 0$.

[2]It is worth noting that the proposed sub-optimal online policy is designed for the case of single-UE. Extending the results to the case of multiple users is beyond the scope of this paper and it is an interesting topic for future work.

repeat
  Set $\beta = \beta^n$.
  Set $n = n + 1$ and $m = 0$.
  Initialize $J_\beta^0(h, b) = 0$ for each $h \in \mathcal{H}$ and $b \in \mathcal{B}$.
  for each $b \in \mathcal{B}$ **do**
    Set $P_{\text{th}}^{\text{I}} = 0$ and $h = 1$.
    while $h \le |\mathcal{H}|$ **do**
      Compute

$$J_\beta^{m+1}(h,b) = \max_{\boldsymbol{a}\in\mathcal{A}''(h,b)} \left\{ (1-\lambda)\widetilde{r}(h,b,\boldsymbol{a};\beta) + \lambda \sum_{h'\in\mathcal{H},b'\in\mathcal{B}} \mathcal{P}(h',b'|h,b,\boldsymbol{a}) J_\beta^m(h',b') \right\}. \tag{36}$$

$$\begin{aligned}\mu_\beta(h,b) =& \{\tau_\beta^{\text{E}}(h,b), \tau_\beta^{\text{I}}(h,b), P_\beta^{\text{I}}(h,b)\} \\ =& \arg \max_{\boldsymbol{a}\in\mathcal{A}''(h,b)} \left\{ (1-\lambda)\widetilde{r}(h,b,\boldsymbol{a};\beta) + \lambda \sum_{h'\in\mathcal{H},b'\in\mathcal{B}} \mathcal{P}(h',b'|h,b,\boldsymbol{a}) J_\beta^{m+1}(h',b') \right\}.\end{aligned} \tag{37}$$

      Set $P_{\text{th}}^{\text{I}} = P_\beta^{\text{I}}(h,b)$ and Set $h = h + 1$.
      Update $\mathcal{A}''(h,b)$.
    end while
  end for
  If $\|\boldsymbol{J}_\beta^m - \boldsymbol{J}_\beta^{m+1}\| < \frac{\varepsilon_J(1-\lambda)}{2\lambda}$, go to line 16. Otherwise, set $m = m + 1$ and go to line 7.
  For each $h \in \mathcal{H}$ and $b \in \mathcal{B}$, select the policy with determined β by

$$\mu_\beta(h,b) = \arg \max_{\boldsymbol{a}\in\mathcal{A}''(h,b)} \left\{ (1-\lambda)\widetilde{r}(h,b,\boldsymbol{a};\beta) + \lambda \sum_{h'\in\mathcal{H},b'\in\mathcal{B}} \mathcal{P}(h',b'|h,b,\boldsymbol{a}) J_\beta^{m+1}(h',b') \right\}. \tag{38}$$

  Compute the stationary distribution $\Psi(h,b)$ induced by $\boldsymbol{\mu}_\beta(h,b) = \{\mu_\beta(h,b), \forall h \in \mathcal{H}, b \in \mathcal{B}\}$.
  if $\sum_{h\in\mathcal{H},b\in\mathcal{B}} \Psi(h,b) e(h,b,\mu_\beta(h,b)) > E_{th}$ **then**
    $\beta^{n+1} = \frac{\beta^+ + \beta^n}{2}$.
    $\beta^- = \beta^n$.
  else
    $\beta^{n+1} = \frac{\beta^- + \beta^n}{2}$.
    $\beta^+ = \beta^n$.
  end if
until $|\beta^{n+1} - \beta^n| < \varepsilon_\beta$.
Select the policies $\mu_{\beta^-}(h,b)$ and $\mu_{\beta^+}(h,b)$ based on (38) with obtained β^- and β^+, respectively.
Compute the stationary distribution $\Psi_{\beta^-}(h,b)$ and $\Psi_{\beta^+}(h,b)$ induced by $\boldsymbol{\mu}_{\beta^-}(h,b)$ and $\boldsymbol{\mu}_{\beta^+}(h,b)$, respectively.
Compute

$$R_{\beta^-} = \sum_{h\in\mathcal{H},b\in\mathcal{B}} \Psi_{\beta^-}(h,b) r(h,b,\mu_{i,\beta^-}(h,b)), \tag{39}$$

$$R_{\beta^+} = \sum_{h\in\mathcal{H},b\in\mathcal{B}} \Psi_{\beta^+}(h,b) r(h,b,\mu_{i,\beta^+}(h,b)), \tag{40}$$

$$E_{\beta^-} = \sum_{h\in\mathcal{H},b\in\mathcal{B}} \Psi_{\beta^-}(h,b) e(h,b,\mu_{i,\beta^-}(h,b)), \tag{41}$$

$$E_{\beta^+} = \sum_{h\in\mathcal{H},b\in\mathcal{B}} \Psi_{\beta^+}(h,b) e(h,b,\mu_{i,\beta^+}(h,b)). \tag{42}$$

Compute q by solving $E_{\text{th}} = qE_{\beta^-} + (1-q)E_{\beta^+}$.
Obtain the optimal reward $R = qR_{\beta^-} + (1-q)R_{\beta^+}$, and the corresponding optimal policy for each $h \in \mathcal{H}$ and $b \in \mathcal{B}$

$$\mu^*(h,b) = \begin{cases} \mu_{\beta^-}(h,b), & \text{w.p. } q \quad (43) \\ \mu_{\beta^+}(h,b), & \text{w.p. } 1-q \quad (44) \end{cases}$$

V. Simulation Results

In this section, numerical simulations are conducted to evaluate the performance of the optimal online policy and our proposed sub-optimal online policy. For the practicality of RF energy transfer, we consider a Rician fading channel between the H-AP and the UE in the simulations. Correspondingly, the PDF of θ_t is

$$\rho(\theta_t) = \frac{1}{2\varrho^2} e^{\frac{-(\theta_t+\varsigma^2)}{2\varrho^2}} I_0\left(\frac{\sqrt{\theta_t}\varsigma}{\varrho^2}\right), \tag{45}$$

where I_0 is the modified Bessel function of the zero-th order, $2\varrho^2$ and ς^2 are the parameters representing the power of multi-path and line-of-sight, respectively. Furthermore, the level crossing rate $\Lambda(\Theta_{\text{b}})$ is given as [23]

$$\Lambda(\Theta_{\text{b}}) = \sqrt{\frac{2\pi(1+\kappa)\Theta_{\text{b}}}{\bar{\theta}}} f_D e^{-(\kappa+\frac{1+\kappa}{\bar{\theta}}\Theta_{\text{b}})} I_0\left(\sqrt{\frac{\kappa(1+\kappa)\Theta_{\text{b}}}{\bar{\theta}}}\right), \tag{46}$$

where f_D is the maximum Doppler shift[3] of the channel, $\bar{\theta} = 2\varrho^2 + \varsigma^2$ represents the local-mean fading power and $\kappa = \frac{\varsigma^2}{2\varrho^2}$. Besides, practical channel parameters setting in [23] is adopted for simulation, where the number of channel states is selected as $K = 3$, the maximum Doppler shift is set as $f_{\text{D}} = 1.34$ Hz, and the block duration is set as $T = 16$ ms, respectively.

The maximum transmit power of a practical UE is usually around 23 dBm with a step size raging from 0.5 dB to 3 dB [36]. In this case, the average step size is about 4.5 mW when a 0.5 dB step size is taken. Therefore, by jointly considering the practicality and the accuracy of the results, we set the step size of the transmit power at the UE as $\Delta P^{\text{I}} = 1$ mW. On the other hand, extensive simulations (not shown here) have revealed that the accuracy of results is guaranteed when $\varepsilon_J = 10^{-5}$, $\varepsilon_\beta = 10^{-4}$, and $\Delta\tau^{\text{I}} = \Delta\tau^{\text{E}} = 0.05T$. Besides, similar to [17], we focus on the case of small devices and express the battery size as a function of the reference value $B_{\text{ref}} = 10^{-3} \times T$ J. Unless otherwise stated, the maximum battery capacity is set as $B_{\max} = 10B_{\text{ref}}$. The battery quantum is set as $Q = B_{\text{ref}}$ (the effect of battery energy quantization will be discussed later).

[3]In practice, the RF signal from a transmitter experiences a multipath fading and arrives at the receiver via direct and reflecting paths. Although the locations of the H-AP and the UE are fixed, the obstacles reflecting the RF signal in the environment could be moving, and thus produce the Doppler shift in communication system, cf. [34] and [35].

TABLE II: Parameters Setting

$P^{\mathrm{E}}_{\max}$	10 W	α	2.8	$P_{\mathrm{C_{AP}}}$	500 mW
$P_{\mathrm{C_U}}$	5 mW	ϑ_{AP}	0.9	ϑ_{U}	0.9
η	0.95	λ	0.9	G_A	8 dBi
ζ	1	W	2 kHz	N_0	-164 dBm/Hz
ς^2	0.75	ϱ^2	0.125		

Other important parameters used in this section are listed in Table II.

In the following, the benchmark used in our simulations is first introduced.

A. The Benchmark: The Myopic Policy

In our simulations, the myopic policy which focuses on maximizing the throughput only in a single block is considered as the benchmark. It makes decision at the beginning of every block according to the observed channel state, the battery state and the current available energy budget. The myopic policy in block t is obtained through solving

$$\max_{\tau_t^{\mathrm{I}},\tau_t^{\mathrm{E}},P_t^{\mathrm{I}}} \quad \tilde{R} = \frac{\int_{\Theta_k}^{\Theta_{k+1}} \tau_t^{\mathrm{I}} W \log_2\left(1+\frac{P_t^{\mathrm{I}}\theta_t d^{-\alpha}}{\zeta\sigma^2}\right)\rho(\theta_t)\mathrm{d}\theta_{\mathrm{t}}}{\pi_k} \tag{47a}$$

$$\text{s.t.} \quad (1-\lambda)\sum_{i=1}^{t}\lambda^{t-1}E_t^{\mathrm{th}} \le E_{\mathrm{th}}, \tag{47b}$$

$$\frac{P_t^{\mathrm{I}}\tau_t^{\mathrm{I}}}{\vartheta_{\mathrm{U}}} + P_{\mathrm{C_U}}\tau_t^{\mathrm{I}} \le B_t + \eta G_{\mathrm{A}}\bar{H}_t P^{\mathrm{E}}_{\max}\tau_t^{\mathrm{E}}, \tag{47c}$$

$$B_t + \eta G_{\mathrm{A}}\bar{H}_t P^{\mathrm{E}}_{\max}\tau_t^{\mathrm{E}} \le B_{\max}, \tag{47d}$$

$$\tau_t^{\mathrm{E}} + \tau_t^{\mathrm{I}} \le T, \tag{47e}$$

$$P_t^{\mathrm{I}}, \tau_t^{\mathrm{E}}, \tau_t^{\mathrm{I}} \ge 0, \tag{47f}$$

where $E_t^{\mathrm{th}} = \frac{P^{\mathrm{E}}_{\max}}{\vartheta_{\mathrm{AP}}}\tau_t^{\mathrm{E}} + P_{\mathrm{C_{AP}}}\tau_t^{\mathrm{E}} + \frac{P_t^{\mathrm{I}}\tau_t^{\mathrm{I}}}{\vartheta_{\mathrm{U}}} + P_{\mathrm{C_U}}\tau_t^{\mathrm{I}} - \eta G_{\mathrm{A}}\bar{H}_t P^{\mathrm{E}}_{\max}\tau_t^{\mathrm{E}}$ is the energy cost in block t. In other words, starting with the first block, the myopic policy is obtained through calculating τ_t^{I}, τ_t^{E} and P_t^{I} ($t = 1, 2, ...$) sequentially until the total system energy consumption meets the system energy budget, i.e., $(1-\lambda)\sum_{i=1}^{t}\lambda^{t-1}E_t^{\mathrm{th}} = E_{\mathrm{th}}$. Let $E_t^{\mathrm{I}} = P_t^{\mathrm{I}}\tau_t^{\mathrm{I}}$, then the objective function of problem (47) can be written as

$$\tilde{R} = \frac{\int_{\Theta_k}^{\Theta_{k+1}} \tau_t^{\mathrm{I}} W \log_2\left(1+\frac{E_t^{\mathrm{I}}\theta_t d^{-\alpha}}{\tau_t^{\mathrm{I}}\zeta\sigma^2}\right)\rho(\theta_t)\mathrm{d}\theta_{\mathrm{t}}}{\pi_k}. \tag{48}$$

As shown in [8] and [33], for a given θ_t, $\tilde{r} = \tau_t^{\mathrm{I}} W \log_2\left(1+\frac{E_t^{\mathrm{I}}\theta_t d^{-\alpha}}{\tau_t^{\mathrm{I}}\zeta\sigma^2}\right)\rho(\theta_t)$ is a concave function of E_t^{I} and τ_t^{I}. Correspondingly, $\tilde{R}$, which can be treated as the summation of multiple $\tilde{r}$ with different positive values of θ_t, is also a concave function of E_t^{I} and τ_t^{I}. Therefore, problem (47) is a convex optimization problem and can be efficiently solved through standard convex optimization methods. Due to the randomization of the channel state transition, the Monte-Carlo method is applied, where the results of the myopic policy are obtained through averaging 200 realizations and are plotted with 95% confidence intervals.

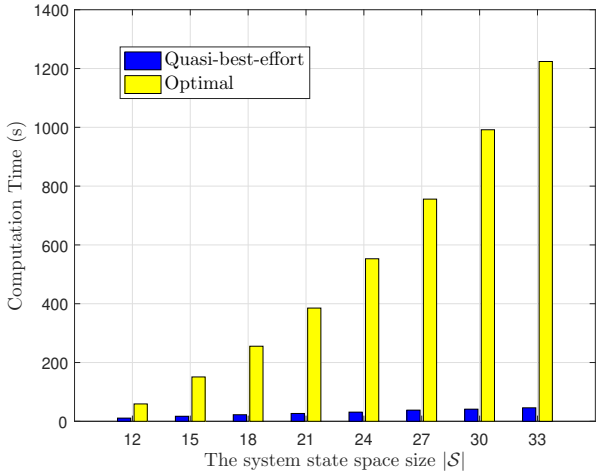


Fig. 2: The comparison on the computation time of the optimal policy and the quasi-best-effort policy.

B. The Performance of Computational Complexity

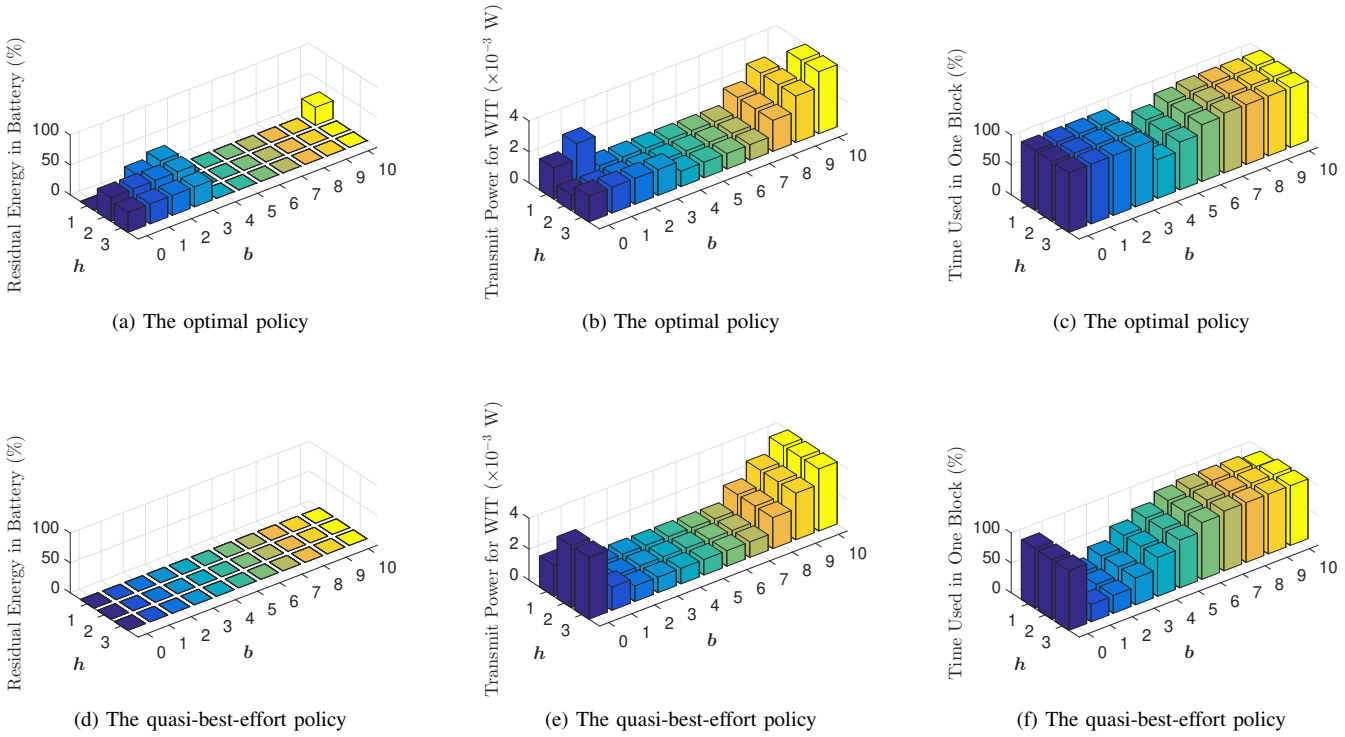
In Sec. III and Sec. IV, the computational complexity of the optimal policy and the quasi-best-effort policy are theoretically analyzed. In this subsection, we compare the computation time of these two policies through numerical simulation. Here we set $|\mathcal{H}| = 3$ and consider various $|\mathcal{B}|$ from 4 to 11. Correspondingly, the system state space size $|\mathcal{S}| = |\mathcal{H}||\mathcal{B}|$ is ranging from 12 to 33. The simulation is conducted on the computer with 16 GB RAM and Inter Core i7-7770 CPU working at 3.6 GHz. The result is recorded in Fig. 2. As shown in the figure, the proposed quasi-best-effort policy greatly outperforms the optimal policy in terms of the computation complexity. Moreover, the larger the $|\mathcal{S}|$, the more obvious the advantage is.

C. The Performance of Long-term Throughput

In this subsection, we investigate the long-term throughput performance of the three policies, i.e, the optimal policy, the quasi-best-effort policy and the myopic policy. We use "Optimal", "Quasi-best-effort" and "Myopic" to represent these three policies in the simulation results, respectively.

The main differences between the optimal policy and the quasi-best-effort policy are first investigated in Fig. 3. In these simulations, the maximum energy budget is set as $E_{\mathrm{th}} = 400B_{\mathrm{ref}}$. The communication distance is set as $d = 6$ m. As stated in Sec. III, the optimal policy of the CMDP is a randomized mixture of two pure policies $\boldsymbol{\mu}_{\beta^-}$ and $\boldsymbol{\mu}_{\beta^+}$. Thus these results are obtained through making a expectation over $\boldsymbol{\mu}_{\beta^-}$ and $\boldsymbol{\mu}_{\beta^+}$.

It can be observed in Fig. 3a and Fig. 3d that, the first difference between the optimal policy and the quasi-best-effort policy is the management of the energy stored in the UE's battery. In particular, for the quasi-best-effort policy, the UE's battery is exhausted in each system state in this case. However, as shown in Fig. 3a, the optimal policy not always uses up the battery's energy at the UE. Remarkably, the behavior of the optimal policy is quite different from that in



(a) The optimal policy (b) The optimal policy (c) The optimal policy

(d) The quasi-best-effort policy (e) The quasi-best-effort policy (f) The quasi-best-effort policy

Fig. 3: The differences between the optimal policy and the quasi-best-effort policy.

the traditional EH (e.g., harvesting energy from solar or wind power) communication systems, where the energy is more likely to be stored in the battery when the channel suffers a deep fading and depleted when channel is of good quality. For the optimal policy in a WPCN, there could be more energy stored in the battery at a high channel state (i.e., $h = 3$ with $b = 1$) rather than at a low channel state (i.e., $h = 1$ with $b = 1$). This is because that the energy delivered from the H-AP suffers dissipation during the WET. It is more reasonable to transfer energy to the UE at a good channel state for a WPCN.

Secondly, the transmit power allocated to the UE of these two policies is illustrated in Fig. 3b and Fig. 3e, respectively. For the quasi-best-effort policy which is structurally designed, the transmit power of the UE is increasing with the channel state. In contrast, the transmit power of the UE of the optimal policy is not monotonic with the channel state (i.e., $b = 1$).

Lastly, we compare the time occupation in a block by these two policies in Fig. 3c and Fig. 3f. As shown in these two figures, due to the maximum allowed energy budget, the time duration of each block is not always fully occupied, which is in contrast to the result in [17] where the time in a block is always totally used, i.e., $\tau^E + \tau^I = T$.

In order to investigate the long-term throughput performance under different scenarios and show the validity of our proposed policy, numerical simulations are conducted below with various energy budget E_{th}, discount factor λ, communication distance d, maximum battery capacity $B_{\max}$, and energy conversion efficiency η, respectively.

The performances of the optimal policy, the quasi-best-effort policy and the myopic policy under different maximum allowed energy cost E_{th} are first examined in Fig. 4. In the simulation, we set $d = 10$ m. As shown in the figure, the performance gap between the quasi-best-effort policy and the optimal policy is small. Compared to the myopic policy, the long-term throughput is considerably improved by adopting the proposed quasi-best-effort policy. Moreover, it is shown to be increased with E_{th}, since a larger E_{th} means a higher system energy budget. However, as E_{th} becomes larger, the long-term throughput tends to be saturated. This is due to the limitation of the battery size and the maximum transmit power constraint of the H-AP.

Besides, we also consider different values of Q to illustrate the impact of quantization of the battery states in Fig. 4. In this simulation, we set $Q = 5B_{\text{ref}}$, B_{ref}, and $B_{\text{ref}}/2$, respectively. Correspondingly, for $B_{\max} = 10B_{\text{ref}}$, the space size of the battery states are $|\mathcal{B}| = 3$, 11, and 21, respectively. As shown in Fig. 4, the performance of the optimal policy is improved when $|\mathcal{B}|$ varies from 3 to 11, but is almost invariant when $|\mathcal{B}|$ increases from 11 to 21, which verifies the validity of using $Q = B_{\text{ref}}$. Therefore, to ensure the accuracy of the results, we have set $Q = B_{\text{ref}}$ in all the simulations.

Fig. 5 depicts the long-term throughput performance of the three policies under different value of λ. Here, we set $E_{\text{th}} = 200B_{\text{ref}}$ and $d = 10$ m. As shown in the figure, the quasi-best-effort policy shows significant improvement over the myopic policy and small performance degradation from the optimal one. Besides, as described in Sec. II, the discount factor λ

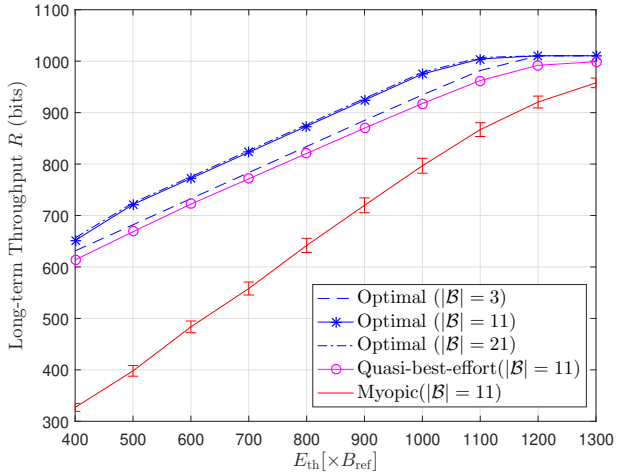


Fig. 4: The long-term throughput versus the maximum available energy budget E_{th}.

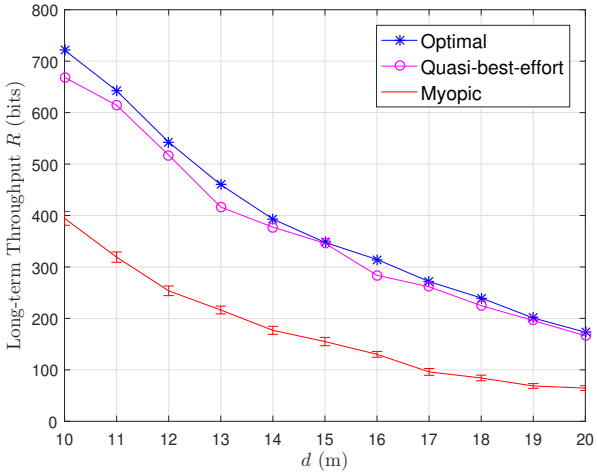


Fig. 6: The long-term throughput versus the communication distance d.

represents the probability of the UE to survive the physical operation failure. Although a larger λ yields a longer average system operation time, the long-term throughput is shown to be almost steady with the increase of λ in Fig. 5. This is quite different from the result in [19], where the sum-throughput is demonstrated to be increased with λ.

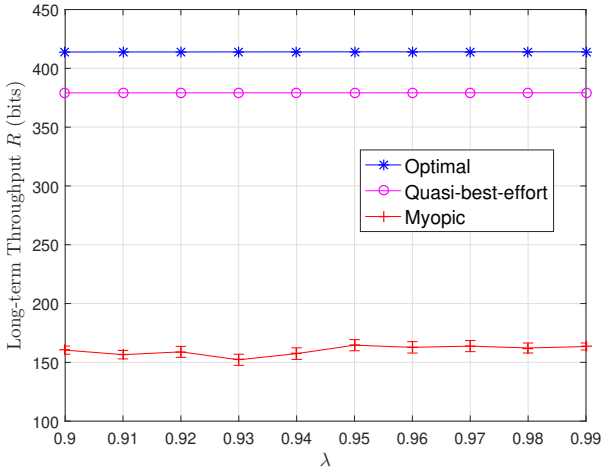


Fig. 5: The long-term throughput versus the discount factor λ.

In Fig. 6, the long-term throughput R is depicted as a function of the communication distances d with $E_{\text{th}} = 500B_{\text{ref}}$. As shown in the figure, R rapidly decreases when d becomes larger. This is for the reason that a longer communication distance results in a more severe signal attenuation due to path-loss during WET and WIT. Similar to the previous results, for different values of d, the quasi-best-effort policy offers a substantial performance gain over the myopic policy and approaches a close-to-optimal performance compared to the optimal policy.

Considering different maximum battery capacity $B_{\max}$, the long-term throughput performances of the three policies are demonstrated in Fig. 7. In this simulation, we set $E_{\text{th}} = 500B_{\text{ref}}$ and $d = 10$ m. As shown in the figure, for the optimal policy and the quasi-best-effort policy, a higher $B_{\max}$ leads to a higher system throughput. Since UE with larger battery capacity can store more energy, it is more likely to fully exploit the channel fluctuation opportunistically. As $B_{\max}$ grows, the performance of the system saturates because the energy budget of system E_{th} is limited. However, the myopic policy shows a different trend. With the growth of $B_{\max}$, the corresponding long-term throughput first increases and then declines. This is because that the myopic policy operates sequentially from block to block and exhausts the battery's energy as much as possible to maximize the current system throughput. In this case, a larger battery capacity $B_{\max}$ brings a higher throughput but leads to a higher energy consumption in each block, and thus a shorter system operation time due to the energy budget constraint. Therefore, there is a trade-off on $B_{\max}$ for the myopic policy. Nevertheless, compared to the myopic policy, considerable improvement can be observed when the quasi-best-effort policy is adopted.

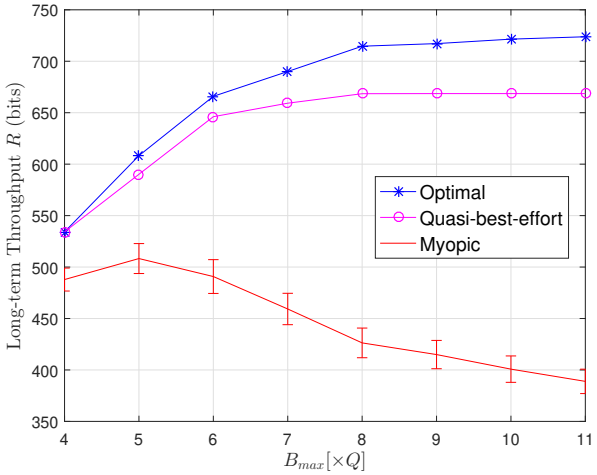


Fig. 7: The long-term throughput versus the maximum battery capacity $B_{\max}$.

Lastly, we investigate the relationship between the energy

conversion efficiency η and the long-term throughput with various circuit power settings. The result is plotted in Fig. 8. Here we set $E_{\text{th}} = 500B_{\text{ref}}$ and $d = 8$ m. As shown in the figure, the long-term throughput grows with increasing η. This is because that a higher energy conversion efficiency comes with more available energy at the UE, thus yielding a better throughput performance. On the other hand, the long-term throughput is shown to be more sensitive to the circuit power at the UE rather than that at the H-AP. For example, with the optimal policy, the long-term throughput achieves a performance gain of 1.2 dB at $\eta = 0.75$ when P_{C_U} decreases 3 dB (from 10 mW to 5 mW) and is almost unchanged when P_{C_AP} drops from 0.5 W to 0.25 W. Nevertheless, for all these three cases of circuit power settings, the proposed quasi-best-effort policy significantly outperforms the myopic policy in terms of the long-term throughput and shows a narrow gap to the optimal policy.

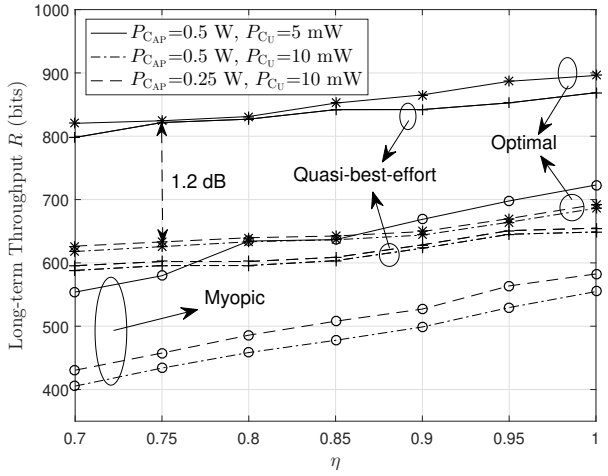


Fig. 8: The long-term throughput versus the energy conversion efficiency η with different circuit power P_{C_AP} and P_{C_U}.

VI. Conclusion

In this paper, we studied the problem of designing online policies in an energy-constrained WPCN to manage the transmit power and time durations for both WET and WIT over fading channels. Aiming at maximizing the system long-term throughput with a limited total energy budget, we first modeled the problem as a CMDP problem and solved it through using the Lagrangian approach. After that, we investigated an offline optimization problem, where the optimal data transmit power was shown to be increasing with the channel gain, and the UE always depletes its battery energy whenever it decides to transmit data in a block. Through mimicking these properties, we proposed a sub-optimal online policy named the quasi-best-effort policy. The simulation results showed that the computation time is considerably reduced when the quasi-best-effort policy is adopted. At the same time, in terms of the long-term throughput performance, the quasi-best-effort policy reaches similar grades to the optimal online policy.

Appendix A
Proof of Lemma 3

By taking the partial derivative of $\mathcal{L}_{\text{QBE}}$ with respect to E_t^{I} and τ_t^{I}, respectively, we obtain

$$\frac{\partial \mathcal{L}_{\text{QBE}}}{\partial E_t^{\text{I}}} = (1-\lambda)\lambda^{t-1}\left(\frac{W}{\ln 2}\frac{\gamma_t}{1+\frac{E_t^{\text{I}}\gamma_t}{\tau_t^{\text{I}}}} - \frac{\delta}{\vartheta_{\text{U}}}\right) - \sum_{i=t}^{N}\frac{\upsilon_i}{\vartheta_{\text{U}}} + \sum_{i=t+1}^{N}\frac{\psi_i}{\vartheta_{\text{U}}}, \tag{49}$$

$$\frac{\partial \mathcal{L}_{\text{QBE}}}{\partial \tau_t^{\text{I}}} = (1-\lambda)\lambda^{t-1}W\left[\log_2(1+\frac{E_t^{\text{I}}}{\tau_t^{\text{I}}}\gamma_t) - \frac{1}{\ln 2}\frac{\frac{E_t^{\text{I}}\gamma_t}{\tau_t^{\text{I}}}}{1+\frac{E_t^{\text{I}}\gamma_t}{\tau_t^{\text{I}}}}\right] - \phi_t - P_{\text{C}_\text{U}}\left[\delta(1-\lambda)\lambda^{t-1} + \sum_{i=t}^{N}\upsilon_i - \sum_{i=t+1}^{N}\psi_i\right]. \tag{50}$$

Let $\frac{\partial \mathcal{L}_{\text{QBE}}}{\partial E_t^{\text{I}}} = 0$, we obtain

$$P_t^{\text{I}*} = \frac{E_t^{\text{I}}}{\tau_t^{\text{I}}} = \left[\frac{W\vartheta_U}{\ln 2}A_1^{-1} - \frac{1}{\gamma_t}\right]^+, \tag{51}$$

where $A_1 = \delta + \sum_{i=t}^{N}\frac{\upsilon_i}{(1-\lambda)\lambda^{t-1}} - \sum_{i=t+1}^{N}\frac{\psi_i}{(1-\lambda)\lambda^{t-1}}$. By submitting (51) into (50), then

$$\frac{\partial \mathcal{L}_{\text{QBE}}}{\partial \tau_t^{\text{I}}} = (1-\lambda)\lambda^{t-1}W\left[\log_2(1+P_t^{\text{I}*}\gamma_t) - \frac{1}{\ln 2}\frac{P_t^{\text{I}*}\gamma_t}{1+P_t^{\text{I}*}\gamma_t}\right] - \phi_t - P_{\text{C}_\text{U}}\left[\delta(1-\lambda)\lambda^{t-1} + \sum_{i=t}^{N}\upsilon_i - \sum_{i=t+1}^{N}\psi_i\right], \tag{52}$$

which shows that $\mathcal{L}_{\text{QBE}}$ is a linear function of τ_t^{I}. Thus we have that

$$\tau_t^{\text{I}}\begin{cases} \in [0,T], & \text{for } \frac{\partial \mathcal{L}_{\text{QBE}}}{\partial \tau_t^{\text{I}}} = 0, \quad (53)\\ = 0, & \text{for } \frac{\partial \mathcal{L}_{\text{QBE}}}{\partial \tau_t^{\text{I}}} < 0. \quad (54)\end{cases}$$

Moreover, (52) can be written as a function of υ_t and γ_t, i.e., $f(\upsilon_t, \gamma_t) = \frac{\partial \mathcal{L}_{\text{QBE}}}{\partial \tau_t^{\text{I}}}$.

Lemma 4. $f(\upsilon_t, \gamma_t)$ is monotonically increasing in γ_t and monotonically decreasing in υ_t under the condition that $A_2 = \frac{W\vartheta_U}{A_1\ln 2} - \frac{1}{\gamma_t} > 0$.

Proof. Taking the partial derivative of $f(\upsilon_t, \gamma_t)$ with respect to γ_t, we have

$$\frac{\partial f}{\partial \gamma_t} = \frac{W(1-\lambda)\lambda^{t-1}}{(1+P_t^{\text{I}*}\gamma_t)\ln 2}\left[1 - \frac{1}{1+P_t^{\text{I}*}\gamma_t}\right]\frac{\partial P_t^{\text{I}*}\gamma_t}{\partial \gamma_t}, \tag{55}$$

where $\frac{\partial P_t^{\text{I}*}\gamma_t}{\partial \gamma_t} = P_t^{\text{I}*} + \frac{\partial P_t^{\text{I}*}}{\partial \gamma_t}$. With the condition of $A_2 > 0$, $\frac{\partial P_t^{\text{I}*}}{\partial \gamma_t} = \frac{1}{\gamma_t^2} > 0$. Thus $\frac{\partial P_t^{\text{I}*}\gamma_t}{\partial \gamma_t} > 0 \Rightarrow \frac{\partial f}{\partial \gamma_t} > 0$, which means $f(\upsilon_t, \gamma_t)$ is monotonically increasing in γ_t.

Similarly, by taking the partial derivative of $f(\upsilon_t, \gamma_t)$ with respect to υ_t, we have that

$$\frac{\partial f}{\partial \upsilon_t} = \frac{W(1-\lambda)\lambda^{t-1}\gamma_t}{(1+P_t^{\text{I}*}\gamma_t)\ln 2}\left[1 - \frac{1}{1+P_t^{\text{I}*}\gamma_t}\right]\frac{\partial P_t^{\text{I}*}}{\partial \upsilon_t} - P_{\text{C}_\text{U}}, \tag{56}$$

where $\frac{\partial P_t^{\mathrm{I}*}}{\partial \upsilon_t} = -\frac{W\vartheta_{\mathrm{U}}}{A_1^2(1-\lambda)\lambda^{t-1}\ln 2} < 0$. Thus we have that $\frac{\partial f}{\partial \upsilon_t} < 0$ and $f(\upsilon_t, \gamma_t)$ is monotonically decreasing in υ_t. □

From *Lemma* 4, we know that, the maximum value of $f(\upsilon_t, \gamma_t)$ in terms of υ_t is achieved at $\upsilon_t = 0$, i.e., $f(0, \gamma_t)$, where

$$f(0,\gamma_t) = (1-\lambda)\lambda^{t-1} W \left[\log_2(1+\gamma_t P^{\mathrm{I}*}_{t,\upsilon_t=0}) - \frac{1}{\ln 2}\frac{\gamma_t P^{\mathrm{I}*}_{t,\upsilon_t=0}}{1+\gamma_t P^{\mathrm{I}*}_{t,\upsilon_t=0}}\right] - \phi_t - P_{C_U}\left[\delta(1-\lambda)\lambda^{t-1} + \sum_{i=t+1}^{N}\upsilon_i - \sum_{i=t+1}^{N}\psi_i\right]. \tag{57}$$

Let $A_1' = \delta + \sum_{i=t+1}^{N}\frac{\upsilon_i}{(1-\lambda)\lambda^{t-1}} - \sum_{i=t+1}^{N}\frac{\psi_i}{(1-\lambda)\lambda^{t-1}}$. With the condition of $A_2 = \frac{W\vartheta_U}{A_1\ln 2} - \frac{1}{\gamma_t} > 0$, from (51), we know that $P_t^{\mathrm{I}*} = \frac{W\vartheta_{\mathrm{U}}}{A_1\ln 2} - \frac{1}{\gamma_t}$. Then $P^{\mathrm{I}*}_{t,\upsilon_t=0} = \frac{W\vartheta_{\mathrm{U}}}{A_1'\ln 2} - \frac{1}{\gamma_t}$ and $f(0,\gamma_t)$ can be rewritten as

$$f(0,\gamma_t) = (1-\lambda)\lambda^{t-1} W \left[\log_2(\frac{W\vartheta_{\mathrm{U}}}{A_1'\ln 2}\gamma_t) - \frac{1}{\ln 2}\frac{\frac{W\vartheta_{\mathrm{U}}}{A_1'\ln 2}\gamma_t - 1}{\frac{W\vartheta_{\mathrm{U}}}{A_1'\ln 2}\gamma_t}\right] - \phi_t - P_{\mathrm{C_U}}\left[\delta(1-\lambda)\lambda^{t-1} + \sum_{i=t+1}^{N}\upsilon_i - \sum_{i=t+1}^{N}\psi_i\right]. \tag{58}$$

Since the function $\log_2(\cdot)$ is defined for the positive real numbers, $f(0,\gamma_t)$ is feasible only if $\frac{W\vartheta_{\mathrm{U}}}{A_1'\ln 2} > 0$, which implies that $A_1' > 0$. Thus we have that $\delta(1-\lambda)\lambda^{t-1} + \sum_{i=t+1}^{N}\upsilon_i - \sum_{i=t+1}^{N}\psi_i > 0$. Then, when $\gamma_t = \frac{A_1'\ln 2}{W\vartheta_{\mathrm{U}}}$, $f(0,\gamma_t) = -\phi_t - P_{\mathrm{C_U}}\left[\delta(1-\lambda)\lambda^{t-1} + \sum_{i=t+1}^{N}\upsilon_i - \sum_{i=t+1}^{N}\psi_i\right] < 0$ and when $\gamma_t \to +\infty$, $f(0,\gamma_t) \to +\infty$. Since $f(\upsilon_t,\gamma_t)$ is a monotonically increasing function of γ_t, there must be a γ_t^* such that $f(0,\gamma_t^*) = 0$. Now we analyse the following two cases through using the monotonically decreasing property of $f(\upsilon_t,\gamma_t)$ with respected to υ_t:

Case I: $\gamma_t < \gamma_t^*$

In this case, it follows that $f(\upsilon_t,\gamma_t) \le f(0,\gamma_t) < f(0,\gamma_t^*) = 0$. According to (54), we have that zero data transmission time is allocated in this case, i.e., $\tau_t^{\mathrm{I}} = 0$.

Case II: $\gamma_t > \gamma_t^*$

When $\gamma_t > \gamma_t^*$, there always exists a $\upsilon_t > 0$ such that $f(\upsilon_t,\gamma_t) = 0$, since $f(\upsilon_t,\gamma_t)$ is decreasing in υ_t. Then according to (53), $\tau_t^{\mathrm{I}} \in [0,T]$ in this case. However, there can also be a $\upsilon_t > 0$ such that $f(\upsilon_t,\gamma_t) < 0$. Then from (54), we know that $\tau_t^{\mathrm{I}} = 0$ and it may leads to $B_0 + \sum_{i=1}^{t}\eta G_{\mathrm{A}} H_i P_{\max}^{\mathrm{E}}\tau_i^{\mathrm{E}} - \sum_{i=1}^{t}\left(\frac{E_i^I}{\vartheta_{\mathrm{U}}} + P_{\mathrm{C_U}}\tau_i^{\mathrm{I}}\right) > 0$, which contradicts the complementary slackness condition (33). Thus this is not the optimal solution. Nevertheless, for $\gamma_t > \gamma_t^*$, $\upsilon_t > 0$ should always come with that $B_0 + \sum_{i=1}^{t}\eta G_{\mathrm{A}} H_i P_{\max}^{\mathrm{E}}\tau_i^{\mathrm{E}} - \sum_{i=1}^{t}\left(\frac{E_i^{\mathrm{I}}}{\vartheta_{\mathrm{U}}} + P_{\mathrm{C_U}}\tau_i^{\mathrm{I}}\right) = 0$, which implies the depletion of the battery at the UE.

REFERENCES

[1] X. Li, X. Zhou, C. Sun, and D. W. K. Ng, "Optimal online transmission policy for energy-constrained wireless powered communication networks," submitted to *Proc. ICC* 2019.

[2] D. Mishra, S. De, S. Jana, S. Basagni, K. Chowdhury, and W. Heinzelman, "Smart RF energy harvesting communications: challenges and opportunities," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 70–78, Apr. 2015.

[3] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless networks with RF energy harvesting: A contemporary survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 757–789, 2015.

[4] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2017.

[5] H. Chen, Y. Li, J. L. Rebelatto, B. F. Ucha-Filho, and B. Vucetic, "Harvest-then-cooperate: Wireless-powered cooperative communications," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1700–1711, Apr. 2015.

[6] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 418–428, Jan. 2014.

[7] C. Zhong, X. Chen, Z. Zhang, and G. K. Karagiannidis, "Wireless-powered communications: Performance analysis and optimization," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 5178–5190, Dec. 2015.

[8] Q. Wu, M. Tao, D. W. K. Ng, W. Chen, and R. Schober, "Energy-efficient resource allocation for wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2312–2327, Mar. 2016.

[9] M. A. Abd-Elmagid, T. ElBatt, and K. G. Seddik, "Optimization of energy-constrained wireless powered communication networks with heterogeneous nodes," *Wireless Networks*, Sep. 2017. [Online]. Available: https://doi.org/10.1007/s11276-017-1587-x.

[10] H. Ju and R. Zhang, "User cooperation in wireless powered communication networks," in *Proc. IEEE Global Communications Conf*, Dec. 2014, pp. 1430–1435.

[11] ——, "Optimal resource allocation in full-duplex wireless-powered communication network," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3528–3540, Oct. 2014.

[12] K. W. Choi and D. I. Kim, "Stochastic optimal control for wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 686–698, Jan. 2016.

[13] Y. L. Che, L. Duan, and R. Zhang, "Dynamic base station operation in large-scale green cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3127–3141, Dec. 2016.

[14] ——, "Spatial throughput maximization of wireless powered communication networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 8, pp. 1534–1548, Aug. 2015.

[15] R. Morsi, D. S. Michalopoulos, and R. Schober, "Performance analysis of near-optimal energy buffer aided wireless powered communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 863–881, Feb. 2018.

[16] X. Zhou, C. K. Ho, and R. Zhang, "Wireless power meets energy harvesting: A joint energy allocation approach in OFDM-based system," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3481–3491, May 2016.

[17] A. Biason and M. Zorzi, "Battery-powered devices in WPCNs," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 216–229, Jan. 2017.

[18] M. A. Abd-Elmagid, A. Biason, T. ElBatt, K. G. Seddik, and M. Zorzi, "On optimal policies in full-duplex wireless powered communication networks," in *Proc. 14th Int. Symp. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2016, pp. 1–7.

[19] S. Mao, M. H. Cheung, and V. W. S. Wong, "Joint energy allocation for sensing and transmission in rechargeable wireless sensor networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2862–2875, Jul. 2014.

[20] S. S. Kalamkar, J. P. Jeyaraj, A. Banerjee, and K. Rajawat, "Resource allocation and fairness in wireless powered cooperative cognitive radio networks," *IEEE Trans. Commun.*, vol. 64, no. 8, pp. 3246–3261, Aug. 2016.

[21] H. S. Wang and N. Moayeri, "Finite-state Markov channel-a useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[22] Y. Kim, "Starvation analysis of CDF-based scheduling over Nakagami-*m* Markov fading channels," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 372–375, Feb. 2018.

[23] F. Babich and G. Lombardi, "A Markov model for the mobile propagation channel," *IEEE Trans. Veh. Technol.*, vol. 49, no. 1, pp. 63–73, Jan. 2000.

[24] P. Sadeghi, R. A. Kennedy, P. B. Rapajic, and R. Shams, "Finite-state Markov modeling of fading channels - a survey of principles and applications," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 57–80, Sep. 2008.

[25] B. Li, W. Guo, Y. Liang, C. An, and C. Zhao, "Asynchronous device detection for cognitive device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2443–2456, Apr. 2018.
[26] R. Zhang, Z. Zhong, Y. Zhang, S. Lu, and L. Cai, "Measurement and analytical study of the correlation properties of subchannel fading for noncontiguous carrier aggregation," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4165–4177, Nov. 2014.
[27] M. Sun, X. Wang, C. Zhao, B. Li, Y. Liang, G. Goussetis, and S. Salous, "Adaptive sensing schedule for dynamic spectrum sharing in time-varying channel," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 5520–5524, Jun. 2018.
[28] E. Altman, *Constrained Markov decision processes*. Chapman & Hall/CRC, 1998.
[29] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2005.
[30] F. J. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint," *Journal of Mathematical Analysis and Applications*, vol. 112, pp. 236–252, Nov. 1985.
[31] M. L. Littman, T. L. Dean, and L. P. Kaelbling, "On the complexity of solving Markov decision problems," in *Proc. the Eleventh Conf. Uncertainty in artificial intelligence - UAI '95*, Aug. 1995, pp. 394–402.
[32] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1732–1743, Sep. 2011.
[33] O. Orhan, D. Gndz, and E. Erkip, "Throughput maximization for an energy harvesting communication system with processing cost," in *Proc. IEEE Information Theory Workshop*, Sep. 2012, pp. 84–88.
[34] S. J. Howard and K. Pahlavan, "Doppler spread measurements of indoor radio channel," *Electron. Lett.*, vol. 26, no. 2, pp. 107–109, Jan. 1990.
[35] S. Thoen, L. V. der Perre, and M. Engels, "Modeling the channel time-variance for fixed wireless communications," *IEEE Commun. Lett.*, vol. 6, no. 8, pp. 331–333, Aug. 2002.
[36] *User Equipment (UE) radio transmission and reception (TDD)*, 3GPP TS 25.102 V15.0.0 Std.