A Cross-Layer Perspective on Energy Harvesting Aided Green Communications over Fading Channels

Tian Zhang, Wei Chen, *Senior Member, IEEE,* Zhu Han, *Senior Member, IEEE,* and Zhigang Cao, *Senior Member, IEEE*

Abstract

In this paper, we consider the power allocation of the physical layer and the buffer delay of the upper application layer in energy harvesting green networks. The total power required for reliable transmission includes the transmission power and the circuit power. The harvested power (which is stored in a battery) and the grid power constitute the power resource. The uncertainty of data generated from the upper layer, the intermittence of the harvested energy, and the variation of the fading channel are taken into account and described as independent Markov processes. In each transmission, the transmitter decides the transmission rate as well as the allocated power from the battery, and the rest of the required power will be supplied by the power grid. The objective is to find an allocation sequence of transmission rate and battery power to minimize the long-term average buffer delay under the average grid power constraint. A stochastic optimization problem is formulated accordingly to find such transmission rate and battery power sequence. Furthermore, the optimization problem is reformulated as a constrained Markov decision process (MDP) problem whose policy is a two-dimensional vector with the transmission rate and the power allocation of the battery as its elements. We prove that the optimal policy of the constrained MDP can be obtained by solving the unconstrained MDP. Then we focus on the analysis of the unconstrained average-cost MDP. The structural properties of the average optimal policy are derived. Moreover, we discuss the relations between elements of the two-dimensional policy. Next, based on the theoretical analysis, the algorithm to find the constrained optimal policy is presented for the finite state space scenario. In addition, heuristic policies (two deterministic policies and a mixed policy) with low-complexity are given for the general state space. Finally, simulations are performed under these policies to demonstrate the effectiveness.

Index Terms

Green communications, energy harvesting, cross-layer design, power allocation, Markov decision process.

I. INTRODUCTION

Rapid wireless communication industry development has led to a dramatic increase of energy consumption in wireless networks, and such an increasing energy consumption produces a series of energetic and environmental problems. Recently, green communications, which aims at enhancing energy efficiency and carbon emission reduction, have received considerable attention [2]-[6]. In the energy-efficient design for wireless communications, the total energy consumption includes not only the transmission energy but also the circuit energy consumption [7].

As a preferred choice supporting green communications, energy harvesting techniques such as photovoltaic solar cells become popular for the ability to prolong the lifetime of the battery and the lifetime of wireless networks thereby. There have been a lot of researches in wireless networks with energy harvesting nodes. In [8], an optimal energy management policy for a solar-powered sensor node was proposed. The policy uses a sleep and wakeup strategy for energy conservation. In [9], throughput optimal and mean delay optimal energy management policies were studied for a single energy harvesting sensor node. The Shannon capacity of an energy harvesting sensor node transmitting over an AWGN channel was obtained in [10]. In [11], the optimal binary transmission policies were studied under i.i.d. Bernoulli energy arrivals. In [12], the long-term average communication reliability optimization problem was studied for the system of energy-harvesting active networked tags (EnHANTs). In [13] and [14], throughput-maximal schemes of energy allocation for wireless communications with energy harvesting constraints are studied.

Resource allocation is a fundamental problem in wireless communications [15]. Generally, resource consumption reduction and quality of service (QoS) improvement are two conflicting objectives in a resource allocation problem. There has been some interests in analyzing the power allocation and delay performance from the cross-layer perspective. In [16] and [17], the tradeoff between the average required power for reliable transmission at the physical layer and the mean delay at the network layer was studied in fading channels. The adaptive control policies utilize information on both queue state and channel state, and some structural results for the optimal policy were derived. In [18], the authors derived the improved results upon these obtained in

[17]. They considered the optimization problem aiming to minimize the delay in the transmitter buffer under an average transmitter power constraint. The existence of stationary average optimal policy was proved and some structural results were obtained. In [19], the fading channel was simplified to a static channel, and the explicit optimal control policy was characterized.

In [17]-[19], only the transmission power is considered. However, as shown in [2], the transmission strategy changes when taking the circuit power into account. Then a natural problem is *what about the power and delay when considering both transmission power and circuit power*. Meanwhile, as energy allocation of the battery plays a central role in the transmission strategy of energy harvesting nodes, *how the energy allocation strategy of the battery will affect the power and delay?*

In this paper, we consider the power allocation in the physical layer and the delay performance in the upper application layer in green wireless networks with energy harvesting nodes. The data are generated in the application layer, and placed in a buffer at the transmitter. The transmitter periodically removes some data from the buffer, and transmits the data to the receiver. The required power for reliable transmission takes both transmission power and circuit power into account, and the power resource makes up of the harvested power and grid power. The harvested energy arrives randomly, and there is a constraint on the average grid power. The objective is to minimize the average delay in the buffer with a constrained average grid power and random battery energy. Since the required power for each transmission can be supplied from both the battery and the grid, the policy is two-dimensional, i.e., the rate as well as the allocation of the battery energy (the grid power allocation is then the total required power minus the allocated battery power), in the formulated optimization problem.

Specifically, the main contributions of the paper can be summarized as follows.

- We consider the delay-optimal power allocation in the framework of green communications over fading channels, where the power comes from both power grid and harvesting devices. The data arrival process, the harvested energy arrival process, and the channel process are Markovian. A stochastic optimization problem is formulated to find a transmission rate and battery power allocation sequence to minimize the long-run average buffer delay under the

constraint on the average grid power.

- We reformulate the optimization problem as a constrained Markov decision process (MDP) problem, in which the state and action are defined. The state includes the queue state, the battery state (i.e., the stored energy in the battery), the channel state, the data arrival, and the harvested energy arrival. The action consists of the transmission rate and the power allocation from the battery. Using the Lagrangian methodology, the constrained MDP can be relaxed to an unconstrained problem (UP), which is an average cost MDP. We prove that the optimal solution of the constrained MDP can be derived by solving the UP with one or two Lagrangian multipliers. Then we focus on the optimal policy of the average cost MDP (i.e., UP). We verify the existence of the optimal stationary policy of the average cost MDP and it can be obtained from the corresponding discount cost MDP. We derive two necessary conditions for the optimal policy of the average cost MDP (average cost optimal policy). Under certain conditions, the policy that serving nothing and allocating no energy from the battery is an average cost optimal policy. We also prove that serving everything combined with allocating the minimal of the total required power and total energy in the battery are an average cost optimal policy under other certain conditions. The monotonicities of the optimal object value with respect to Lagrangian multiplier and optimal policy regarding the state are investigated, respectively.

- We analyze the relations between the transmission rate and the power allocation from the battery. We find that given the transmission rate policy, the optimal battery power allocation policy is the greedy policy in some scenario. For general scenario, we propose a sufficient condition under which the optimal policy of two-dimensional MDP problem can be decomposed to the optimal policy of an MDP problem with the policy to be the transmission rate only in addition with the greedy battery power allocation policy.

- On the basis of the theoretical investigation, we propose an algorithm to find the constrained optimal policy under the finite state case. In addition, we propose three heuristic policies for the constrained MDP with the general state case: radical policy, conservative policy, and mixed policy.

The remainder of the paper is organized as follows. In Section II, the system model is described, and a mean buffer delay minimization problem with average grid power constraint is formulated. In Section III, the optimization problem is re-formulated as a constrained MDP and the optimal two-dimensional policy of the constrained MDP is investigated. Next, we discuss the relations between elements of the two-dimensional policy in Section IV. Based on the theoretical analysis, the algorithm to find the constrained optimal policy under the finite state space and heuristic policies for the general state space are proposed in Section V. Simulations are performed in Section VI. Finally, Section VII concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a slotted-time model of a point-to-point block fading channel. The length of a time-slot is τ units. The n-th time-slot is the time interval $\big[n\tau, (n+1)\tau\big)$. The channel gain remains static in each slot, and changes between different slots. The sequence of the channel gains is a finite-state ergodic Markov chain $\{H[n]\}$. The transmitter is assumed to have perfect channel state information (CSI). As shown in Fig. 1, at the end of the n-th slot, the higher layer generates $A[n]$ packets and they are stored in a buffer before transmission. It is assumed that each packet is with b bits and $\{A[n]\}$ is a finite-state ergodic Markov chain. We assume that the transmitter is equipped with an energy harvesting device and it can also get power from the power grid.[1] The harvested energy arrives at each end of the slot according to a finite-state ergodic Markov chain $\{E[n]\}$, and the harvested energy will be stored in a battery before consumption. There exists a long run average constraint on the grid power at the transmitter. At the beginning of the n-th time slot, the transmitter chooses $R[n]$ packets from the buffer and transmits to the receiver.[2] We assume the additive white Gaussian noise (AWGN) at the receiver is with zero mean and variance σ^2. In green communications, the total power required

---

[1]Grid power with average constraint is to guarantee user's QoS (delay). Specifically, due to the causality of harvested energy, the transmitter should accumulate a sufficient amount of energy before each packet transmission. Then the waiting time could be undesirably long since the randomness of harvested energy arrival. In contrast, when the grid power is available, even if the battery energy is insufficient, the transmitter could use the grid power to transmit packet. Hence, the user's QoS can be guaranteed.

[2]$R[n]$ is the transmission rate of the n-th timeslot with unit packets/timeslot.
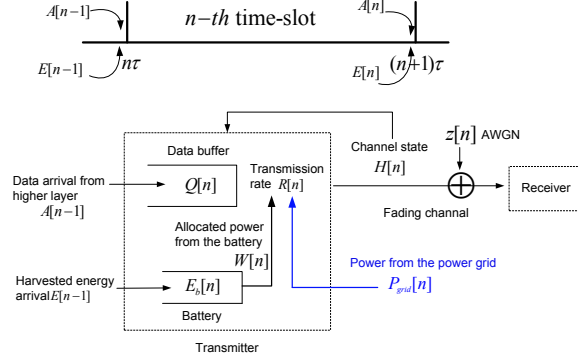
Fig. 1. System model

for reliable transmissions[3] of $R[n]$ packets in the n-th time-slot is [2]

$$P(X[n], R[n]) = \rho \frac{\sigma^2}{H[n]} (e^{\theta R[n]} - 1) + \Delta(R[n]), \tag{1}$$

where $X[n]$ is the system state that will be defined later, $\rho \geq 1$ is a constant, $\theta = \frac{2\ln(2)b}{N}$ with N being the channel uses in each time-slot, and

$$\Delta(R[n]) = \begin{cases} C, R[n] \neq 0; \\ 0, R[n] = 0, \end{cases} \tag{2}$$

where $C \geq 0$ is a constant. In particular, $\rho = 1$ and $C = 0$ when no circuit power is taken into account. In the transmission during the n-th timeslot, the transmitter allocates $W[n]$ power from the battery, and the remaining power will be supplied by the power grid. Denote $Q[n]$ as the queue length of the buffer at instance $n\tau$, the evolution equation for the buffer length is

$$Q[n+1] = Q[n] - R[n] + A[n]. \tag{3}$$

Assume that the capacity for the battery is E_{max}. Denote the battery's stored energy at instance $n\tau$ as $E_b[n]$, then the evolution equation for harvested energy in the battery can be given by

$$E_b[n+1] = \min\{E_b[n] - W[n]\tau + E[n], E_{max}\} := (E_b[n] - W[n]\tau + E[n])^-. \tag{4}$$

[3]In the paper, "reliable transmission" means totally error-free according to capacity arguments.

The objective is to find a rate and battery power allocation sequence that minimizes the mean buffer delay under the constraint on the long-run average grid power $\bar{\mathcal{P}}$, and the stochastic optimization problem is given by

$$\min_{\{(R[n],W[n])\}_{n=1}^{\infty}} \limsup_{n \to \infty} \frac{1}{n}\mathbb{E}\left[\sum_{k=0}^{n-1} Q[k]\right] \tag{5}$$

$$\text{s.t.} \begin{cases} \limsup_{n \to \infty} \frac{1}{n}\mathbb{E}\left[\sum_{k=0}^{n-1} P_{grid}[k]\right] \leq \bar{\mathcal{P}}, & \text{(6a)} \\[2ex] R[k] \leq Q[k], & \text{(6b)} \\[2ex] W[k]\tau \leq E_b[k], & \text{(6c)} \end{cases}$$

where $P_{grid}[k]$ is the power from the power grid,

$$P(X[k], R[k]) = P_{grid}[k] + W[k]. \tag{7}$$

III. ANALYSIS OF THE FORMULATED STOCHASTIC OPTIMIZATION PROBLEM

In this section, we first reconstruct the problem (5) as a constrained two-dimensional (i.e., rate and battery power allocation) MDP. Second, we prove that the constrained two-dimensional MDP can be transformed to unconstrained MDP by the Lagrangian method in Section III-B. Then we focus on the analysis of the unconstrained MDP in Section III-C. We verify the existence of the stationary policy for the unconstrained MDP (which is an average cost MDP) in Section III-C1. Next, we investigate the optimal policy of the average cost MDP, and structural properties of the average cost optimal policy are derived in Section III-C2. For better readability, the analysis flowchart for this section is illustrated in Fig. 2.
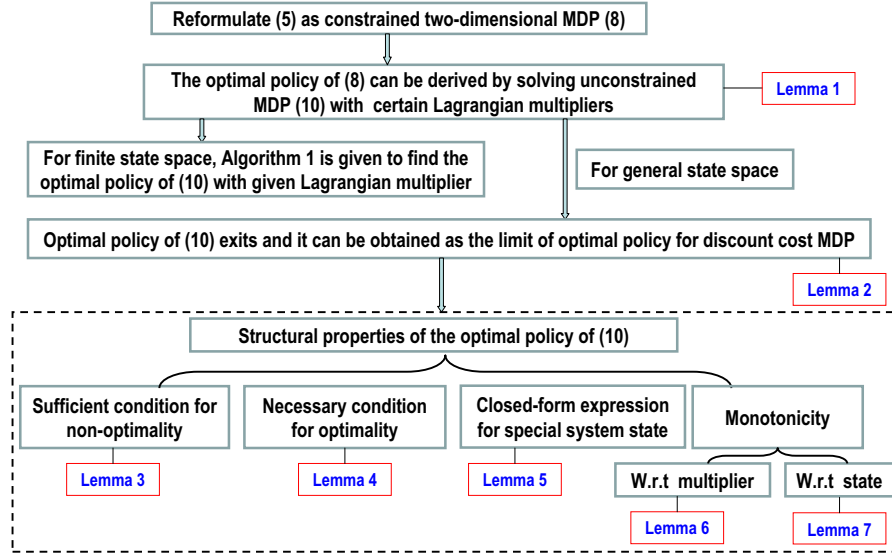
Fig. 2.   Analysis structure of Section III

A. Reconstruction as a constrained two-dimensional MDP

Define the state as $X[n] := (Q[n], H[n], A[n], E_b[n], E[n])$ with state space $\mathcal{X}$ and the action as $\mathfrak{A}[n] := (R[n], W[n])$ with action space $\mathcal{A}$, respectively.[4] Then $\{X[n], \mathfrak{A}[n]\}$ can be viewed as a Markov decision process (MDP). The feasible action (r, w) in a state $x = (q, h, a, e_b, e) \in \mathcal{X}$ belongs to $\mathcal{A}(x) = \{0, 1, \cdots, q\} \times \{0, \frac{1}{\tau}, \cdots, \frac{e_b}{\tau}\}$.[5] Define a policy $\pi = (\pi_0, \pi_1, \cdots)$ that π_n generates an action $(r[n], w[n])$ with a probability at instant $n\tau$ [20][21]. We denote the set of all policies as Π. Specially, a stationary deterministic policy is $\pi = (g, g, \cdots)$, where g is a measurable mapping from $\mathcal{X}$ to $\mathcal{A}$ such that $g(x) \in \mathcal{A}(x)$ for all $x \in \mathcal{X}$. Then, (5) is reformulated as the constrained MDP to find the two-dimensional (i.e., rate and battery power

---

[4]The system state includes the buffer queue length, channel gain, data arrival, energy in the battery, and harvested energy arrival. The action includes the allocated rate and the allocated battery energy.

[5]The harvested energy has been discretized.

allocation) optimal policy.

$$\min_{\pi \in \Pi} B_x^\pi = \limsup_{n \to \infty} \frac{1}{n} \mathbb{E}_x^\pi \left[\sum_{k=0}^{n-1} Q[k] \right] \tag{8}$$

$$\text{s.t.} \quad K_x^\pi = \limsup_{n \to \infty} \frac{1}{n} \mathbb{E}_x^\pi \left[\sum_{k=0}^{n-1} P_{grid}[k] \right] \leq \bar{\mathcal{P}}, \tag{9}$$

where the subscript $x = (q, h, a, e_b, e) \in \mathcal{X}$ is the initial system state.

B. Transformation to unconstrained MDP

Define $P_{grid}(x, r, w) := \max\{P(x, r) - w, 0\} := (P(x, r) - w)^+$ and $f_\beta(x, r, w) := q + \beta P_{grid}(x, r, w)$ with $\beta > 0$. Then we have a family of the following unconstrained problem (UP_β).

$$\min_\pi J_x^\pi(\beta) := \limsup_{n \to \infty} \frac{1}{n} \mathbb{E}_x^\pi \left[\sum_{k=0}^{n-1} f_\beta(X[k], R[k], W[k]) \right]. \tag{10}$$

In UP_β, $f_\beta(X[k], R[k], W[k])$ is the one-step cost in the k-th time-slot.

Remark: UP_β is an average cost MDP. Its optimal solution is called the average cost optimal policy.

The following lemma gives the relation between UP_β and the constrained two-dimensional MDP (8).

Lemma 1. When there exists a $\beta_0 > 0$ that the optimal policy of UP_{β_0} has an average grid power consumption equal to $\bar{\mathcal{P}}$, the optimal solution of UP_β is optimal for the constrained MDP in (8). Otherwise, there exit a $\beta^+ > 0$ and a $\beta^- > 0$. The optimal policy for the constrained MDP (8) is as follows: at each decision epoch, choose π^- with a certain probability q and π^+ with probability $1 - q$, where π^+ and π^- are the optimal policies obtained for $UP_{\beta+}$ and $UP_{\beta-}$, respectively. q depends on $\bar{\mathcal{P}}$ and the grid power consumptions of the two policies.

Proof: See Appendix A. ∎

Lemma 1 reveals that the solution of (8) can be obtained by solving UP_β with one or two β. In the following, we focus on the analysis of the unconstrained MDP, UP_β.

C. Analysis of the unconstrained MDP

1) Existence of the optimal policy: Define a discount cost MDP with discount factor $\alpha \in (0,1)$ corresponding to UP_β for each initial state $x = (q, h, a, e_b, e)$, with value function

$$V_\alpha(x) = \min_\pi \mathbb{E}_x^\pi \left[\sum_{k=0}^\infty \alpha^k \left(Q[k] + \beta P_{grid}(X[k], R[k], W[k]) \right) \right]. \tag{11}$$

The optimal solution for the discounted problem is referred to as a discount optimal policy.

The following lemma reveals the existence of the stationary policy. Furthermore, it derives how to obtain the optimal solution.

Lemma 2. There exists a stationary deterministic policy that solves UP_β with a $\beta > 0$, and it can be obtained as a limit of discount optimal policies as the discount factor increases to one.

Proof: See Appendix B. ∎

Following the proof of Lema 2, we can also derive that the optimal $J_x^{\pi^*}(\beta)$ is independent of the initial state x. Thus we can rewrite $J_x^{\pi^*}(\beta)$ as $J^{\pi^*}(\beta)$.

If the state is finite (Specifically, the data buffer state is finite), the relative value iteration algorithm (Algorithm 1) [26] can be utilized to find the optimal policy of the unconstrained MDP UP_β with given β. However, we are interested in deriving structural results on the optimal policies under general state space[6] and not simply solving the unconstrained problem with finite state space. Furthermore, some structural results are useful to solve the constrained MDP (Section V).

2) Structural properties: The average optimal policy are discussed in the subsection. First, the sufficient condition for non-optimality, necessary condition for optimality, and the closed-form expressions of optimal policy in special system states are given.

Lemma 3. In state $x = (q, h, a, e_b, e)$, $(r(x), w(x))$ is not the average cost optimal policy if $q - r(x) \neq 0$ and $e_b - w(x) + e > E_{max}$.

---

[6]The number of data buffer states can be infinite, then the state number can be infinite in the paper.

TABLE I

Algorithm 1: Relative value iteration algorithm of finding the optimal policy for UP$_\beta$

Step 1: Select initial value V^0, choose reference state $x^* \in \mathcal{X}$, specify ϵ, and set $n = 0$

Step 2: For each $x = (q, h, a, e_b, e) \in \mathcal{X}$, compute $V^{n+1}(x, \beta)$ by

$$V^{n+1}(x, \beta) = \min_{(r,w) \in \mathcal{A}(x)} \left\{ f_\beta(x, r, w) + \sum_{x' = (q', h', a', e_b', e') \in \mathcal{X}} p(x'|x, (r, w)) V^n(x', \beta) \right\}$$

where $p(x'|x, (r, w)) = \delta(q - r + a - q')\delta(e_b - w + e - e_b')p(h'|h)p(a'|a)p(e'|e)$

is the transition probability, $\delta(0) = 1$ and $\delta(x) = 0$ when $x \neq 0$.

Step 3: Normalize $V^{n+1}(x, \beta)$ for each $x \in \mathcal{X}$ as $V^{n+1}(x, \beta) = V^{n+1}(x, \beta) - V^{n+1}(x^*, \beta)$

Step 4: If $|V^{n+1} - V^n| < \epsilon$, go to next Step. Otherwise, $n = n + 1$ and go to Step 2.

Step 5: For each $x \in \mathcal{X}$, choose the policy according to

$$\pi(x, \beta) = \arg \min_{(r,w) \in \mathcal{A}(x)} \left\{ f_\beta(x, r, w) + \sum_{x' \in \mathcal{X}} p(x'|x, (r, w)) V^n(x', \beta) \right\}$$

Proof: When a policy results in battery overflow (i.e., $e_b - w(x) + e > E_{max}$) and non-emptiness of the buffer (i.e., $q - r(x) \neq 0$), then in terms of the average cost performance, the policy can be improved by using the overflowed energy for transmitting some (parts or all) remaining buffer data. The reasons are as follows. First, using overflowed energy for transmitting some (parts or all) remaining buffer data will not increase one-step cost since no extra grid power is utilized. Second, using overflowed energy for transmitting some (parts or all) remaining buffer data will decrease the initial buffer data for future while the initial battery energy for future does not change (remains E_{max}). Using Property 1 in Appendix C1, we derive that the average cost will be decreased. ∎

Remark: Lemma 3 means if a policy results in battery overflow but non-emptiness of the buffer, there are (is) polices (policy) that can achieve better average cost performance definitely.

Remark: Lemma 3 gives a sufficient condition for the non-optimality. Meanwhile, Lemma 3 can be also viewed as the necessary condition for the optimality. That is to say, any average optimal policy should not incur battery overflow and non-emptiness of the buffer simultaneously.

Next, based on Lemma 2 and Proposition 1 in Appendix C2, we have the following lemma.

Lemma 4. Given state $x = (q, h, a, e_b, e)$, the average cost optimal policy $(r^*(x), w^*(x))$ should

satisfy the following inequality array

$$\tilde{Z}_1(q, q - r^*, h, a, e_b - w^*, e) \le \beta\rho\frac{\sigma^2}{h}e^{\theta q}(e^\theta - 1) \le \tilde{Z}_1(q, q - r^* + 1, h, a, e_b - w^*, e), \quad (12)$$

$$\tilde{Z}_2(q - r^*, h, a, e_b - w^*, e) \le \frac{-\beta}{\tau} \le \tilde{Z}_2(q - r^*, h, a, e_b - w^* + 1, e), \quad (13)$$

$$\tilde{Z}_3(q, q - r^*, h, a, e_b - w^*, e) \le \beta\rho\frac{\sigma^2}{h}e^{\theta q}(e^\theta - 1) \le \tilde{Z}_3(q, q - r^* + 1, h, a, e_b - w^* + 1, e), \quad (14)$$

where $\tilde{Z}_1(q, u, h, a, \eta, e) = \lim_{\alpha \to 1} Z_1(q, u, h, a, \eta, e)$, $\tilde{Z}_2(u, h, a, \eta, e) = \lim_{\alpha \to 1} Z_2(u, h, a, \eta, e)$, and $\tilde{Z}_3(q, u, h, a, \eta, e) = \lim_{\alpha \to 1} Z_3(q, u, h, a, \eta, e)$. $Z_i(\cdot)$ $(i = 1, 2, 3)$ is defined in Proposition 1.

Remark: Lemma 4 reveals a necessary condition for the average cost optimality, i.e., the optimal transmit rate r^* and the optimal battery energy allocation w^* should satisfy the condition.

Remark: When (r^*, w^*) is on the boundary of the feasible set, corresponding conditions can also be obtained similarly.

Combining Lemma 2 and Proposition 2 in Appendix C2, we derive the following lemma.

Lemma 5. For $x = (q, h, a, e_b, e)$ satisfying

$$\tilde{Z}_1(q, 0, h, a, \tau\max\{0, \frac{e_b}{\tau} - P(x, q)\}, e) > \beta\rho\frac{\sigma^2}{h}e^{\theta q}(e^\theta - 1) \quad (15)$$

and

$$\tilde{Z}_2(0, h, a, \tau\max\{0, \frac{e_b}{\tau} - P(x, q)\}, e) > \frac{-\beta}{\tau}, \quad (16)$$

$(q, e_b - \tau\max\{0, \frac{e_b}{\tau} - P(x, q)\})$ is the average cost optimal policy. In addition, for (q, h, a, e_b, e) satisfying

$$\tilde{Z}_1(q, q, h, a, e_b, e) < \beta\rho\frac{\sigma^2}{h}e^{\theta q}(e^\theta - 1) \quad (17)$$

and

$$\tilde{Z}_2(q, h, a, e_b, e) < \frac{-\beta}{\tau}, \tag{18}$$

$(0, 0)$ is the average cost optimal policy.

Remark: $(q, e_b - \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\})$ *means transmit all the data in the buffer and the allocated battery energy is* $e_b - \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\}$*). That is to say, transmit all data in the buffer and allocate as much energy as possible from the battery. Specifically, if the required power for transmitting all buffer data is less than the power stored in the battery, allocate all the required power from the battery. Otherwise, allocate all the battery's energy (the rest of the required power will be allocated from the power grid).* $(0, 0)$ *means transmit no buffer data and allocate no battery energy.*

Remark: (15) and (16) give the set of states, for which transmit all the buffer data the together with allocate as much energy as possible from the battery is the two-dimensional average cost optimal policy. (17) and (18) give the set of states, for which transmit no buffer data together with allocate no battery energy is the two-dimensional average cost optimal policy.

In the following, we investigate the monotonicity.

Lemma 6. Denote the optimal stationary deterministic policy for UP_β as g_β, we have

- $J^{g_\beta}(\beta)$ is non-decreasing in β.
- B^{g_β} is non-decreasing in β, and K^{g_β} is monotone non-increasing in β

Proof: See Appendix F. ∎

Lemma 7. The average cost optimal transmit rate policy $r(q, h, a, e_b, e)$ is non-decreasing in q and e_b, respectively; The average cost optimal battery energy allocation policy $w(q, h, a, e_b, e)$ is non-decreasing in q and e_b, respectively.

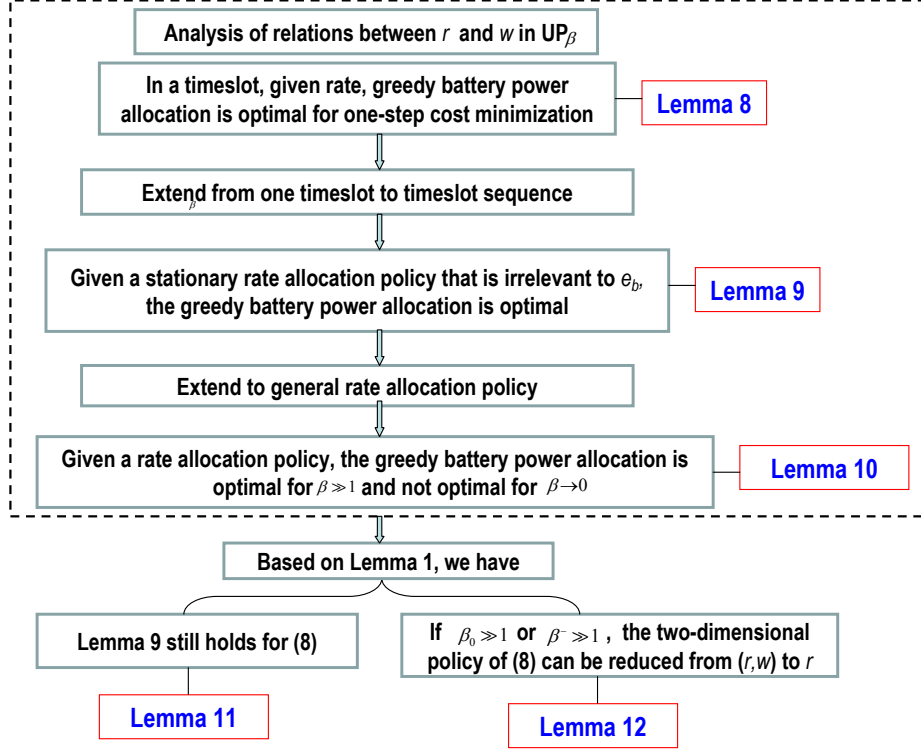*Proof:* The lemma can be proved by the second half of Lemma 2 and Proposition 3 in Appendix C2. ∎

Fig. 3. Analysis structure of Section IV, the derived results are also shown correspondingly

IV. RELATIONS BETWEEN RATE ALLOCATION AND THE BATTERY POWER ALLOCATION

The rate allocation r and the power allocation from the battery w are coupled together, they affect each other. In this section, we investigate the relations between the rate allocation r and the power allocation from the battery w. We first focus on the relation between r and w in UP_β. Next, we derive that under a condition, the policy of the constrained two-dimensional MDP problem (8) can be reduced to the rate policy only. To make the presentation clear, the analysis structure of this section is drawn in Fig. 3.

A. The relation between r and w in UP_β

If we assume that rate $r[n]$ has been chosen at the n-th time-slot, then the required total power has been fixed. In this case, to minimize the immediate one-step cost $q[n] + \beta \left[\rho \frac{\sigma^2}{h[n]} (e^{\theta r[n]} - 1) + \Delta(r[n]) - w[n] \right]^+$, we will allocate as much power as possible from the battery to meet the

required total power, i.e., the greedy policy for the battery power allocation. This is because the power from the battery is "free".[7] Formally, we have the following claim.

Lemma 8. In a timeslot, if the rate allocation r is chosen, the greedy battery power allocation is optimal for the immediate one-step cost minimization.

Proof: See Appendix D. ∎

Remark: Lemma 8 reveals the optimality of greedy battery power allocation for given rate in a time-slot.

In the following, we consider the extension from one time-slot to the timeslot sequence. First, we have the following lemma.

Lemma 9. For a given rate allocation policy $r(q, h, a, e_b, e)$ that is irrelevant to e_b, i.e., $r(q, h, a)$,[8] the greedy battery power allocation is optimal (for UP_β).

Proof: See Appendix E. ∎

Remark: The irrelevance to e_b of rate allocation policy is sufficient condition for the optimality of greedy battery power allocation. Lemma 9 guarantees the optimality of greedy battery power under any given rate allocation policy irrelevant to e_b.

Next, a natural question is *whether the greedy allocation strategy of battery power is optimal given general rate allocation policy $r(x = (q, h, a, e_b, e))$?* The following lemma gives the answer.

Lemma 10. Given a rate allocation policy $r(x)$,[9]

- When β is large enough, e.g., $\beta \gg 1$, the greedy policy is the optimal battery power allocation policy in UP_β.
- If β is sufficiently small, e.g., $\beta \to 0$, the greedy battery power allocation policy is NOT optimal for UP_β.

---

[7] Please refer to (10). the price of the grid power is β.

[8] According to (4), if a policy is irrelevant to e_b, then it is irrelevant to e.

[9] According to Lemma 7, it is reasonable to assume that $r(x)$ is non-decreasing in q and e_b, respectively.

Proof: See Appendix G. ∎

Remark: Lemma 10 reveals that the greedy policy is NOT the optimal battery power allocation policy in UP$_\beta$ with arbitrary β. The optimality of greedy battery power allocation depends on the value of β. It can be explained as follows: Since β is the "price" of grid power in UP$_\beta$, when the grid power is very cheap, the profit of reserving some battery power for future timeslot[10] is more than the cost of buying the same amount of grid power in current timeslot. Thus, reserving some battery energy but using the grid power instead is optimal. When the price is high, the cost of buying the grid power is more than the profit of reserving some battery energy, then allocate as much energy as possible from the battery to fulfill the required power (i.e., greedy battery allocation policy) is optimal.

Remark: As the remaining battery energy will affect action and cost in future timeslot for given rate policy (e.g., battery power allocation $w[n]$ at the n-th time-slot will affect the rate allocation $r[n+1]$ at the $(n+1)$ times-lot), the optimality of greedy battery power allocation can not extend from one timeslot (Lemma 8) to time-slot sequence.

B. Dimension reduction for the two-dimensional policy of the constrained MDP under a sufficient condition

According to Lemma 1, the two-dimensional optimal policy of constrained MDP (8) can be derived by the optimal policy of the UP$_\beta$ with one or two values of β. Then we have

Lemma 11. For a given rate allocation policy $r(q, h, a, e_b, e)$ that is irrelevant to e_b, i.e., $r(q, h, a)$, the greedy battery power allocation is optimal (for the constrained MDP (8)).

Furthermore, the following lemma reveals that the two-dimensional policy of the constrained MDP can be reduced to the rate policy when β_0 or β^- satisfies a condition.

---

[10]Based on Property 2 in Appendix C1, there exists profit for reserving some battery power for future timeslot. The price of using grid energy is constant over time, the cost of using grid power is constant. But reserving battery energy can incur more data transmission in future (Observe that the rate policy has been given already, more battery power leads to more data transmission). That is to say, delaying the use of battery energy has profits in minimizing data delay. All in all, there are profits for the first part of $J_x^\pi(\beta)$.

Lemma 12. If $\beta_0 \gg 1$ or $\beta^- \gg 1$, the greedy policy is the optimal battery power allocation policy of the two-dimensional constrained MDP (8). Furthermore, view $(X[n], R[n])$ as an MDP with state $X[n]$ and action $R[n]$.[11] The feasible action r in state $x = (q, h, a, e_b, e)$ belongs to $\{0, 1, \cdots, q\}$. Define $\pi_r = (\pi_r[0], \pi_r[1], \cdots)$ to be a policy that $\pi_r[n]$ generates an action $r[n]$ at $n\tau$, the optimal policy of the following MDP problem is the optimal rate policy of (8).

$$\min_{\pi_r} \limsup_{n \to \infty} \frac{1}{n} \mathbb{E}_x^{\pi_r} \left[\sum_{k=0}^{n-1} Q[k] \right] \tag{19}$$

$$\text{s.t.} \quad \limsup_{n \to \infty} \frac{1}{n} \mathbb{E}_x^{\pi_r} \left[\sum_{k=0}^{n-1} P_{grid}[k] \right] \leq \bar{\mathcal{P}}, \tag{20}$$

where $P_{grid}[k] = P(X[k], R[k]) - \min \left\{ P\left(X[k], R[k]\right), \frac{E_b[k]}{\tau} \right\}$, and the evolution of energy in the battery becomes $E_b[k+1] = \left(E_b[k] - \tau \min \left\{ P\left(X[k], R[k]\right), \frac{E_b[k]}{\tau} \right\} + E[k] \right)^{-}$.

Proof: See Appendix H. ∎

Remark: When the condition $\beta_0 \gg 1$ or $\beta^- \gg 1$ holds, the two-dimensional policy can be obtained as follows. We can first derive the optimal battery allocation policy of the two-dimensional policy to be greedy policy, and then the optimal rate policy can be solved through an MDP whose policy includes the rate allocation only (i.e., (19)). The dimension of the policy has reduced from (r, w) to r.

Remark: If $\beta_0 \gg 1$ or $\beta^- \gg 1$, the dimension reduction can be implemented. In contrast, if $\beta_0 \to 0$ or $\beta^+ \to 0$, the dimension reduction in Lemma 12 can not be accomplished (See the second half of Lemma 10). For other cases, we do not know whether the dimension reduction can be implemented. $\beta_0 \gg 1$ or $\beta^- \gg 1$ is only a sufficient condition for dimension reduction in Lemma 12.

Since there is a condition $\beta_0 \gg 1$ or $\beta^- \gg 1$ in Lemma 12 and the dimension reduction does not hold for $\beta_0 \to 0$ or $\beta^+ \to 0$, formulating the original optimization problem (5) directly as

---

[11]The state includes the buffer queue length, channel gain, data arrival, energy in the battery, and harvested energy arrival. The action includes the allocated rate only.

Fig. 4. Structure of Section V

(19) is NOT convincing.[12]

V. POLICY OF THE CONSTRAINED MDP

Based on the previous theoretical results, an algorithm to find the constrained optimal policy is proposed for the finite state space, and heuristic polices are given for the general state space. The structure of this section is illustrated in Fig. 4.

A. Algorithm to find the optimal policy for finite state space

In this subsection, we give the algorithm to find the constrained optimal policy when the state is finite.

According to Lemma 6, smaller β results in better delay performance B. Meanwhile, the decrease of β will increase the grid power consumption K. Too small β will violate the grid

---

[12]If we can prove that the condition $\beta_0 \gg 1$ or $\beta^- \gg 1$ holds, (5) can be reformulated as (19).

TABLE II

Algorithm 2: Algorithm of finding the constrained optimal policy for finite state
Step 1:   Using iteration algorithm (21) to find β^*, and the corresponding average grid   power $K^{g_{\beta^*}}$, in which the relative value iteration algorithm (Algorithm 1) is applied.   **Step 2:**   If $K^{g_{\beta^*}} = \bar{\mathcal{P}}$, then g_{β^*} is the optimal policy of the constrained MDP. Otherwise, go to next Step.   **Step 3:**   Perturb β^* by ν: $\beta^+ = \beta^* + \nu$ and $\beta^- = \beta^* - \nu$. Find the optimal policies g_{β^+} and g_{β^-} for UP_{β^+} and UP_{β^+} as well as the corresponding grid power $K^{g_{\beta^+}}$ and $K^{g_{\beta^-}}$, respectively, by using   Algorithm 1. The optimal policy is taking g_{β^+} with probability ξ and g_{β^-} with probability $1 - \xi$ at each decision stage. ξ is determined by $\xi K^{g_{\beta^+}} + (1 - \xi)K^{g_{\beta^-}} = \bar{\mathcal{P}}$.

power constraint. Then we should find the smallest β that satisfying the average grid power first. Denote $\beta^* = \inf\{\beta : K^{g_\beta} \leq \bar{\mathcal{P}}\}$, where g_β is the optimal policy of UP_β. We can use the following method to find β^*. Let

$$\beta_{n+1} = \beta_n + \frac{1}{n}\big(K^{g_{\beta_n}} - \bar{\mathcal{P}}\big) \tag{21}$$

with β_1 is a sufficiently large number. $K^{g_{\beta_n}}$ is computed by using the relative value iteration algorithm for each β_n. Then $\{\beta_n\}$ converges to β^* [25]. Based on Lemma 1, if the average grid power $K^{g_{\beta^*}}$ equals to the grid power constraint, the obtained optimal policy is also optimal for the constrained MDP. Otherwise, we should find β^+ and β^-. The detailed algorithm for finite state is listed in Table II.

B. Proposed heuristic policies

Algorithm 2 is only for the finite state space. Meanwhile, it is time-consuming when the number of states is large. In this subsection, we propose low-complex heuristic policies for general state space. The paper has derived the structural properties of the optimal policy. Particularly, we have proved that the optimal policy exists, and it is a stationary deterministic policy or a mixed policy of two stationary deterministic policies. Moreover, we have proved that the greedy battery power allocation MAY BE optimal (in Section IV). Based on these properties and in

light of Algorithm 2, we propose heuristic policies as follows (a summary is given in Table III).

The first is named radical policy. Under radical policy, the action is $(r = q, w = \min\{e_b, P(x, r)\})$ for state $x = (q, h, a, e_b, e)$. That is to say, all the buffer data are served at each time-slot, and use the greedy strategy for the battery energy allocation, i.e., if the required power is not greater than the battery power, then all the power will be supplied from the battery and no grid power will be used. Otherwise, allocate all the battery power, and the rest will be supplied from the power grid.

Remark: When there is no average grid power constraint, the radical policy is the optimal policy to minimize the mean buffer delay. Furthermore, given an average grid power constraint, when the mean date arrival, mean energy arrival, and mean channel gain satisfy a condition, the grid power constraint can be obeyed under radical policy, the radical policy is the optimal policy even when considering the average grid power constraint.

In the radical policy, the average grid power constraint is not considered. Then we propose another policy (i.e., the conservative policy) that guarantees the average grid power constraint through satisfying the constraints in each time-slot. Define $P^{-1}(\cdot)$ as the inverse function of $P(x, r)$ with respect to r. We call the policy $(r(x), w(x)) = \left(\min\left\{ q, P^{-1}\left(\bar{P} + \frac{e_b}{\tau}\right) \right\}, \min\{\frac{e_b}{\tau}, P(x, r)\} \right)$ the conservative policy. That is to say, we first guarantee that the grid power utilized in each time is less than the average grid constraint, then transmit as many packets as possible and utilize the greedy policy for the battery energy allocation.

The third policy is a random policy referred to as mixed policy. In the mixed policy, the radical policy and conservative policy are utilized randomly with probability ξ and $1 - \xi$, respectively. Denote the average grid power consumptions of the radical policy and conservative as G_r and G_c, respectively. ξ is determined by $\xi * G_r + (1 - \xi) * G_c = \bar{\mathcal{P}}$.

VI. NUMERICAL RESULTS

In this section, simulation results are presented under the radical policy, conservative policy and mixed policy. We consider the i.i.d. Rayleigh fading channel (i.e., the power gain H is exponentially distributed). In addition, unless otherwise specified, we set $\tau = 1$, $b = 1$, $N = 5$,

TABLE III

Policy name	Strategy $(r(x), w(x))$ for $x = (q, h, a, e_b, e)$
Radical policy	$(q, \min\{e_b, P(x, r)\})$
Conservative policy	$\left(\min\left\{q, P^{-1}\left(\bar{P} + \frac{e_b}{\tau}\right)\right\}, \min\{\frac{e_b}{\tau}, P(x, r)\}\right)$
Mixed policy	Apply the radical policy and conservative policy with probability ξ and $1 - \xi$, respectively

and $\rho = 1$. Both the initial battery energy and initial buffer length are zero.

Fig. 5 plots the average grid power consumption with respect to the average data arrival ($\bar{A}$) under radical policy. We can observe that when $\bar{A}$ is small, the grid power consumption is nearly zero. However, when $\bar{A}$ is large the grid power consumption grows rapidly with the increase of $\bar{A}$ roughly according to exponential relation. This can be explained as follows: when $\bar{A}$ is small, the required power is small and the battery can supply the power. Then no grid power will be consumed. Once $\bar{A}$ is large, the required power is much larger than the battery power, and the grid power becomes the main power source. Since the required power roughly varies with the transmission rate according to the exponential function, the grid power consumption varies exponentially with $\bar{A}$. Meanwhile, we can see that the better channel conditions lead to less grid power consumption.

Furthermore, from Fig. 5, it can be derived that if $\bar{A}$ is less than a certain value, the grid power will be less than a certain value. Since the radical policy is optimal for the buffer delay minimization without the average grid power constraint, if $\bar{A}$ is less than some value to make the average grid power be no more than the constraint, i.e., the average grid power constant is satisfied, then the radical policy is also optimal when considering the grid power constraint. For example, when $\bar{\mathcal{P}} = 2000$, according to Fig. 5, the strategy is optimal for $\bar{A} = 1, 2, \cdots, 8$. The reason is that when the average power grid plus the harvested power is large enough to serve all the data, then serving all is optimal.

Fig. 6 illustrates the average buffer length performance for conservative policy. A takes values from $\{0, 10, 20, 30\}$ with probabilities $\{0.1, 0.3, 0.5, 0.1\}$, respectively. E takes values
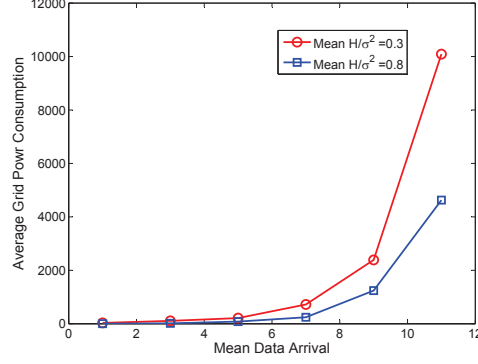
Fig. 5. Average grid power consumptions v.s. $\bar{A}$. $C = 1$ and $E_{max} = 2500$. A takes 0 and $2*\bar{A}$ with equal probability 0.5. E takes values $\{200, 800, 1000, 2000\}$ with probabilities $\{0.1, 0.6, 0.2, 0.1\}$, respectively. The mean grid power is average over 10^6 time-slots.

$\{200, 800, 1000, 2000\}$ with probabilities $\{0.1, 0.6, 0.2, 0.1\}$, respectively. In Fig. 6(a), the buffer length is averaged over 10^5 time-slots. From the figure, we can see that the mean buffer length decreases fast when $\bar{\mathcal{P}}$ is small (e.g., $\bar{\mathcal{P}} \leq 1000$), and the decrease becomes slow when $\bar{\mathcal{P}}$ is large (e.g., $\bar{\mathcal{P}} > 2000$). This can be explained as follows: when the upper bound of the average grid power (i.e., $\bar{\mathcal{P}}$) increases, there are more available grid power in a time-slot in average, sense and we can transmit more (at least no less) buffer data, then the average buffer length becomes shorter. When $\bar{\mathcal{P}}$ is small, $r = \min\left\{q, P^{-1}\left(\bar{\mathcal{P}} + \frac{e_b}{\tau}\right)\right\} = P^{-1}\left(\bar{\mathcal{P}} + \frac{e_b}{\tau}\right)$ with a high chance, hence r increases apparently with the increase of $\bar{\mathcal{P}}$, and the average buffer length decreases quickly. Once $\bar{\mathcal{P}}$ is large enough, $r = \min\left\{q, P^{-1}\left(\bar{\mathcal{P}} + \frac{e_b}{\tau}\right)\right\} = q$ with a high probability, and r becomes static with respect to $\bar{\mathcal{P}}$. Then, the average buffer length decreases slowly. Furthermore, we can observe that more extra circuit power consumption (i.e., C) and smaller battery capacity can respectively result in worse mean buffer length performance (i.e., longer length). Meanwhile, by comparing $(C = 1, E_{max} = 850)$ with $(C = 100, E_{max} = 2500)$, we can find that the mean buffer length performance for $(C = 1, E_{max} = 850)$ is better when $\bar{\mathcal{P}}$ is small. But when $\bar{\mathcal{P}}$ is large, $(C = 100, E_{max} = 2500)$ has slightly better performance.

In Fig. 6(b), for each curve, we can observe that the buffer length performance decreases with the increase of $\overline{H/\sigma^2}$, fast when $\overline{H/\sigma^2}$ is small (e.g. $0.1, 0.2, 0.3$), moderately when $\overline{H/\sigma^2}$ is large (e.g., $0.4, \cdots, 0.7$), and slowly when $\overline{H/\sigma^2}$ is very large (e.g., $0.8, 0.9$). The reason is as follows. When $\overline{H/\sigma^2}$ is not very large, $q > P^{-1}(\frac{e_b}{\tau})$ with a high probability, i.e., $r(x) = P^{-1}(\frac{e_b}{\tau})$. The
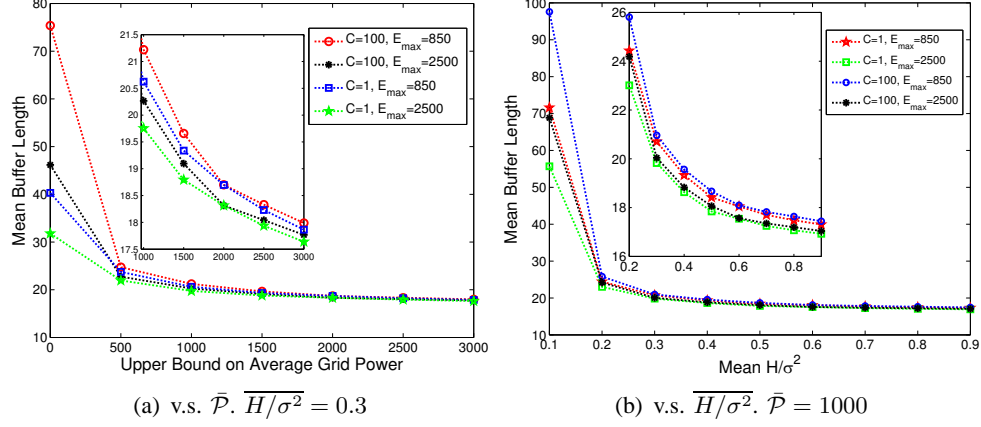
Fig. 6. The mean buffer length performance for conservative policy.

remaining buffer length $u(x) = q - r(x) = q - P^{-1}(\frac{e_b}{\tau})$ will decrease with the increase of h/σ^2 approximately according to minus logarithmic relation.[13] Thus, the mean buffer length decreases harshly at first and moderate then. Once $\overline{H/\sigma^2}$ is larger than a certain value, $q < P^{-1}(\frac{e_b}{\tau})$ with a high probability. Then, $r(x) = q$ and the remaining buffer length becomes zero with a high probability. In this case, the increase of $\overline{H/\sigma^2}$ will not have great effects on the mean buffer length (the mean length is nearly the average data arrival, 16).

Fig. 7 compares the buffer length performance of the heuristic policies with respect to $\bar{H}/\sigma^2$. In the simulations, A takes values from $\{0, 10, 20, 30\}$ with probabilities $\{0.1, 0.5, 0.3, 0.1\}$, respectively. E takes values $\{200, 800, 1000, 2000\}$ with probabilities $\{0.1, 0.6, 0.2, 0.1\}$, respectively.. $E_{max} = 2500$ and $\bar{\mathcal{P}} = 3000$. Based on the average grid power consumptions of radical policy and conservative policy (as plotted in Fig. 8), we compute the probability of using radical policy in the mixed policy, $\xi = [0.9468\ 0.8615\ 0.7933\ 0.7463\ 0.7053\ 0.6608\ 0.6689\ 0.6452]$. We can see that in terms of the buffer length performance, the radical policy is better than the mixed policy, which is better than the conservative policy. For the conservative policy and mixed policy, the buffer length decreases with the increase of $\bar{H}/\sigma^2$ first harshly and then moderately. The explanations for the conservative policy are similar to Fig. 6(b). As the usage probability of the conservative policy in mixed policy is high, the buffer length of the mixed policy is similar as

---

[13]$P^{-1}(\cdot)$ is increasing with h/σ^2 according to logarithmic relation.
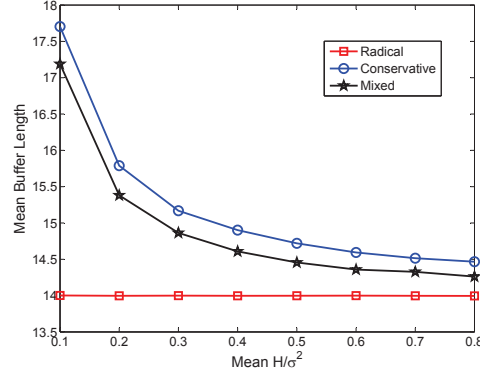
Fig. 7. Buffer length performance of the radical policy, conservative policy, and mixed policy

(a) Radical policy
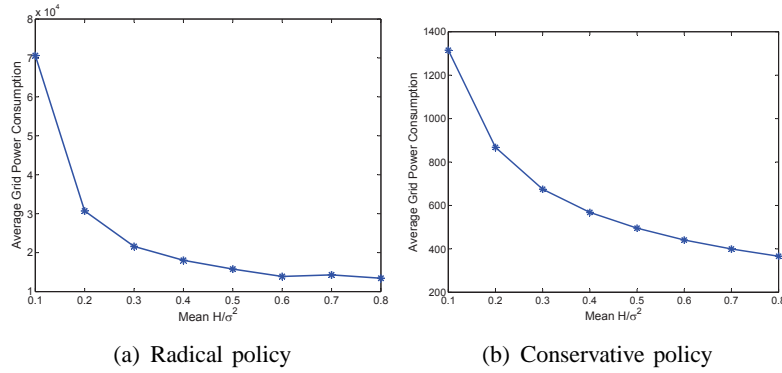
(b) Conservative policy

Fig. 8. The average grid power consumptions of the radical policy and conservative policy.

the conservative policy. Meanwhile, as there is chance of using the radical policy in the mixed policy, the buffer length performance of the mixed policy is better than the conservative policy. The mean buffer length of the radical policy is approximately the mean data arrival and remains static. As the radical policy is the optimal policy without the grid power constraint, the buffer length of the radical policy is the lower bound of the optimal policy.

VII. CONCLUSION

In this paper, we have studied the power allocation of the physical layer together with the optimal mean buffer delay of the upper layer in green networks with energy harvesting nodes. The physical power allocation contains two aspects: power allocation from the power grid and power allocation from the battery. The rate allocation can represent the total power allocation and the grid power allocation is the total power subtract the battery power, then the physical power allocation is equivalent to rate allocation and battery power allocation. For the purpose of

modeling and analyzing the conflicting relation between power and delay as well as the coupling between rate allocation and battery power allocation, we reformulate a constrained MDP with a two-dimensional policy. The analysis of the constrained MDP is transformed to that of the corresponding unconstrained MDP. Structural properties of the optimal policy are derived. In addition, the relations between elements of the two-dimensional policy are also investigated. According to the theoretical study, an algorithm to find the constrained optimal policy is proposed for finite state space. Furthermore, heuristic policies (i.e., the radical policy, the conservative policy and the mixed policy) are presented for general state. In the end, simulations are carried out under these policies. We have observed the interactions among the channel, the data arrival, the harvested energy arrival, the power grid, and the data buffer length.

APPENDIX

A. Proof of Lemma 1

If for some β (denoted as β_0), the optimal stationary policy π^* of UP_{β_0} satisfies: 1) π^* yields B^{π^*} and K^{π^*} as limits for all $x \in \mathcal{X}$; 2) $K^{\pi^*} = \bar{\mathcal{P}}$. Then π^* is optimal for the constrained MDP (8) according to [22][23]. Otherwise, there are β^+ and β^-. The optimal policy π^- that obtained for UP_{β^-} has a grid power consumption slightly larger than $\bar{\mathcal{P}}$. $\beta^+ > \beta^-$ will instead lead to a less aggressive policy π^+ with a grid power consumption slightly smaller than $\bar{\mathcal{P}}$. The optimal policy for the constrained MDP (8) is as follows: at each decision epoch, choose π^- with a certain probability q and π^+ with probability $1 - q$, where q depends on $\bar{\mathcal{P}}$ and the grid power consumptions of the two policies [23][24].[14]

B. Proof of Lemma 2

We prove the lemma by applying Theorem 3.8 in [27]. First, we can prove that the conditions of Proposition 2.1 in [27] holds. Next, the discounted cost optimality equation [28] for $V_\alpha(x)$ is

$$V_\alpha(q, h, a, e_b, e) = \min_{r \in \{0, 1, \cdots, q\}, w \in \{0, \frac{1}{\tau}, \cdots, \frac{e_b}{\tau}\}} \left\{ q + \beta \left[\rho \frac{\sigma^2}{h} (e^{\theta r} - 1) + \Delta(r) - w \right]^+ + \alpha \right.$$

$$\left. \times \mathbb{E}_{h, a, e} \left[V_\alpha(q - r + A, H, A, (e_b - w\tau + E)^-, E) \right] \right\}. \quad (22)$$

---

[14]The state space is countable.

We can see that $V_\alpha(q, h, a, e_b, e)$ is increasing in q and non-increasing in e_b given (h, a, e) since the larger the initial buffer the larger will be the cost to go, and the larger the initial battery energy the smaller will be the cost.[15] Thus, $\arg\inf_{y \in \mathcal{X}} V_\alpha(y) = (0, h_0, a_0, E_{max}, e_0) := x_0$, i.e., the infimum is obtained when the system begins with an empty buffer, a full battery, and for some channel sate h_0, arrival state a_0, and harvested energy arrival state e_0. When the buffer is empty, the set of feasible rate is $\{0\}$. Then $f(x_0, 0, w) = 0$, we get

$$
\begin{aligned}
V_\alpha(x_0) &= \min_{w \in \{0, \frac{1}{\tau}, \cdots, \frac{E_{max}}{\tau}\}} \alpha \mathbb{E}_{h_0, a_0, e_0} \left[V_\alpha(A, H, A, (E_{max} - w\tau + E)^-, E) \right] \\
&= \alpha \mathbb{E}_{h_0, a_0, e_0} \left[V_\alpha(A, H, A, E_{max}, E) \right].
\end{aligned}
\tag{23}
$$

Meanwhile, since policy $(q, 0)$ is feasible for state (q, h, a, e_b, e), then

$$
V_\alpha(x) \leq q + \rho \frac{\sigma^2}{h} (e^{\theta q} - 1) + C + \alpha \mathbb{E}_{h,a,e} \left[V_\alpha(A, H, A, (e_b + E)^-, E) \right].
\tag{24}
$$

Let the system start in state $(a, h, a, e_b + e, e)$, we take the action $r[n] = a[n]$ and $w[n] < e[n]$ for all n. Let $\xi(h, a, e_b, e)$ be the expected number of slots to hit the state $(a_0, h_0, a_0, E_{max}, e_0)$.[16] Observe that $\xi(h, a, e_b, e)$ is finite. Let $c_{max} = \max_{h,a} \left\{ a + \rho \frac{\sigma^2}{h} (e^{\theta a} - 1) \right\} + C$. Applying the Wald's lemma [29], we get

$$
\begin{aligned}
\alpha \mathbb{E}_{h,a,e} \left[V_\alpha(A, H, A, (e_b + E)^-, E) \right] &\leq c_{max} \xi(h, a, e_b, e) + \alpha \\
\times \quad \mathbb{E}_{h_0, a_0, e_0} \left[V_\alpha(A, H, A, E_{max}, E) \right] &= c_{max} \xi(h, a, e_b, e) + V_\alpha(x_0).
\end{aligned}
\tag{25}
$$

In (25), we have used $(E_{max} + E)^- = E_{max}$. Next, combining (24) and (25), we have $V_\alpha(x) \leq q + \rho \frac{\sigma^2}{h} (e^{\theta q} - 1) + C + c_{max} \xi(h, a, e_b, e) + V_\alpha(x_0)$. Thus, $V_\alpha(x) - V_\alpha(x_0) \leq q + \rho \frac{\sigma^2}{h} (e^{\theta q} - 1) + C + c_{max} \xi(h, a, e_b, e) < \infty$. Third, there exits a policy $\pi \in \mathcal{A}$ and an initial state $x \in \mathcal{X}$ such that $J_x^\pi(\beta) < \infty$ in the practical problem. Otherwise, the cost is infinite for all policies and any policy is optimal. Based on the above analysis, the conditions in Theorem 3.8 in [27] hold, and then we prove the lemma.

---

[15]See the formal proof at Property 1 and Property 2 in Appendix C1.

[16]When $w[n] < e[n]$, E_{max} is the absorbing state of the battery energy.

C. Optimal policy for the discount cost MDP

1) Properties of $V_\alpha(q, h, a, e_b, e)$: Property 1 - Property 3 give the properties of the value function $V_\alpha(q, h, a, e_b, e)$.

Property 1. $V_\alpha(q, h, a, e_b, e)$ is an increasing function of q.

Proof: We verify the increasing property by induction. The value iteration algorithm (or successive approximation method) corresponding to (32) is

$$
\begin{aligned}
V_{\alpha,n}(q, h, a, e_b, e) &= \min_{u \in \{0,1,\cdots,q\}, \eta \in \{0,1,\cdots,e_b\}} \left\{ q + \beta \left[\rho \frac{\sigma^2}{h}(e^{\theta(q-u)} - 1) + \Delta(q - u) - \frac{e_b - \eta}{\tau} \right]^+ + \alpha \right. \\
&\left. \times\quad \mathbb{E}_{h,a,e} \left[V_{\alpha,n-1}(u + A, H, A, (\eta + E)^-, E) \right] \right\}
\end{aligned} \tag{26}
$$

with $V_{\alpha,0}(q, h, a, e_b, e) = 0$. Accordingly, $V_{\alpha,0} = 0$, and $V_{\alpha,1} = q$. The increasing property in q holds. Assume $V_{\alpha,n-1}(q, h, a, e_b, e)$ is increasing in q. Fix (h, a, e_b, e), in the state $(q + 1, h, a, e_b, e)$, the set of feasible u is $\{0, 1, \cdots, q + 1\}$ whereas it is $\{0, 1, \cdots, q\}$ for state (q, h, a, e_b, e). Consider state $(q + 1, h, a, e_b, e)$, let the optimal action be (u^*, η^*) with $u^* \in \{0, 1, \cdots, q\}$, hence $V_{\alpha,n}(q + 1, h, a, e_b, e) = q + 1 + \beta \left[\rho \frac{\sigma^2}{h}(e^{\theta(q+1-u^*)} - 1) + \Delta(q + 1 - u^*) - \frac{e_b-\eta^*}{\tau} \right]^+ + \alpha \mathbb{E}_{h,a,e} \left[V_{\alpha,n-1}(u^* + A, H, A, (\eta^* + E)^-, E) \right]$. As (u^*, η^*) is feasible in state (q, h, a, e_b, e), $V_{\alpha,n}(q, h, a, e_b, e) \leq q + \beta \left[\rho \frac{\sigma^2}{h}(e^{\theta(q-u^*)} - 1) + \Delta(q - u^*) - \frac{e_b-\eta^*}{\tau} \right]^+ + \alpha \mathbb{E}_{h,a,e} \left[V_{\alpha,n-1}(u^* + A, H, A, (\eta^* + E)^-, E) \right] \leq V_{\alpha,n}(q + 1, h, a, e_b, e)$. If (u^*, η^*) with $u^* = q + 1$,

$$
V_{\alpha,n}(q + 1, h, a, e_b, e) = q + 1 + \alpha \mathbb{E}_{h,a,e} \left[V_{\alpha,n-1}(q + 1 + A, H, A, (\eta^* + E)^-, E) \right]. \tag{27}
$$

Meanwhile, since (q, η^*) is feasible in state (q, h, a, e_b, e),

$$
V_{\alpha,n}(q, h, a, e_b, e) \leq q + \alpha \mathbb{E}_{h,a,e} \left[V_{\alpha,n-1}(q + A, H, A, (\eta^* + E)^-, E) \right] \overset{(a)}{\leq} V_{\alpha,n}(q + 1, h, a, e_b, e),
$$

where (a) holds since the induction hypothesis. ∎

Property 2. $V_\alpha(q, h, a, e_b, e)$ is a non-increasing function of e_b.

Proof: We verify this by induction. According to (26), $V_{\alpha,0} = 0$, and then $V_{\alpha,1} = q$. The non-increasing property holds. Assume $V_{\alpha,n-1}(q, h, a, e_b, e)$ is non-increasing in e_b. Given

(q, h, a, e), consider state (q, h, a, e_b, e), let (u^*, η^*) be the optimal policy, i.e., $V_{\alpha,n}(q, h, a, e_b, e) = q + \beta \left[\rho \frac{\sigma^2}{h}(e^{\theta(q-u^*)} - 1) + \Delta(q-u^*) - (e_b - \eta^*)/\tau \right]^+ + \alpha \mathbb{E}_{h,a,e} \left[V_{\alpha,n-1}(u^* + A, H, A, (\eta^* + E)^-, E) \right]$.

For state $(q, h, a, e_b + 1, e)$, (u^*, η^*) is feasible, then we have $V_{\alpha,n}(q, h, a, e_b + 1, e) \leq q + \beta \left[\rho \frac{\sigma^2}{h}(e^{\theta(q-u^*)} - 1) + \Delta(q-u^*) - (e_b + 1 - \eta^*)/\tau \right]^+ + \alpha \mathbb{E}_{h,a,e} \left[V_{\alpha,n-1}(u^* + A, H, A, (\eta^* + E)^-, E) \right] \leq V_{\alpha,n}(q, h, a, e_b, e)$. ■

In the practical case, the allocated harvested power will not surpass the required total power. Thus, we assume the (u, η) always guarantees that

$$\rho \frac{\sigma^2}{h}(e^{\theta(q-u)} - 1) + \Delta(q - u) \geq \frac{e_b - \eta}{\tau}. \tag{28}$$

Based on this assumption, $P_{grid}(x, r, w) = P(x, r) - w$. The following property gives the joint convexity of $V_\alpha(q, h, a, e_b, e)$ in (q, e_b).

Property 3. $V_\alpha(q, h, a, e_b, e)$ is convex in (q, e_b).

Proof: First, we prove the following claim.

Claim 1. For $\phi \in (0, 1)$ and $\forall x_1, x_2, y$, $\phi \min\{x_1, y\} + (1 - \phi) \min\{x_2, y\} \leq \min\{\phi x_1 + (1 - \phi)x_2, y\}$.

Proof: The claim can be proved by considering $\min\{x_1, x_2\} > y$, $\max\{x_1, x_2\} < y$, and $\min\{x_1, x_2\} \leq y \leq \max\{x_1, x_2\}$, respectively. ■

The convexity is proved by induction. For $n = 0$, $V_{\alpha,0} = 0$, and it is convex. Assume $V_{\alpha,n-1}(q, h, a, e_b, e)$ is convex in (q, e_b). Fix (q, h, a, e_b, e), let (u_1, η_1) and (u_2, η_2) be the optimal

policy for (q_1, e_{b1}) and (q_2, e_{b2}). Then, we get

$$\phi V_{\alpha,n}(q_1, h, a, e_{b1}, e) + (1-\phi)V_{\alpha,n}(q_2, h, a, e_{b2}, e) = \phi\Big[q_1 + \beta(\rho\frac{\sigma^2}{h}(e^{\theta(q_1-u_1)} - 1)$$

$$+ \quad \Delta(q_1 - u_1) - \frac{e_{b1} - \eta_1}{\tau})\Big] + (1-\phi)[q_2 + \beta(\rho\frac{\sigma^2}{h}(e^{\theta(q_2-u_2)} - 1) + \Delta(q_2 - u_2) - \frac{e_{b2} - \eta_2}{\tau})]$$

$$+ \quad \alpha\mathbb{E}_{h,a,e}\Big[\phi V_{\alpha,n-1}(u_1 + A, H, A, (\eta_1 + E)^-, E) + (1-\phi)V_{\alpha,n-1}(u_2 + A, H, A, (\eta_2 + E)^-, E)\Big]$$

$$\overset{(b)}{\geq} \quad \phi q_1 + (1-\phi)q_2 + \beta\Big[\rho\frac{\sigma^2}{h}(e^{\theta[\phi(q_1-u_1)+(1-\phi)(q_2-u_2)]} - 1) + \Delta(\phi(q_1 - u_1) + (1-\phi)(q_2 - u_2))$$

$$- \quad \frac{1}{\tau}(\phi(e_{b1} - \eta_1) + (1-\phi)(e_{b2} - \eta_2))\Big] + \alpha\mathbb{E}_{h,a,e}\Big[V_{\alpha,n-1}(\phi u_1 + (1-\phi)u_2 + A, H, A, \phi(\eta_1 + E)^-$$

$$+ \quad (1-\phi)(\eta_2 + E)^-, E) \overset{(c)}{\geq} \phi q_1 + (1-\phi)q_2 + \beta\Big[\rho\frac{\sigma^2}{h}(e^{\theta[\phi(q_1-u_1)+(1-\phi)(q_2-u_2)]} - 1) + \Delta(\phi(q_1 - u_1)$$

$$+ \quad (1-\phi)(q_2 - u_2)) - \frac{1}{\tau}(\phi(e_{b1} - \eta_1) + (1-\phi)(e_{b2} - \eta_2))\Big] + \alpha\mathbb{E}_{h,a,e}\Big[V_{\alpha,n-1}(\phi u_1 + (1-\phi)u_2$$

$$+ \quad A, H, A, (\phi\eta_1 + (1-\phi)\eta_2 + E)^-, E) \overset{(d)}{\geq} V_{\alpha,n}(\phi q_1 + (1-\phi)q_2, h, a, \phi e_{b1} + (1-\phi)e_{b2}, e),$$

where (b) holds because of the convexity of $e^{\theta(q-u)}+\Delta(q-u)$ (with respect to u) and $V_{\alpha,n-1}(q, h, a, e_b, e)$, (c) holds because of Claim 1 as well as Property 2, and (d) holds since $(\phi u_1 + (1-\phi)u_2, \phi\eta_1 + (1-\phi)\eta_2)$ is feasible for $\phi(q_1, h, a, e_{b1}, e) + (1-\phi)(q_2, h, a, e_{b2}, e)$. The proof completes. ∎

2) On the discount optimal policy: For a state-action pair $(x = (q, h, a, e_b, e), (r, w)) \in \mathcal{X} \times \mathcal{A}(x)$, define $u := q - r$ and $\eta := e_b - w\tau$, i.e., let u and η denote the remaining data in the buffer and the remaining energy in the battery, respectively. Then $(u(x), \eta(x))$ can also define a stationary policy. We can analysis the policy in terms of the remaining data in the buffer u and the remaining energy in the battery η.

Proposition 1. Denote the discount optimal policy in state $x = (q, h, a, e_b, e)$ as $(u^*(x), \eta^*(x))$. Then, $(u^*(x), \eta^*(x))$ satisfies the following inequality array

$$Z_1(q, u^*, h, a, \eta^*, e) \leq \beta\rho\frac{\sigma^2}{h}e^{\theta q}(e^\theta - 1) \leq Z_1(u^* + 1, h, a, \eta^*, e), \tag{29}$$

$$Z_2(u^*, h, a, \eta^*, e) \leq \frac{-\beta}{\tau} \leq Z_2(u^*, h, a, \eta^* + 1, e), \tag{30}$$

$$Z_3(q, u^*, h, a, \eta^*, e) \leq \beta\rho\frac{\sigma^2}{h}e^{\theta q}(e^\theta - 1) \leq Z_3(u^* + 1, h, a, \eta^* + 1, e), \tag{31}$$

where $Z_1(q, u, h, a, \eta, e) = e^{\theta u}\Big[\alpha\mathbb{E}_{h,a,e}\big[G_1(u + A, H, A, (\eta + E)^-, E)\big] + \beta\big[\Delta(q-u) - \Delta(q-u+1)\big]\Big]$ with $G_1(q, h, a, e_b, e) = V_\alpha(q, h, a, e_b, e) - V_\alpha(q - 1, h, a, e_b, e)$ being the partial backward difference of V_α regarding q. $Z_2(u, h, a, \eta, e) = \alpha\mathbb{E}_{h,a,e}\big[G_2(u + A, H, A, (\eta + E)^-, E)\big]$ with $G_2(q, h, a, e_b, e) = V_\alpha(q, h, a, e_b, e) - V_\alpha(q, h, a, e_b - 1, e)$ being the partial backward difference of V_α regarding e_b. $Z_3(q, u, h, a, \eta, e) = e^{\theta u}\Big[\alpha\mathbb{E}_{h,a,e}\big[G_{12}(u + A, H, A, (\eta + E)^-, E)\big] + \beta\big[\Delta(q - u) - \Delta(q - u + 1)\big] + \frac{\beta}{\tau}\Big]$ with $G_{12}(q, h, a, e_b, e) = V_\alpha(q, h, a, e_b, e) - V_\alpha(q - 1, h, a, e_b - 1, e)$ being the backward difference of V_α regarding (q, e_b).

Proof: First, the discounted cost optimality equation becomes

$$V_\alpha(q, h, a, e_b, e) = \min_{u \in \{0,1,\cdots,q\}, \eta \in \{0,1,\cdots,e_b\}} \left\{ q + \beta\Big[\rho\frac{\sigma^2}{h}(e^{\theta(q-u)} - 1) + \Delta(q - u) - \frac{e_b - \eta}{\tau}\Big]^+ \right.$$
$$\left. + \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(u + A, H, A, (\eta + E)^-, E)\big] \right\}, \tag{32}$$

Let $S(u, \eta) = q + \beta\Big[\rho\frac{\sigma^2}{h}(e^{\theta(q-u)} - 1) + \Delta(q-u) - \frac{e_b - \eta}{\tau}\Big] + \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(u + A, H, A, (\eta + E)^-, E)\big]$. First, we have

$$S(u + 1, \eta) - S(u, \eta) = \beta\rho\frac{\sigma^2}{h}(e^{\theta(q-u-1)} - e^{\theta(q-u)}) + \beta[\Delta(q - u - 1) - \Delta(q - u)]$$
$$+ \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(u + 1 + A, H, A, (\eta + E)^-, E) - V_\alpha(u + A, H, A, (\eta + E)^-, E)\big] \tag{33}$$

and

$$S(u - 1, \eta) - S(u, \eta) = \beta\rho\frac{\sigma^2}{h}(e^{\theta(q-u+1)} - e^{\theta(q-u)}) + \beta[\Delta(q - u + 1) - \Delta(q - u)]$$
$$+ \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(u - 1 + A, H, A, (\eta + E)^-, E) - V_\alpha(u + A, H, A, (\eta + E)^-, E)\big]. \tag{34}$$

Then applying $S(u^* + 1, \eta^*) - S(u^*, \eta^*) \geq 0$ and $S(u^* - 1, \eta^*) - S(u^*, \eta^*) \geq 0$, we obtain (29). Similarly, as $S(u, \eta + 1) - S(u, \eta) = \frac{\beta}{\tau} + \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(u + A, H, A, (\eta + 1 + E)^-, E) - V_\alpha(u + A, H, A, (\eta + E)^-, E)\big]$ and $S(u, \eta - 1) - S(u, \eta) = \frac{-\beta}{\tau} + \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(u + A, H, A, (\eta - 1 + E)^-, E) - V_\alpha(u + A, H, A, (\eta + E)^-, E)\big]$, we can reach (30) from $S(u^*, \eta^* + 1) - S(u^*, \eta^*) \geq 0$

and $S(u^*, \eta^* - 1) - S(u^*, \eta^*) \geq 0$. In addition,

$$S(u+1, \eta+1) - S(u, \eta) = \beta \rho \frac{\sigma^2}{h}(e^{\theta(q-u-1)} - e^{\theta(q-u)}) + \frac{\beta}{\tau} + \beta[\Delta(q-u-1) - \Delta(q-u)]$$

$$+ \quad \alpha \mathbb{E}_{h,a,e}\big[V_\alpha(u+1+A, H, A, (\eta+1+E)^-, E) - V_\alpha(u+A, H, A, (\eta+E)^-, E)\big] \quad (35)$$

and

$$S(u-1, \eta-1) - S(u, \eta) = \beta \rho \frac{\sigma^2}{h}(e^{\theta(q-u+1)} - e^{\theta(q-u)}) - \frac{\beta}{\tau} + \beta[\Delta(q-u+1) - \Delta(q-u)]$$

$$+ \quad \alpha \mathbb{E}_{h,a,e}\big[V_\alpha(u-1+A, H, A, (\eta-1+E)^-, E) - V_\alpha(u+A, H, A, (\eta+E)^-, E)\big]. \quad (36)$$

Then, (31) can be obtained by applying $S(u^*-1, \eta^*-1) - S(u^*, \eta^*) \geq 0$ and $S(u^*+1, \eta^*+1) - S(u^*, \eta^*) \geq 0$. ∎

Remark: When (u^*, η^*) is on the boundary of the feasible set, corresponding conditions can also be obtained by following the proof of Proposition 1.

Proposition 2. For $x = (q, h, a, e_b, e)$ satisfying

$$Z_1\big(q, 0, h, a, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\}, e\big) > \beta \rho \frac{\sigma^2}{h} e^{\theta q}(e^\theta - 1) \quad (37)$$

and

$$Z_2(0, h, a, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\}, e) > \frac{-\beta}{\tau}, \quad (38)$$

$(0, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\})$ is the discount optimal policy. In addition, for (q, h, a, e_b, e) satisfying

$$Z_1(q, q, h, a, e_b, e) < \beta \rho \frac{\sigma^2}{h} e^{\theta q}(e^\theta - 1) \quad (39)$$

and

$$Z_2(q, h, a, e_b, e) < \frac{-\beta}{\tau}, \quad (40)$$

(q, e_b) is the discount optimal policy.

Proof: Using Property 3 in Appendix C1, we can derive that $Z_1(q, u, h, a, \eta, e) \leq Z_1(q, u+1, h, a, \eta, e)$, $Z_1(q, u, h, a, \eta, e) \leq Z_1(q, u, h, a, \eta+1, e)$, $Z_2(u, h, a, \eta, e) \leq Z_2(u, h, a, \eta+1, e)$,

$Z_2(u, h, a, \eta, e) \leq Z_2(u+1, h, a, \eta, e)$, and $Z_3(q, u, h, a, \eta, e) \leq Z_3(q, u+1, h, a, \eta+1, e)$.[17]
On the other hand, (28) should be satisfied. Thus, given (q, h, a, e), $Z_1(q, 0, h, a, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\}, e)$, $Z_2(0, h, a, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\}, e)$, and $Z_3(q, 0, h, a, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\}, e)$
are the smallest respectively. Following the proof of proposition 1, we can prove the first half
of the proposition by contradiction. Specifically, suppose $(0, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\})$ is not the
optimal solution, then $S(u^* - 1, \eta^*) - S(u^*, \eta^*) \geq 0$ or $S(u^*, \eta^* - 1) - S(u^*, \eta^*) \geq 0$ should
hold. We have $Z_1(q, 0, h, a, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\}, e) < Z_1(q, u^*, h, a, \eta^*, e) \leq \beta\rho\frac{\sigma^2}{h}e^{\theta q}(e^\theta - 1)$
or $Z_2(0, h, a, \tau \max\{0, \frac{e_b}{\tau} - P(x, q)\}, e) < Z_2(u^*, h, a, \eta^*, e) \leq \frac{-\beta}{\tau}$, and the contradiction occurs.

We can prove the second half of the proposition similarly by using contradiction. First,
given (q, h, a, e), $Z_1(q, q, h, a, e_b, e)$ and $Z_2(q, h, a, e_b, e)$ are the largest values of Z_1 and Z_2,
respectively. Assume (q, e_b) is not the optimal solution, then $S(u^* + 1, \eta^*) - S(u^*, \eta^*) \geq 0$ or
$S(u^*, \eta^* + 1) - S(u^*, \eta^*) \geq 0$ should be satisfied. Consequently, we get $Z_1(q, q, h, a, e_b, e) \geq$
$Z_1(q, u^* + 1, h, a, \eta^*, e) \geq \beta\rho\frac{\sigma^2}{h}e^{\theta q}(e^\theta - 1)$ or $Z_2(q, h, a, e_b, e) \geq Z_2(u, h, a, \eta^* + 1, e) \geq \frac{-\beta}{\tau}$. The
contradiction occurs then. ∎

*Remark: In Proposition 1 and Proposition 2, to compute $Z_i(\cdot)$ $i = 1, 2, 3$, we need to compute
$V_\alpha(\cdot)$. It can be obtained by value iteration (26).*

Proposition 3. Denote $x = (q, h, a, e_b, e)$. The discount optimal transmit rate policy $r(x) = q - u^*(x)$ is non-decreasing in q and e_b, respectively; The discount optimal battery energy
allocation policy $w(x) = e_b - \eta^*(x)$ is non-decreasing in q and e_b, respectively.

Proof: First, it is easy to see that $r(x)$ is nondecreasing in e_b and $w(x)$ is non-decreasing
in q. Next, we prove the non-decreasing of $r(x)$ in q by contradiction. Consider two states
$x_1 = (q_1, h, a, e_b, e)$ and $x_2 = (q_2, h, a, e_b, e)$. We write $r(x_1)$ and $r(x_2)$ as $r(q_1)$ and $r(q_2)$ for
brevity. Assume $q_1 < q_2$ but $r(q_1) > r(q_2)$, then $0 \leq r(q_2) < r(q_1) \leq q_1 < q_2$. $r(q_2), w(q_2)$ is

---

[17]It is assumed that $\alpha\mathbb{E}_{h,a,e}\big[G_1(q + A, H, A, (\eta + E)^-, E)\big] - e^{-\theta}\alpha\mathbb{E}_{h,a,e}\big[G_1(q - 1 + A, H, A, (\eta + E)^-, E)\big] \geq \beta C$ and $\alpha\mathbb{E}_{h,a,e}\big[G_{12}(q + A, H, A, (\eta + E)^-, E)\big] - e^{-\theta}\alpha\mathbb{E}_{h,a,e}\big[G_{12}(q - 1 + A, H, A, (\eta + E)^-, E)\big] + \frac{\beta}{\tau}(1 - e^{-\theta}) \geq \beta C$. This assumption can be definitely satisfied when C is small.

feasible in x_1 and $r(q_1), w(q_1)$ is feasible in x_2. Since $r(\cdot)$ and $w(\cdot)$ are optimal, we have

$$
\begin{aligned}
& q_1 + \beta\big[\rho\frac{\sigma^2}{h}(e^{\theta r(q_1)} - 1) + C - w(q_1)\big] \\
& \quad + \quad \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(q_1 - r(q_1) + A, H, A, (e_b - w(q_1)\tau + E)^-, E)\big] \\
& \leq \quad q_1 + \beta\big[\rho\frac{\sigma^2}{h}(e^{\theta r(q_2)} - 1) + \Delta(r(q_2)) - w(q_2)\big] \\
& \quad + \quad \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(q_1 - r(q_2) + A, H, A, (e_b - w(q_2)\tau + E)^-, E)\big]
\end{aligned}
\tag{41}
$$

$$
\begin{aligned}
& q_2 + \beta\big[\rho\frac{\sigma^2}{h}(e^{\theta r(q_2)} - 1) + \Delta(r(q_2)) - w(q_2)\big] \\
& \quad + \quad \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(q_2 - r(q_2) + A, H, A, (e_b - w(q_2)\tau + E)^-, E)\big] \\
& \leq \quad q_2 + \beta\big[\rho\frac{\sigma^2}{h}(e^{\theta r(q_1)} - 1) + C - w(q_1)\big] \\
& \quad + \quad \alpha\mathbb{E}_{h,a,e}\big[V_\alpha(q_2 - r(q_1) + A, H, A, (e_b - w(q_1)\tau + E)^-, E)\big]
\end{aligned}
\tag{42}
$$

Add (41) and (42), we have

$$
\begin{aligned}
& \mathbb{E}_{h,a,e}\big[V_\alpha(q_1 - r(q_2) + A, H, A, (e_b - w(q_2)\tau + E)^-, E)\big] \\
& \quad - \quad \mathbb{E}_{h,a,e}\big[V_\alpha(q_1 - r(q_1) + A, H, A, (e_b - w(q_1)\tau + E)^-, E)\big] \\
& \quad > \quad \mathbb{E}_{h,a,e}\big[V_\alpha(q_2 - r(q_2) + A, H, A, (e_b - w(q_2)\tau + E)^-, E)\big] \\
& \quad - \quad \mathbb{E}_{h,a,e}\big[V_\alpha(q_2 - r(q_1) + A, H, A, (e_b - w(q_1)\tau + E)^-, E)\big]
\end{aligned}
\tag{43}
$$

As $V_\alpha(q, h, a, e_b, e)$ is convex in (q, e_b), $\mathbb{E}_{h,a,e}\big[V_\alpha(y + A, H, A, (z + E)^-, E)\big]$ is convex in (y, z). (43) contradicts the convexity. Then we prove the non-decreasing of $r(x)$ in q. The non-decreasing of $w(x)$ in e_b can be verified similarly. ∎

D. Proof of Lemma 8

The lemma can be proved intuitively as follows. Given a transmission rate, the required power is known from the inverse of (1). Out of this power, as much as possible shall be supplied by the battery, since battery energy is "free". In other words, any policy that draws power from the grid while energy is still available in the battery cannot outperform an equivalent one which

strictly uses battery energy first, that has the same total power.

E. Proof of Lemma 9

Since $r(q, h, a, e_b, e)$ is irrelevant to e_b, given rate policy $r(q, h, a)$, the rate is determined independent of the battery allocation in each timeslot. Then greedy battery allocation is optimal for one-step cost in each timeslot according to lemma 8. Thus, the greedy battery allocation policy is the optimal for (10).

F. Proof of Lemma 6

Since the optimal policy of UP_β is g_β, we have

$$J^{g_\beta}(\beta + \lambda) - J^{g_\beta}(\beta) \geq J^{g_{\beta+\lambda}}(\beta + \lambda) - J^{g_\beta}(\beta) \geq J^{g_{\beta+\lambda}}(\beta + \lambda) - J^{g_{\beta+\lambda}}(\beta) \tag{44}$$

for any positive $\beta > 0$ and $\lambda > 0$. Thus,

$$\lambda K^\beta \geq J^{g_{\beta+\lambda}}(\beta + \lambda) - J^{g_\beta}(\beta) \geq \lambda K^{\beta+\lambda} > 0. \tag{45}$$

The monotonicity of $J^{g_\beta}(\beta)$ and K^{g_β} with respect to β are verified. In the following, we prove the non-decreasing of B^{g_β} in β. First, similarly as in [18], we can prove that $u^*(x)$ is non-decreasing in β. Next, as $A[n]$ is an independent process, then using (3), we claim that B^{g_β} is also non-decreasing in β.

G. Proof of Lemma 10

We can verify the lemma through (22) together with Lemma 2. When $\beta \gg 1$, we have $\beta \gg \alpha$. Then $V_\alpha = \min_{r \in \{0, 1, \cdots, q\}, w \in \{0, \frac{1}{\tau}, \cdots, \frac{e_b}{\tau}\}} \left\{ q + \beta \left[\rho \frac{\sigma^2}{h} (e^{\theta r} - 1) + \Delta(r) - w \right]^+ \right\}$. Given rate $r(x)$, we have $w(x) = \min\{\frac{e_b}{\tau}, P(x, r)\}$ (i.e., greedy policy) is discount optimal for state x. When β is sufficient small, we have $\beta \ll \alpha$. Thus, $V_\alpha = \min_{r \in \{0, 1, \cdots, q\}, w \in \{0, \frac{1}{\tau}, \cdots, \frac{e_b}{\tau}\}} \left\{ q + \alpha \mathbb{E}_{h, a, e} \left[V_\alpha(q - r + A, H, A, (e_b - w\tau + E)^-, E) \right] \right\}$. Using Property 2 in Appendix C1, $w = 0$ is discount optimal. Since limitation will not change the partial order, utilizing the second half of Lemma 2, we reach the lemma.

H. Proof of Lemma 12

Since the constrained MDP (8) is equivalent to UP_{β_0} or the constrained optimal policy is a mixed policy of optimal policies for UP_{β^+} and UP_{β^-}. When $\beta_0 \gg 1$ or $\beta^- \gg 1$, according to the first half of Lemma 10, we can derive the greedy policy is the optimal battery power allocation policy under given rate policy. Fix the greedy policy as the battery power allocation policy in (8), we arrive at (19) for solving the optimal rate policy.

REFERENCES

[1] T. Zhang, W. Chen, Z. Han, and Z. Cao, "A cross-layer perspective on energy harvesting aided green communications over fading channel," *Proc. the 2nd IEEE INFOCOM Workshop on Communications and Control for Smart Energy Systems (CCSES 2013)*, Turin, Italy, Apr. 2013.

[2] G. Miao, N. Himayat, Y. (Geoffrey) Li, and A. Swami, "Cross-layer optimization for energy-efficient wireless communications: a survey," *Wirel. Commun. Mob. Comput.,* vol. 9, no. 4, pp. 529-542, Apr. 2009.

[3] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C.-X. Wang, T. A. Le, M. R. Nakhai, J. Zhang, and L. Hanzo, "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.,* vol 49, no. 6, pp. 46-54, Jun. 2011.

[4] Y. Chen, S. Zhang, S. Xu, and Y. (Geoffrey) Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.,* vol 49, no. 6, pp. 30-37, Jun. 2011.

[5] B. Wang, Y. Wu, F. Han, Y.-H. Yang, and K. J. R. Liu, "Green wireless communications: A time-reversal paradigm," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1698- 1710, Sep. 2011.

[6] G. Auer, V. Giannini, C. Desset, I. Gdor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?," *IEEE Wireless Commun.*, vol. 10, no. 5, pp. 40-49, Oct. 2011.

[7] S. Cui, A. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.,* vol. 4, no.5, pp. 2349-2360, Sep. 2005.

[8] D. Niyato, E. Hossain, M. Rashid, and V. Bhargava, "Wireless sensor networks with energy harvesting technologies: a game-theoretic approach to optimal energy management," *IEEE Wireless Commun.,* vol. 14, no. 4, pp. 90-96, Aug. 2007.

[9] V. Sharma, U. Mukherji, V. Joseph, and S. Gupta, "Optimal energy management policies for energy harvesting sensor nodes," *IEEE Trans. Wireless Commun.,* vol. 9, no. 4, pp. 1326-1336, Aug. 2010.

[10] R. Rajesh, V. Sharma, and P. Viswanath, "Information capacity of energy harvesting sensor nodes," *Proc. IEEE ISIT'11*, Saint-Petersburg, Russia, Jul. 31-Aug. 5, 2011.

[11] N. Michelusi, K. Stamatiou, and M. Zorzi, "On optimal transmission policies for energy harvesting devices," *Proc. Information Theory and Applications Workshop 2012 (ITA'12)*, San Diego, CA, Feb. 2012.

[12] Z. Wang, A. Tajer, and X. Wang, "Communication of energy harvesting tags," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 1159-1166, Apr. 2012.

[13] C. K. Ho and R. Zhang, "Optimal energy allocation for wireless communications with energy harvesting constraints," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4808-4818, Sep. 2012.

[14] J. Xu and R. Zhang, "Throughput optimal policies for energy harvesting wireless transmitters with non-ideal circuit power," *IEEE J. Sel. Areas Commun.*, accepted for publiation, 2013.

[15] Z. Han and K. J. R. Liu, *Resource Allocation for Wireless Networks: Basics, Techniques, and Applications*, NewYork, NY: Cambridge University Press, 2008.

[16] R. Berry, "Power and delay trade-offs in fading channels," Ph.D. dissertation, MIT, Cambridge, MA, Jun. 2000.

[17] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inform. Theory,*, vol. 48, no. 5, pp. 1135-1149, May 2002.

[18] M. Goyal, A. Kumar, and V. Sharma, "Optimal cross-layer scheduling of transmissions over a fading multi-access channel," *IEEE Trans. Inform. Theory,* vol. 54, no. 8, pp. 3518-3537, Aug. 2008.

[19] B. Ata, "Dynamic power control in a wireless static channel subject to a quality-of-service constraint," *Oper. Res.,* vol. 53, no. 5, pp. 842-851, Sep.-Oct. 2005.

[20] E. Altman, *Constrained Markov decision processes,* London/Boca Raton: Chapman & Hall/CRC Press, 1999.

[21] E. A. Feinberg and A. Shwartz, *Handbook of Markov Decision Processes: Methods and Applications,* edited, Boston: Kluwer Academic Publishers, 2002.

[22] D. J. Ma, A. M. Makowski, and A. Shwartz, "Estimation and optimal control for constrained Markov chains," *Proc. IEEE Conf. Decision and Control,* Athens, Greece, Dec. 1986.

[23] F. J. Beutlerand and K. W. Ross, "Optimal policies for controlled markov chains with a constraint," *Journal of Mathematical Analysis and Applicationsm*, vol. 112, no. 1, pp. 236-252, 1985

[24] L. I. Sennott, "Constrained average cost Markov decision chains," *Probability in the Engineering and Informational Sciences*, vol. 7, no. 1, pp. 69-83, Jan. 1993

[25] D. V. Djonin and V. Krishnamurthy, "Structural results on optimal transmission scheduling over dynamical fading channels: A constrained markov decision process approach," *The IMA Volumes in Mathematics and its Applications*, vol. 143, pp 75-98, 2007.

[26] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynammic Programming*. Hoboken, New Jersey: John Wiley & Sons, 1994.

[27] M. Schal, "Average optimality in dynamic programming with general state space," *Math. Oper. Res.*, vol. 18, no. 1, pp. 163-172, Feb. 1993.

[28] O. H. Lerma and J. B. Lassere, *Discrete-Time Markov Control Processes: Basic Optimality Criteria,* New York: Springer Verlag, 1996.

[29] A. Wald, "Some generalizations of the theory of cumulative sums of random variables," *The Ann. Math. Statist.*, vol. 16, no. 3, pp. 287-293, Sep. 1945.