2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)

Q-Learning Based Energy Harvesting for Heterogeneous Statistical QoS Provisioning Over Multihop Big-Data Relay Networks

Xi Zhang, Jingqing Wang, and Qixuan Zhu

Networking and Information Systems Laboratory
Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
E-mail: {*xizhang@ece.tamu.edu*, *wang12078@tamu.edu*, *qixuan@tamu.edu*,}

***Abstract*—With the increasing demand for the data-intensive wireless multimedia services over the time-varying wireless channels, the big-data based wireless networks demand the 5G candidate framework to process such massive amount of multimedia data without causing extra burden to the backhaul links in supporting the heterogeneous statistical delay-bounded quality-of-service (QoS) provisionings. Due to the benefits of energy harvesting (EH) technologies, wireless devices are able to support the data-intensive wireless multimedia services by harvesting energy from the environment. Energy harvesting has emerged as the promising technology to solve the energy supply problem while bringing new challenges due to the stochastic nature of the harvested energy in supporting the heterogeneous statistical quality-of-service (QoS) provisionings. However, due to the unknown dynamics of the harvested energy as well as the channel state information (CSI), it is challenging to design the efficient routing protocol for selecting the optimal routing and power allocation policies under the statistical delay-bounded QoS constraints. To overcome the aforementioned problems, in this paper we propose the Q-learning based optimal routing and power allocation policies through learning from the history of the energy harvesting process while satisfying the heterogeneous statistical delay-bounded QoS constraints over multihop big-data relay networks. In particular, under the heterogeneous statistical delay-bounded QoS requirements, we formulate the end-to-end effective-capacity optimization problem for the battery-free energy harvesting based big-data multihop relay networks. Then, we apply the Markov decision process and Q-learning methods for deriving the optimal multihop routing algorithms over big-data multihop relay networks. Also conducted is a set of simulations which evaluate the system performances and show that our proposed Q-learning based multihop routing scheme outperforms the other existing schemes under the heterogeneous statistical delay-bounded QoS constraints over multihop big-data relay networks.**

***Index Terms*—Q-learning, energy harvesting, heterogeneous statistical delay-bounded quality of service (QoS), effective capacity, multihop big-data relay networks.**

I. Introduction

In order to support the explosively growth of the multimedia data-intensive wireless services and applications, the energy supply problem has received a great deal of research attention from both academia and industry. Consequently, in order to efficiently manage the energy consumption and prolong the lifetime of the wireless devices, researchers have proposed various 5G candidate techniques to support the current demand of the real-time multimedia big-data transmissions for guaranteeing the statistical delay-bounded quality-of-service (QoS) [1] [2] provisionings.

To relieve the energy consumption of the network operators over big-data relay networks, a promising solution has been proposed by applying energy harvesting (EH) devices at base stations and utilize clean and renewable energy (such as solar, thermal, RF radiation etc.) as alternative energy resources. Energy harvesting scheme, which enables the wireless devices to harvest energy from the environment, has been proposed as one of the 5G promising candidate techniques to tackle such energy supply problem. In this case, each wireless device is equipped with one or more energy harvesters, as well as an energy buffer, storing the harvested energy for future use. However, due to dynamic nature of the energy sources, it is challenging to adapt the transmit power of the wireless devices while guaranteeing the statistical delay-bounded QoS requirements, in order to maximize the system capacity. Furthermore, the resources may be wasted and the service will be suspended in the absence of the prior knowledge of the system statistics, accordingly, how to learn from the historical behaviors and efficiently allocate the available energy is a significant concern for network operation and optimization problems.

Toward this end, various machine-learning based optimal routing protocols have been proposed to learn from the historical behaviors by using game-theoretic approach, differentially private online learning, Q-learning method, and so on. According to the machine learning theory, there is a network controller who knows about the amount of traffics associated with BSs in the current network state, updates the network state according to a controlled discrete-time Markov decision process (DTMDP), and selects the optimal traffic offloading strategy that can maximize the total reward while satisfying the statistical delay-bounded QoS requirements. The works of [3] formulated the joint optimization problem by combining different perspectives

This work was supported in part by the U.S. National Science Foundation under Grants ECCS-1408601 and CNS-1205726, and the U.S. Air Force under Grant FA9453-15-C-0423.

978-1-7281-2980-8/19/$31.00 ©2019 IEEE
DOI 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00147

from Markov approximation and non-cooperative game theory. The authors apply the log-linear learning algorithm to find the equilibrium of a non-cooperative game. However, when considering the real-time multimedia data transmissions for the coordination among the competing interests of a large number of mobile users, how to efficiently design the machine-learning based multihop routing algorithms for the EH based data transmissions under the statistical delay-bounded QoS constraints still remains as a challenging and open problem.

To effectively overcome the aforementioned problems, in this paper we propose the Q-learning based multihop routing scheme for designing the optimal routing and power allocation policies through learning from the history of the EH process while satisfying the heterogeneous statistical delay-bounded QoS constraints over multihop big-data relay networks. In particular, we establish the wireless communication model and EH model over multihop big-data relay networks. Under the heterogeneous statistical delay-bounded QoS provisionings, we formulate the end-to-end effective-capacity optimization problem for the battery-free EH based multihop relay scheme. Then, we develop the Markov decision process and Q-learning based multihop routing and power allocation algorithms. We also conduct a set of simulations which evaluate the system performances and show that our proposed Q-learning based routing scheme outperforms the other existing schemes under the heterogeneous statistical delay-bounded QoS constraints over multihop big-data relay networks.

The rest of this paper is organized as follows. Section II establishes the system models for wireless communications and EH. Section III formulates the end-to-end effective capacity optimization problem under the heterogeneous statistical delay-bounded QoS constraints. Section IV proposes the Markov decision process based multihop routing algorithm. Section V proposes the Q-learning based multihop routing algorithm. Section VI evaluates and compares the performances of our proposed schemes with the other existing schemes. The paper concludes with Section VII.

II. The System Models

Consider an energy harvesting (EH) based big-data multihop relay network where the source node transmits its data to the destination node via multiple decode-and-forward (DF) relays, as shown in Fig. 1. Assume that there are K users and one base station (BS) that can transmit energy to all the users. Assume that both data and energy arrive in packets at each time slot. Define $T_{\max}$ as the maximum delay that can be tolerated by at the destination mobile user.

A. The Wireless Communication System Model

The Nakagami-*m* fading model, where m is the shape factor of the Nakagami-*m* model, is applied in our proposed system. In the special case, $m = 1$ represents Rayleigh fading, and $m = \infty$ corresponds to the Gaussian channel. The channel's impulse response function, denoted by $h_k(t)$, from mobile user k to mobile user $(k+1)$ at time slot t can be expressed as

$$h_k(t) = \alpha_k \delta\left(t - \xi_k\right) \exp\left(-\jmath \phi_k\right), \tag{1}$$

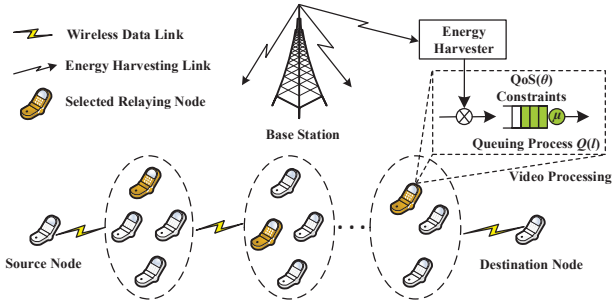

Fig. 1. The system architecture for the energy harvesting based big-data multihop relay network.

where $\jmath = \sqrt{-1}$; ξ_k is the path delay for mobile user k; $\delta(\cdot)$ is the unit impulse function; α_k is a random variable representing the path envelope for mobile user k's channel and follows the Nakagami-*m* distribution; and ϕ_k is a random variable representing the phase-shift for mobile user k. Note that all random variables $\phi_k, \forall k$, are independent and identically distributed (i.i.d) which uniformly distributed between $[0, 2\pi)$, and all random variables $\alpha_k, \forall k$, are i.i.d.

Correspondingly, the received signal, denoted by $y_k(t)$, from mobile user k to mobile user $(k+1)$ at time slot t can be expressed as in the following equation:

$$y_k(t) = \sqrt{\mathcal{P}_k(t)} h_k(t) s_k(t) + \sum_{j=1, j\neq k}^{K} \sqrt{\mathcal{P}_j(t)} h_j(t) s_j(t) + n_k(t), \tag{2}$$

where $s_k(t)$ and $s_j(t)$ denote the source signals sent from mobile user k and mobile user j at time slot t, respectively; $\mathcal{P}_k(t)$ and $\mathcal{P}_j(t)$ are the transmit power at mobile user k and mobile user j, respectively; $h_k(t)$ and $h_j(t)$ denote the channel's impulse response from mobile user k to mobile user $(k+1)$ and mobile user j to mobile user $(k+1)$ at time slot t, respectively; and $n_k(t)$ represents the additive white Gaussian noise (AWGN) with zero mean and variance σ^2.

Consider the multihop relay network, where the required file is transmitted through multiple relay nodes to the destination. Define multihop routing indicator variable at time slot t, denoted by $x_k(t)$, subject to the following constraints:

$$\begin{cases} x_k(t) = 1, \text{ mobile user } k \text{ is selected for next hop;} \\ x_k(t) = 0, \text{ otherwise.} \end{cases} \tag{3}$$

B. The Energy Harvesting System Model

We apply the radio-frequency (RF) based EH technique in light of its flexible and sustainable characteristics compared with the conventional solar and wind EH techniques. Assume that each user is equipped with one single antenna and operates in half-duplex mode such that it can only transmit or receive at one time. In this way, the users work under the "harvest-then-transmit" protocol [4], i.e., there are two phases in a frame duration T_f:

1) Phase 1: Energy Harvesting: During the energy harvesting phase, the RF energy harvesting source will transmit energy to user k. The duration of Phase 1 is $\tau_k T_f$ where $0 < \tau_k < 1$.

2) Phase 2: Wireless Data Transmission: For the wireless data transmission phase, user k will transmit data using the energy harvested in *Phase 1* in the remaining time duration $(1-\tau_k)T_f$.

For the proposed EH model, we can ignore the noise energy since it is too small as compared with the overall harvested energy. We assume that all mobile users are equipped with energy harvester that can harvest energy from RF signals. Accordingly, the received power, denoted by E_k, at user k can be formulated as follows [5]:

$$E_k = \kappa \tau_k \mathcal{P}_t g_{E,k}, \tag{4}$$

where κ is the EH efficiency decided by the energy harvester, $\mathcal{P}_t$ denotes the transmit power at the energy harvester, and $g_{E,k}$ represents the channel's impulse response from the power transmitter to user k.

Assume that all users are battery-free, i.e., the users are not equipped with any constant energy supplies, but they can harvest energy from the RF signals. Thus, the users' operations completely depend on the amount of energy harvested from the RF signals. Correspondingly, the transmit power need to satisfy the following constraint:

$$\mathbb{E}_{\gamma_k}\left[(1-\tau_k)\mathcal{P}_k\right] \leq \kappa \tau_k T_f \mathcal{P}_t g_{E,k}. \tag{5}$$

III. Maximizing Effective Capacity Under Heterogeneous Statistical Delay-Bounded QoS Provisionings

In this section, we consider the battery-free scenario for our proposed EH scheme. By jointly optimizing the EH time and the transmit power, we formulate the effective-capacity maximization problem for our proposed EH scheme under the heterogeneous statistical delay-bounded QoS constraints over multihop big-data relay networks.

A. Preliminary for Effective Capacity Under Homogeneous Statistical Delay-Bounded QoS Constraints

The statistical QoS guarantees [6] has been extensively studied for analyzing the queuing behavior for time-varying arrival and service processes. Based on large deviation principle (LDP), under sufficient conditions, the queue length process $Q(t)$ with the average arrival rate $\widetilde{\lambda}$ and average service rate $\widetilde{\mu}$ converges in distribution to a random variable $Q(\infty)$ such that [7]

$$-\lim_{Q_{th}\to\infty} \frac{\log\left(\Pr\left\{Q(\infty) > Q_{th}\right\}\right)}{Q_{th}} = \theta. \tag{6}$$

where $\theta > 0$ is defined as the QoS exponent and plays a critically important role for statistical delay-bounded QoS provisionings and Q_{th} denotes the queue-length bound.

The *effective capacity* [8] is defined as the maximum constant arrival rate for a given service process subject to the statistical delay-bounded QoS constraints. We can derive the effective capacity, denoted by $C_k(\theta_k, t)$, for the communication between the mobile user k and the mobile user $(k+1)$ at time slot t as follows [8]:

$$C_k(\theta_k, t) = -\frac{1}{\theta_k}\log\left(\mathbb{E}_{\gamma_k}\left[e^{-\theta_k R_k(t)}\right]\right) \tag{7}$$

where $\mathbb{E}_{\gamma_k}(\cdot)$ is the expectation operation with respect to the random variable Γ_k. Γ_k is the mobile user k's signal-to-noise-plus-interference ratio (SINR) whose value is γ_k and θ_k is the QoS exponent for the mobile user k.

Furthermore, we need to derive the power allocation policy, denoted by $\boldsymbol{\nu}_k \triangleq \boldsymbol{\nu}_k(\theta_k, \gamma_k)$, which is a function of *both* the SINR γ_k and the QoS exponent θ_k [9]. Assume that all mobile users are heterogeneous, which implies that they are assigned different power allocations according to different channel state information. Define the mean transmit power of mobile user k as $\overline{\mathcal{P}}_k$. Applying the power allocation policy, the instantaneous transmit power of the user becomes $\mathcal{P}_k(\boldsymbol{\nu}_k) = \boldsymbol{\nu}_k \overline{\mathcal{P}}_k$. As a result, the power allocation policy need to satisfy the mean power constraint given as follows:

$$\int_0^\infty \mathcal{P}_k(\boldsymbol{\nu}_k) p_{\Gamma_k}(\gamma_k) d\gamma_k = \overline{\mathcal{P}}_k, \qquad \forall k, \tag{8}$$

where $p_{\Gamma_k}(\gamma_k)$ denotes the probability density function (pdf) of the random variable Γ_k over a Nakagami-m fading channel given by

$$p_{\Gamma_k}(\gamma_k) = \frac{\gamma_k^{m-1}}{\Gamma(m)}\left(\frac{m}{\overline{\gamma}_k}\right)^m \exp\left(-\frac{m\gamma_k}{\overline{\gamma}_k}\right), \tag{9}$$

where $\overline{\gamma}_k$ denotes the average SINR for mobile user k and $\Gamma(\cdot)$ is the gamma function.

B. Joint Optimization of Power and Time Allocation Under Heterogeneous Statistical Delay-Bounded QoS Guarantees

Considering our proposed EH based multihop relaying model, it is unrealistic to assume that all the different hops have the homogeneous statistical QoS provisionings. Accordingly, the diverse delay-bounded QoS provisionings for different hops need to be considered, which represents the new heterogeneous statistical QoS provisioning framework and imposes many new challenges. Define $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_K]$ as the QoS exponent vector, $\boldsymbol{\tau} = [\tau_1, \tau_2, \ldots, \tau_K]$ as the time scheduling vector, and $\boldsymbol{\mathcal{P}} \triangleq [\mathcal{P}_1, \ldots, \mathcal{P}_K]$ as the power allocation vector for all K users.

We can derive the data transmission rate $R_k(t)$ in Eq. (7) for mobile user k at time slot t as follows:

$$R_k(t) = x_k(t) T_f B \log_2\left(1+\gamma_k\right), \tag{10}$$

where B denotes the bandwidth for the users. We assume that all users are assigned with the same bandwidth B. Using Eq. (10), we can further derive the single-hop effective capacity $C_k(\theta_k, t)$ in Eq. (7) between user k and user $(k+1)$ at time slot t as follows:

$$C_k(\theta_k, t) = -\frac{1}{\theta_k}\log\left(\mathbb{E}_{\gamma_k}\left[\exp\left\{-\theta_k x_k(t)\tau_k T_f \times B\log_2\left(1+\mathcal{P}_k\gamma_k\right)\right\}\right]\right). \tag{11}$$

Since the system capacity of multihop transmissions is dominated by the bottlenecked link, the end-to-end effective capacity of the multihop relay networks can be derived as follows:

$$C(\theta,t) \triangleq \min_{1\le k\le K}\left\{C_k(\theta_k,t)\right\}. \tag{12}$$

Then, using Eqs. (11) and (12), we construct an optimization problem $\mathbf{P_1}$ for maximizing the end-to-end effective capacity under the heterogeneous statistical delay-bounded QoS provisionings can be formulated as follows:

$$\mathbf{P_1}: C^{\text{opt}}(\theta,t) \triangleq \arg\max_{\{\boldsymbol{\mathcal{P}},\boldsymbol{\tau}\}}\left\{\min_{1\le k\le K}\left\{-\frac{1}{\theta_k}\log\left(\mathbb{E}_{\gamma_k}\left[\exp\left\{-\theta_k \times(1-\tau_k)x_k(t)T_f B\log_2\left(1+\mathcal{P}_k\gamma_k\right)\right\}\right]\right)\right\}\right\}, \tag{13}$$

$$\begin{aligned}\text{s.t.}\quad & C1:\ \mathbb{E}_{\gamma_k}\left[(1-\tau_k)\mathcal{P}_k\right]\le \kappa\tau_k T_f\mathcal{P}_t g_{E,k};\\ & C2:\ 0<\tau_k<1,\forall k;\\ & C3:\ \mathcal{P}_k\ge 0,\forall k;\\ & C4:\ x_k(t)\in\{0,1\},\ \forall k,\end{aligned}$$

where $C^{\text{opt}}(\theta,t)$ is the maximized end-to-end effective capacity obtained from the optimization problem $\mathbf{P_1}$. According to Eq. (13), we can observe that the above mixed-integer optimization problem is NP-hard [10] and hence is challenging to solve. Correspondingly, in the next section, the Markov decision process is considered to solve the optimization problem $\mathbf{P_1}$.

IV. Markov Decision Process Based Multihop Routing Algorithm

A. Optimal Multihop Routing Protocol

Our goal is to find the optimal multihop routing protocol $\boldsymbol{\pi}(\boldsymbol{x}(t)) \triangleq [\boldsymbol{\pi}(x_1(t)),\ldots,\boldsymbol{\pi}(x_K(t))]$ in order to maximize the reward function. Accordingly, we define the reward function, denoted by $\mathcal{R}(\boldsymbol{\pi}(\boldsymbol{x}(t)))$, as follows:

$$\mathcal{R}(\boldsymbol{\pi}(\boldsymbol{x}(t))) = \max_{\boldsymbol{\pi}(\boldsymbol{x}(t))}\left\{\mathbb{E}\left[\sum_t \eta_{t-1}\mu(\boldsymbol{\pi}(\boldsymbol{x}(t-1)))\right]\right\}, \tag{14}$$

where $\eta_{t-1}\in[0,1)$ represents the discount factor, $\mu(\boldsymbol{\pi}(\boldsymbol{x}(t-1)))$ is the utility function defined as the maximum end-to-end effective capacity at time slot $(t-1)$ given the total transmission energy $\boldsymbol{\pi}(\boldsymbol{x}(t-1))$, i.e.,

$$\mu(\boldsymbol{\pi}(\boldsymbol{x}(t))) \triangleq \arg\max_{\{\boldsymbol{\mathcal{P}},\boldsymbol{\tau}\}}\left\{\min_{1\le k\le K}\left\{-\frac{1}{\theta_k}\log\left(\mathbb{E}_{\gamma_k}\left[\exp\left\{-\theta_k \times(1-\tau_k)x_k(t)T_f B\log_2\left(1+\mathcal{P}_k\gamma_k\right)\right\}\right]\right)\right\}\right\}, \tag{15}$$

subject to the constraints $C1$-$C4$ of problem $\mathbf{P_1}$ in Eq. (13). Since the mobile users only know their own energy storage state and EH state, the utility function $\mu(\boldsymbol{\pi}(\boldsymbol{x}(t)))$ is a stochastic function, where the statistic nature is introduced by the energy storage state, EH state, and CSI from other mobile users. Assume that each user knows the minimum amount of energy for transmitting the arriving data packet at time slot t. The optimization problem in Eq. (15) can be considered as a discrete-time Markov decision process (DTMDP), consisting of K independent Markov chains.

To solve the above-mentioned optimization problem, first, we define the following four elements:

- Agents: K mobile users.
- State: The system state at time slot t is characterized by the aggregation of channel state information (CSI), and energy state information (ESI), which are denoted by $\boldsymbol{S}(t) = (\boldsymbol{H}(t),\boldsymbol{E}(t))$, where $\boldsymbol{H}(t) = \{h_k(t)\}_{k=1}^K$, and $\boldsymbol{E}(t) = \{E_k(t)\}_{k=1}^K$, respectively.
- Action: At each time slot t, a multihop routing action $\boldsymbol{x}(t) = \{x_k(t)\in\{0,1\}\}_{k=1}^K$ is taken from the set of actions in the action space $\mathcal{A}$ based on the current state $\boldsymbol{S}(t)$. Note that the transmission power allocation action and the routing action are correlated.
- Reward: Reward is defined as the end-to-end effective capacity, specified by Eq. (14).

Given the EH policy $\boldsymbol{\pi}(\boldsymbol{x}(t))$ and the network state $\boldsymbol{S}(t)$, the future network state at time slot $(t+1)$ can be derived using the state transition probability. Since state transitions depend only on the current state and the current action at each mobile user, our proposed model follows the discrete-time Markov decision process. Accordingly, we define the action function as in the following equation:

$$Q^{\pi}(s_k(t),x_k(t)) \triangleq \sum_{s_j(t)\in\mathcal{S}(t)} p_{x_k(t)}(s_j(t),s_k(t))\left[\mu(\boldsymbol{\pi}(\boldsymbol{x}(t))) + V^{\pi}(s_k(t))\right], \tag{16}$$

where $p_{x_k(t)}(s_j(t),s_k(t))$ denotes the transition probability from network state $s_j(t)$ to network state $s_k(t)$ when action $x_k(t)$ is taken at time slot t. Then, the DTMDP based effective-capacity optimization problem can be solved by applying the dynamic programming [11]. Define the Bellman's equation [12] as follows:

$$V^{\pi}(s_k(t)) \triangleq \max_{x_k(t)\in\mathcal{X}} Q^{\pi}(s_k(t),x_k(t)). \tag{17}$$

Define $\mathcal{G}$ as the group of the selected relay nodes for the EH based multihop relay networks. The DTMDP based routing algorithm is proposed in Algorithm 1. The optimal value of the equation $\mathcal{V}(s_k(t))$ can be calculated by working backwards from $T_{\max}$ to 1.

B. Optimal Joint Power and Time Allocation Under Heterogeneous Statistical Delay-Bounded QoS Provisionings

Using Algorithm 1, we can derive the optimal joint power and time allocation policy under the heterogeneous statistical delay-bounded QoS constraint. However, the optimization problem $\mathbf{P_1}$ is a non-convex problem. In order to derive a feasible solution,

Algorithm 1 DTMDP Based Multihop Routing Algorithm Over EH Based Multihop Relay Networks

Input: $T_f, T_{\max}, B, K, \overline{\mathcal{P}}_1, \overline{\mathcal{P}}_2, \ldots, \overline{\mathcal{P}}_K$
Initialization: $[\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K] = [\overline{\mathcal{P}}_1, \overline{\mathcal{P}}_2, \ldots, \overline{\mathcal{P}}_K]$
Step 1:
Set $t = T_{\max}$
while $t \geq 1$ **do**
  for $j = 1 : K$ **do**
    Calculate $Q^{\pi}(s_k(t), x_k(t))$ by using Eq. (16).
    Calculate $V^{\pi}(s_k(t))$.
  end for
  Set $t \leftarrow t - 1$
end while
Step 2:
Set $t = 1$
while $t \leq T_{\max}$ **do**
  for $j = 1 : K$ **do**
    Observe the current state $s_k(t)$ and select the corresponding routing action $x_k(t)$ with the maximum value of $V^{\pi}(s_k(t))$.
    if $x_k(t) = 1$ **then**
      $k \rightarrow \mathcal{G}$
    end if
  end for
  Set $t \leftarrow t + 1$
end while

we define a new variable $e_k \triangleq \tau_k \mathcal{P}_k$. Accordingly, we can convert $\mathbf{P_1}$ into the following optimization problem $\mathbf{P_2}$:

$$\mathbf{P_2}: C^{\text{opt}}(\theta, t) = \arg \max_{\{\boldsymbol{e}, \boldsymbol{\tau}\}} \left\{ \min_{1 \leq k \leq K} \left\{ -\frac{1}{\theta_k} \log \left(\mathbb{E}_{\gamma_k} \left[\exp \left\{ -\theta_k \times (1-\tau_k) T_f B \log_2 \left(1 + \frac{e_k}{\tau_k} \gamma_k \right) \right\} \right] \right) \right\} \right\}, \tag{18}$$

subject to the constraints $C1$-$C4$ of problem $\mathbf{P_1}$ in Eq. (13).

Lemma 1: $\mathbf{P_2}$ is a jointly convex optimization problem with respect to $\boldsymbol{\tau}$ and $\boldsymbol{e}$.

Proof: Due to the lack of space, we omit the proof of *Lemma 1*. ■

According to *Lemma 1*, the non-convex optimization problem $\mathbf{P_1}$ is converted to a convex problem $\mathbf{P_2}$ with respect to time and power allocations. Then, we can further convert $\mathbf{P_2}$ in Eq. (18) into an equivalent problem $\mathbf{P_3}$ as follows:

$$\mathbf{P_3}: C^{\text{opt}}(\theta, t) = \arg \min_{\{\boldsymbol{e}, \boldsymbol{\tau}\}} \left\{ \max_{1 \leq k \leq K} \left\{ \mathbb{E}_{\gamma_k} \left[\exp \left\{ -\theta_k \times (1-\tau_k) T_f B \log_2 \left(1 + \frac{e_k}{\tau_k} \gamma_k \right) \right\} \right] \right\} \right\}, \tag{19}$$

subject to the constraints $C1$-$C4$ of problem $\mathbf{P_1}$ in Eq. (13). For simplicity, we define a new function as follows:

$$\widetilde{C}_k(\tau_k, e_k) \triangleq \mathbb{E}_{\gamma_k} \left[\exp \left\{ -\theta_k (1-\tau_k) T_f B \log_2 \left(1 + \frac{e_k}{\tau_k} \gamma_k \right) \right\} \right], \tag{20}$$

and a new variable $\widetilde{C}$ such that $\widetilde{C}_k(\tau_k, e_k) \leq \widetilde{C}$. Then, we can further reformulate the optimization problem $\mathbf{P_3}$ as follows:

$$\mathbf{P_4}: C^{\text{opt}}(\theta, t) = \arg \min_{\{\boldsymbol{e}, \boldsymbol{\tau}\}} \left\{ \widetilde{C} \right\}, \tag{21}$$

$$\text{s.t.} \quad C0: \ \widetilde{C}_k(\tau_k, e_k) \leq \widetilde{C}; \quad C1, C2, \text{and } C3. \tag{22}$$

Then, we can derive the partial Lagrangian function of $\mathbf{P_4}$ with respect to the constraint $C0$ given in Eq. (22) as follows:

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{e}, \boldsymbol{\tau}) = -\sum_{k=1}^{K} \lambda_k \left(\widetilde{C}_k(\tau_k, e_k) - \widetilde{C} \right), \tag{23}$$

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]$ with λ_k representing the non-negative Lagrangian multiplier associated with the constraint $C0$. Let $\mathcal{F}$ denote the feasible set of $(\boldsymbol{\tau}, \boldsymbol{e})$ given in the constraints $C1$, $C2$, and $C3$. We can determine the Lagrange dual problem $\mathbf{P_4}$ as follows:

$$\mathcal{D}(\boldsymbol{\lambda}) = \arg \min_{\{\boldsymbol{e}, \boldsymbol{\tau}\} \in \mathcal{F}} \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{e}, \boldsymbol{\tau}), \tag{24}$$

We can ignore the term $\lambda_k \widetilde{C}$ in Eq. (23) since it does not affect the optimization with respect to set $(\boldsymbol{\tau}, \boldsymbol{e})$. Therefore, we can rewrite the Lagrange dual problem as follows:

$$\mathbf{P_5}: \ \arg \max_{\{\boldsymbol{e}, \boldsymbol{\tau}\} \in \mathcal{F}} \sum_{k=1}^{K} \lambda_k \widetilde{C}_k(\tau_k, e_k), \tag{25}$$

subject to the constraints $C1$-$C4$ of problem $\mathbf{P_1}$ in Eq. (13).

Using the decomposition technique [13], we can relax the coupling constraints by dual decomposition and formulate the partial Lagrangian function of optimization problem $\mathbf{P_5}$ as follows:

$$\begin{aligned} \widetilde{\mathcal{L}}(\boldsymbol{e}, \boldsymbol{\tau}, \boldsymbol{\mu}) &= \sum_{k=1}^{K} \lambda_k \widetilde{C}_k(\tau_k, e_k) + \frac{\mu_k}{\tau_k} \left\{ \mathbb{E}_{\gamma_k} \left[\left(\frac{1}{\tau_k} - 1 \right) e_k \right] - \kappa \tau_k T_f \mathcal{P}_t g_{E,k} \right\} \\ &= \prod_{k=1}^{K} \lambda_k \mathbb{E}_{\gamma_k} \left[\left(1 + \frac{e_k}{\tau_k} \gamma_k \right)^{-(1-\tau_k)\beta_k} \right] + \frac{\mu_k}{\tau_k} \left\{ \mathbb{E}_{\gamma_k} \left[\left(\frac{1}{\tau_k} - 1 \right) e_k \right] - \kappa T_f \mathcal{P}_t g_{E,k} \right\}, \end{aligned} \tag{26}$$

where μ_k is the non-negative Lagrangian multiplier associated with constraints $C1$-$C4$ of problem $\mathbf{P_1}$ in Eq. (13), $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]$ represents the vector of Lagrangian multipliers, and $\beta_k \triangleq \theta_k T_f B / \log 2$ is defined as the *normalized QoS exponent* at user k. Due to the convexity of optimization problem $\mathbf{P_5}$, the duality gap between $\mathbf{P_5}$ and its dual problem

is zero. Correspondingly, we can formulate the Lagrangian dual function as follows:

$$\mathcal{D}(\boldsymbol{\mu}) = \arg\max_{\{\boldsymbol{e},\boldsymbol{\tau}\}} \widetilde{\mathcal{L}}(\boldsymbol{e},\boldsymbol{\tau},\boldsymbol{\mu}), \tag{27}$$

Therefore, the Lagrange dual problem $\mathbf{P_6}$ is given as follows:

$$\mathbf{P_6}: \ \arg\min_{\boldsymbol{\mu}} \mathcal{D}(\boldsymbol{\mu}), \tag{28}$$

$$\text{s.t.} \quad \boldsymbol{\mu} \geq 0. \tag{29}$$

Theorem 1: The optimal joint power and time allocation that maximizes the end-to-end effective capacity for our proposed EH based multihop scheme in high SINR regime can be derived as follows:

$$\begin{cases} \tau_k^{\text{opt}} = \dfrac{\mathcal{W}\left(-\log\left(e_k^{\text{opt}}\gamma_k\right)\right)}{-\log\left(e_k^{\text{opt}}\gamma_k\right)}; \\ e_k^{\text{opt}} = \dfrac{\tau_k^{\text{opt}}\beta_k}{\widetilde{\lambda}_k \prod\limits_{k=1}^{K}\left(\frac{\tau_k^{\text{opt}}\beta_k\gamma_k\lambda_k^{\text{opt}}}{\mu_k}\right)^{\frac{(1-\tau_k^{\text{opt}})\beta_k}{1+(1-\tau_k^{\text{opt}})K\beta_k}}} - \dfrac{1}{\gamma_k}, \end{cases} \tag{30}$$

where $\mathcal{W}(\cdot)$ denotes the Lambert W function [14] and λ_k^{opt} is the optimal Lagrange multiplier and can be numerically obtained by substituting Eq. (30) back into constraint $C1$ for problem $\mathbf{P_1}$ given by Eq. (13).

Proof: Applying the Karush-Kuhn-Tucker (KKT) condition, we can take the derivative of $\widetilde{\mathcal{L}}(\boldsymbol{e},\boldsymbol{\tau},\boldsymbol{\mu})$ with respect to e_k and τ_k $(1 \leq k \leq K)$ and set the results to zero as follows:

$$\begin{aligned} \frac{\partial\widetilde{\mathcal{L}}(\boldsymbol{e},\boldsymbol{\tau},\boldsymbol{\mu})}{\partial e_k} =& -(1-\tau_k)\beta_k\frac{\gamma_k}{\tau_k}\left(1+\frac{e_k}{\tau_k}\gamma_k\right)^{-1} \\ &\times\prod_{k=1}^{K}\lambda_k\left(1+\frac{e_k}{\tau_k}\gamma_k\right)^{-(1-\tau_k)\beta_k} p_{\Gamma_k}(\gamma_k) \\ &+\frac{\mu_k(1-\tau_k)}{(\tau_k)^2}p_{\Gamma_k}(\gamma_k) = 0. \end{aligned} \tag{31}$$

and

$$\begin{aligned} \frac{\partial\widetilde{\mathcal{L}}(\boldsymbol{e},\boldsymbol{\tau},\boldsymbol{\mu})}{\partial \tau_k} =& \left(\beta_k\log\left(1+\frac{e_k}{\tau_k}\gamma_k\right)+\frac{(1-\tau_k)\beta_k e_k\gamma_k}{\left(1+\frac{e_k}{\tau_k}\gamma_k\right)(\tau_k)^2}\right) \\ &\times\prod_{k=1}^{K}\lambda_k\left(1+\frac{e_k}{\tau_k}\gamma_k\right)^{-(1-\tau_k)\beta_k} p_{\Gamma_k}(\gamma_k) \\ &+\mu_k e_k\left(\frac{-2}{(\tau_k)^3}+\frac{1}{(\tau_k)^2}\right)p_{\Gamma_k}(\gamma_k) = 0, \end{aligned} \tag{32}$$

where $p_{\Gamma_k}(\gamma_k)$ is defined in Eq. (9). According to Eq. (31), we have

$$\left(1+\frac{e_k}{\tau_k}\gamma_k\right)^{-1} = \frac{\mu_k}{\tau_k\beta_k\gamma_k\prod\limits_{k=1}^{K}\lambda_k\left(1+\frac{e_k}{\tau_k}\gamma_k\right)^{-(1-\tau_k)\beta_k}}. \tag{33}$$

Multiplying K equations in Eq. (33), we can derive the relationship between the optimal power allocation e_k^{opt} and the optimal time allocation τ_k^{opt} as follows:

$$e_k^{\text{opt}} = \frac{\tau_k^{\text{opt}}\beta_k}{\widetilde{\lambda}_k\prod\limits_{k=1}^{K}\left(\frac{\tau_k^{\text{opt}}\beta_k\gamma_k\lambda_k^{\text{opt}}}{\mu_k}\right)^{\frac{(1-\tau_k^{\text{opt}})\beta_k}{1+(1-\tau_k^{\text{opt}})K\beta_k}}} - \frac{1}{\gamma_k}, \tag{34}$$

where λ_k^{opt} is the optimal Lagrange multiplier and can be numerically obtained by substituting Eq. (34) back into constraint $C1$ for problem $\mathbf{P_1}$ given by Eq. (13). Then, combing Eq. (31) and Eq. (32), we have

$$\log\left(1+\frac{e_k}{\tau_k}\gamma_k\right) = \frac{e_k\gamma_k}{(\tau_k)^2}\left(1+\frac{e_k}{\tau_k}\gamma_k\right)^{-1}. \tag{35}$$

In high SINR regime, we have $1+e_k\gamma_k/\tau_k \gg 1$. Correspondingly, we can approximately rewrite Eq. (35) as follows:

$$\log\left(\frac{e_k}{\tau_k}\gamma_k\right) = \frac{e_k\gamma_k}{(\tau_k)^2}\left(\frac{e_k}{\tau_k}\gamma_k\right)^{-1}. \tag{36}$$

Then, we apply the Lambert W function [14] to solve Eq. (35) for $k = 1, 2, \ldots, K$. Then, we can derive the optimal time allocation that maximizes the end-to-end effective capacity for our proposed EH based multihop scheme as follows:

$$\tau_k^{\text{opt}} = \frac{\mathcal{W}\left(-\log\left(e_k^{\text{opt}}\gamma_k\right)\right)}{-\log\left(e_k^{\text{opt}}\gamma_k\right)}, \tag{37}$$

where $\mathcal{W}(\cdot)$ denotes the Lambert W function. By substituting Eq. (37) into Eq. (34), we can derive the optimal power allocation policy for our proposed EH based scheme as in Eq. (30). Therefore, we complete the proof of Theorem 1. ■

V. Q-Learning Based Multihop Routing Algorithm Under Heterogeneous Statistical Delay-Bounded QoS Provisionings

In the previous section, we derive the optimal joint power and time allocation policy by applying the DTMDP based algorithm. However, in order to use DTMDP based algorithm, we assume that the transition probability can be calculated in advance for deriving the optimal multihop routing policy in Algorithm 1 over EH based multihop relay networks. In practical scenarios, it is usually infeasible to derive an exact model of the transition probability for our proposed EH based multihop relay scheme because of the large state space and the uncertainty of the energy state information (ESI), CSI, and the mobile users' mobility [15]. As a result, we propose to apply the reinforcement learning method which allows the mobile users to try different actions in different network states within infinite number of times until it can gradually adapt to the dynamically changing environment according to the received feedbacks. In this section, we propose the Q-learning based multihop routing algorithm in order to solve the optimization problem $\mathbf{P_1}$ with unknown transition probability. In particular, we apply the Q-learning based multihop routing algorithm at the source node to choose a route with the lowest maximum end-to-end effective capacity. Q-learning method can gradually learn an optimal decision policy without knowing the transition probabilities.

The goal of the Q-learning based multihop routing algorithm is to learn to select the best action given state $\boldsymbol{S}(t)$ for users such that the best reward (maximum end-to-end effective capacity) can be achieved. Given state $\boldsymbol{S}(t)$ at time slot t, a finite number of possible actions can be selected to perform at the users.

Algorithm 2 Q-Learning Based Multihop Routing Algorithm

Input: $T_f, B, K, \overline{\mathcal{P}}_1, \overline{\mathcal{P}}_2, \ldots, \overline{\mathcal{P}}_K$
Initialization: $[\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K] = [\overline{\mathcal{P}}_1, \overline{\mathcal{P}}_2, \ldots, \overline{\mathcal{P}}_K]$
for each state $\boldsymbol{S}$ and action $\boldsymbol{a}$ **do**
  $Q(\boldsymbol{S}, \boldsymbol{a}) = 0$
end for
Learning:
if $rand(\cdot) < \epsilon$ **then**
  randomly select a routing action
else
  Observe the current state $\boldsymbol{S}(t)$ and select an action $\boldsymbol{a}(t)$ with maximum Q value.
end if
Observe the transition state $\boldsymbol{S}(t) \rightarrow \boldsymbol{S}(t+1)$ and calculate the immediate reward.
Update the Q value function using Eq. (39).
Set $t \leftarrow (t+1)$.

Correspondingly, user k need to guarantee that the transmit power satisfies the following constraint:

$$\begin{cases} \mathcal{P}_k(t) \leq B_k(t); \\ B_k(t+1) = B_k(t) - \min\{\mathcal{P}_k(t), B_k(t)\} + E_k(t). \end{cases} \tag{38}$$

With the unknown users' mobility, it is obvious that transition probability is no longer needed in Q-learning algorithm. Accordingly, using the Q-learning algorithm, the network controller need to calculate Q-function $Q(\boldsymbol{S}(t), \boldsymbol{a}(t))$ at each time slot t in order to learn which action is optimal for the corresponding state. The Q value updating depends on the states $\{\boldsymbol{S}(t+1), \boldsymbol{S}(t)\}$ and actions $\boldsymbol{a}(t)$, and thus the corresponding reward can be derived as follows:

$$Q(\boldsymbol{S}(t), \boldsymbol{a}(t)) \leftarrow (1-\widehat{\alpha}) Q(\boldsymbol{S}(t), \boldsymbol{a}(t)) + \widehat{\alpha}\Bigg[\sum_{k=1}^{K} C_k(\theta_k, t) + \zeta \max_{\boldsymbol{a} \in \boldsymbol{A}} \sum_{\widetilde{\boldsymbol{S}} \in \mathcal{S}} Q\left(\widetilde{\boldsymbol{S}}, \boldsymbol{a}\right)\Bigg], \tag{39}$$

where $0 < \widehat{\alpha} \leq 1$ denotes the learning rate, ζ is the discount factor, where $\widetilde{\boldsymbol{S}}$ denotes the next state at time slot $(t+1)$, and $\max_{\boldsymbol{a} \in \boldsymbol{A}} Q\left(\widetilde{\boldsymbol{S}}, \boldsymbol{a}\right)$ represents the estimation of the optimal future Q value based on the best selected actions. In addition, the ε-greedy policy is applied at the network controller in each time slot for the action selection, in which the action with the maximum value of the effective capacity is chosen with a high possibility $(1-\varepsilon)$, and the rest actions are selected randomly with a small probability ε. Accordingly, the Q-learning based multihop routing algorithm for our proposed EH scheme is described in Algorithm 1.

According to Algorithm 1, using ϵ-greedy policy, the proposed Q-learning based multihop routing algorithm converges with probability one as $t \rightarrow \infty$ if the learning rate $\widehat{\alpha}$ satisfies the following conditions:

$$\left\{0 \leq \widehat{\alpha} < 1; \sum_{t} \widehat{\alpha} = \infty; \sum_{t} \widehat{\alpha}^2 < \infty\right\}. \tag{40}$$

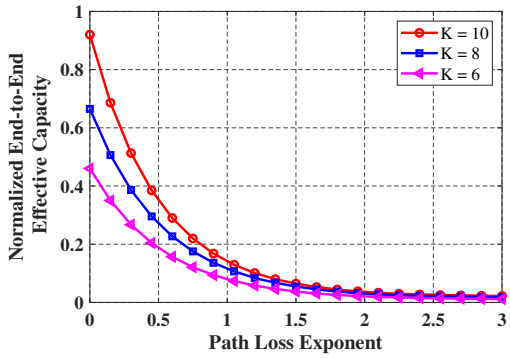

Fig. 2. The normalized end-to-end effective capacity v.s. path loss exponent $\widetilde{\alpha}$ over multihop big-data relay networks.

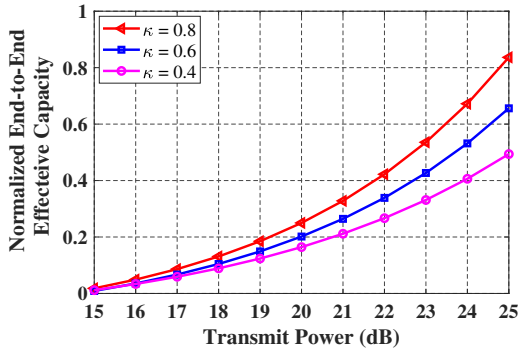

Fig. 3. The normalized end-to-end effective capacity v.s. transmit power at the energy harvester over multihop big-data relay networks.

For brevity, the convergence of the proposed Q-learning algorithm can be found in [16] when the learning rate satisfies the conditions in Eq. (40).

VI. Performance Evaluations

We use simulations to validate and evaluate our proposed schemes. Throughout our simulations, we set the bandwidth $B = 5$ MHz for all mobile users, the time frame $T_f = 2$ s, and the average transmit power $\overline{\mathcal{P}}_k = 1$ Watt for mobile user k.

Define the normalized effective capacity as the effective capacity divided by B and T_f, which then has the unit of bits/sec/Hz. Consider the three-hop EH based multihop scheme, Fig. 2 and Fig. 3 plot the normalized end-to-end effective capacity v.s. path loss exponent $\widetilde{\alpha}$ and transmit power at the energy harvester, respectively. As shown in Fig. 2, the normalized end-to-end effective capacity decreases as the quality of wireless channel becomes worse. We can observe from Fig. 2 that the normalized end-to-end effective capacity increases with the increase of the number of users. Also, Fig. 3 shows that our proposed scheme performs better with more harvested energy.

Figure 4 depicts the normalized end-to-end effective capacity with different number of hops for our proposed EH based big-

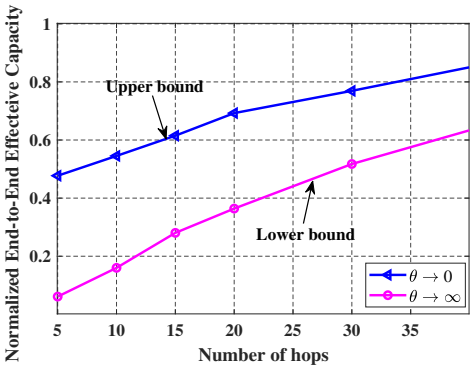

Fig. 4. The upper bound and lower bound for the normalized end-to-end effective capacity with different number of hops.

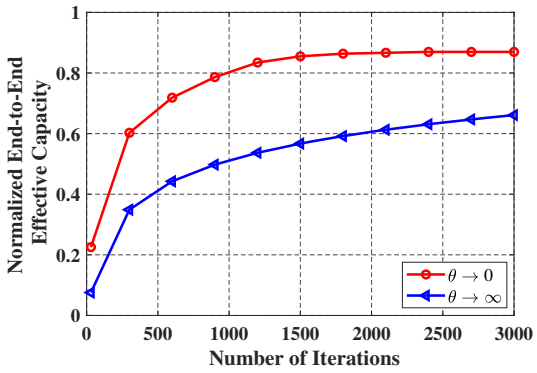

Fig. 5. The upper bound and lower bound for the normalized end-to-end effective capacity with different number of hops.

data multihop relay networks. Fig. 4 shows that the loose QoS exponent ($\theta \to 0$) and the stringent QoS exponent ($\theta \to \infty$) set the upper bound and lower bound for the normalized effective capacity, respectively. Also as shown in Fig. 4, the normalized end-to-end effective capacity increases as the number of hops increases, which implies that our proposed multihop EH scheme can outperform the traditional single-hop schemes in terms of the normalized end-to-end effective capacity.

Set the number of users $K = 8$ and the energy harvesting efficiency $\kappa = 0.6$. Using Algorithm 1, Fig. 5 plots the upper bound and lower bound of the normalized end-to-end effective capacity for our proposed Q-learning based multihop routing algorithm with different number of mobile user over big-data multihop relay networks. We can observe from Fig. 5 that the normalized end-to-end effective capacity increases as the number of iterations increases, and finally achieves the optimal multihop routing strategy.

VII. Conclusions

We have designed the Q-learning based algorithm for optimizing power and routing policies through learning from the history of the EH process while satisfying the heterogeneous statistical delay-bounded QoS constraints over multihop big-data relay networks. In particular, we have established and analyzed the wireless communication model as well as the EH model. Under the heterogeneous statistical delay-bounded QoS requirements, we have formulated the end-to-end effective-capacity optimization problem for the battery-free EH based multihop relay networks. Then, we have developed DTMDP and Q-learning based multihop routing algorithms. We have also conducted a set of simulations which show that our proposed Q-learning based EH scheme outperforms other existing schemes under the heterogeneous statistical delay-bounded QoS constraints over multihop big-data relay networks.

References

[1] H. Su and X. Zhang, "Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE Journal on Selected Areas in Comm.*, vol. 26, no. 1, pp. 118–129, Jan. 2008.

[2] J. Wang and X. Zhang, "Heterogeneous QoS-driven resource adaptation over full-duplex relay networks," in *IEEE GLOBECOM 2016*.

[3] T. Z. Oo, N. H. Tran, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Offloading in HetNet: A coordination of interference mitigation, user association, and resource allocation," *IEEE Transactions on Mobile Computing*, vol. 16, no. 8, pp. 2276–2291, Aug. 2017.

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

[5] S. Akbar, Y. Deng, A. Nallanathan, M. Elkashlan, and A.-H. Aghvami, "Simultaneous wireless information and power transfer in K-tier heterogeneous cellular networks," *IEEE Trans. on Wireless Comm.*, vol. 15, no. 8, pp. 5804–5818, Aug. 2016.

[6] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Auto. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[7] X. Zhang, W. Cheng, and H. Zhang, "Heterogeneous statistical QoS provisioning over 5G mobile wireless networks," *IEEE Network Magazine*, vol. 28, no. 6, pp. 46–53, Nov. 2014.

[8] W. Cheng, X. Zhang, and H. Zhang, "Heterogeneous statistical QoS provisioning for downlink transmissions over mobile wireless cellular networks," in *IEEE GLOBECOM 2014*, pp. 4757–4763.

[9] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. on Wireless Comm.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[10] T. J. Van and L. A. Wolsey, "Solving mixed integer programming problems using automatic reformulation," *IEEE Transactions on Wireless Communications*, vol. 35, no. 1, pp. 45–57, Aug. 2007.

[11] R. Bellman, "Dynamic programming," in *Princeton University Press*, 1957.

[12] M. L. Littman, T. L. Dean, and L. P. Kaelbling, "On the complexity of solving Markov decision problems," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 394–402.

[13] C. Xu, M. Zheng, W. Liang, H. Yu, and Y.-C. Liang, "End-to-end throughput maximization for underlay multi-hop cognitive radio networks with RF energy harvesting," *IEEE Transactions on Wireless Comm.*, vol. 6, no. 6, pp. 3561–3572, June 2017.

[14] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the lambert w function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[15] X. Chen, J. Wu, Y. Cai, H. Zhang, and T. Chen, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 4, pp. 3520–3535, April 2015.

[16] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.