IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received May 13, 2016, accepted May 30, 2016, date of publication June 6, 2016, date of current version September 28, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2577020

Mode Selection and Resource Allocation in Device-to-Device Communications With User Arrivals and Departures

LEI LEI, (Member, IEEE), QINGYUN HAO, AND ZHANGDUI ZHONG
State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: L. Lei (leil@bjtu.edu.cn)

This work was supported in part by the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University under Grant RCS2016ZT020 and in part by the National Natural Science Foundation of China under Grant 61272168 and Grant U1334202.

ABSTRACT The pervasive increasing mobile devices and explosively increasing data traffic pose imminent challenges on wireless network design. Device-to-device (D2D) communication is envisioned to play a key role in the fifth generation cellular networks to efficiently support much larger and more diverse set of devices. This paper investigates the mode selection and resource allocation for D2D communications with dynamic user arrivals and departures. We formulate the optimal resource control problem to minimize the average energy consumption of flow transmission into an infinite horizon average reward Markov decision process. In order to deal with the well-known curse of dimensionality problem and facilitate distributed implementation, we approximate the mode selection Q-factor by the sum of per-queue mode selection Q-factors. Moreover, we apply distributive stochastic online learning to estimate the per-queue Q-factors. Simulation results show that the proposed approach outperforms various existing baseline algorithms.

INDEX TERMS Device-to-device communications, flow-level model, Markov decision process.

I. INTRODUCTION

Device-to-Device (D2D) communications commonly refer to a type of technologies that enable devices to communicate directly with each other without the communication infrastructure, e.g., access points (APs) or base stations (BSs). With the emergence of context-aware applications and the accelerating growth of Machine-to-Machine (M2M) applications, D2D function plays an increasingly important role. This is because it facilitates the discovery of geographically close devices, and enables direct communications between these proximate devices, which increases communication capability and reduces communication delay and power consumption. To seize the emerging market that requires D2D function, the mobile operators and vendors are exploring the possibilities of introducing D2D communications in the cellular networks [1]. D2D communication is currently accepted as a part of fourth generation (4G) Long Term Evolution (LTE)-Advanced standard in 3rd Generation Partnership Project (3GPP) Release 12 [2], and it is also envisioned to continuously evolve into the fifth generation (5G) cellular networks to efficiently support much larger and more diverse set of devices [3].

When D2D communications are supported in cellular networks, the data between a pair of D2D user equipments (UEs) can either be routed along a one-hop route of D2D link (direct over-the-air link) in *D2D mode*, or a two-hop route of cellular links in *cellular mode*. Moreover, two resource sharing paradigms are defined for D2D communications: either *overlay* where cellular and D2D links use orthogonal time/frequency resources, or *underlay* where D2D links can access the time/frequency resources occupied by cellular users [4]. Mode selection and resource allocation are two important resource control functions in D2D communications.

Recently, there is a vast body of work on resource control design for D2D communications under the packet-level model, which considers that the user number is fixed, and the traffic pattern is usually assumed to be saturated with infinite backlogs (i.e., each user always has data to transmit). However, the number of users varies with time in practice, where new users arrive according to a stochastic process, and each user has a certain amount of data for transmission [5]. A user leaves the system when the data transmission is completed. The optimal resource control policies and corresponding performance under the above flow-level model can be very different from those derived and predicted under the packet-level model. As a simple example, consider a cellular network with overlay D2D communications. Under the infinite backlog traffic model, the mode selection decision is mostly related to the UE positions, which impact the time-average transmission rates of the D2D links and cellular links. On the other hand, the optimal mode under flow-level model also depends on the user population of cellular and D2D communications, i.e., the mode with smaller user population is favored.

The flow-level models are extremely difficult to solve for optimality due to the time-varying service rates and dynamic user population. There are several important works that analyze the flow-level performance of cellular networks, which use a simple constant-rate service process to approximate the time-varying service rate due to fast fading [5], [6]. However, to the best of our knowledge, there is few mathematical model to optimize the resource control functions of cellular networks under flow-level model with time-varying service rate [7]. Moreover, the optimization of D2D communications under the flow-level model is barely studied in existing work.

Compared with traditional cellular networks, the design of resource control for D2D communications has some unique requirements: 1) the time scale of mode selection needs to be determined, since the selected mode should be updated with the varying radio condition and network load for better performance, but not updated too often to avoid large signaling overhead; 2) it is preferable to perform the resource control distributively by the UEs with proper help and control from the BS, since a fully centralized solution as in traditional cellular networks brings an exponential computational complexity and huge signaling overhead; 3) the resource control functions should be jointly optimized to achieve better performance since they are closely related, e.g., the mode selection decision depends on the system or user performance under each mode, which is greatly impacted by the resource allocation and power control schemes.

In this paper, we consider an Orthogonal Frequency Division Multiple Access (OFDMA) cellular network with overlay D2D communications and dynamic cellular and D2D flow arrivals. Our objective is to design a joint distributed mode selection and resource allocation policy with two time scales to minimize the average energy consumption of flow transmission. We formulate an infinite horizon average reward Markov Decision Process (MDP) problem for the dynamic optimization of mode selection and resource allocation over frequency-selective fading channel with Adaptive Modulation and Coding (AMC) scheme in the physical layer under dynamic flow arrival and departure. The mode selection is only adaptive to the flow dynamics, i.e., the user population variation, while the resource allocation is adaptive to both the channel state variation and flow dynamics. Since it is well-known that there is generally no simple solution for the MDP problem because the brute-force value iterations or policy iterations could not lead to any viable solution due to the curse of dimensionality, we use approximate dynamic programming and online stochastic learning to reduce complexity and facilitate distributed implementation. In the proposed solution, every user receives a couple of per-queue Q-factors from the BS when it first arrives, and it distributively makes mode selection decision and computes bids for resource allocation based on the Q-factors. Then, it submits the bids to the BS which makes the resource allocation decision. The per-queue Q-factors are distributively updated at the UEs, which submit the updated Q-factors to the BS after they finish transmission and are ready to leave the system. Different from our previous works in [8] and [9] which are based on the packet-level model with fixed number of users, we focus on the flow-level model with time-varying user population in this paper. The main contribution of this paper lies in the following aspects:

1) We define a new performance metric under the flow-level model, i.e., the mean flow transmission energy consumption, which is the product of the mean flow transmission delay and transmission power. The optimization of this performance metric achieves the best trade-off between throughput maximization and power minimization, which are the two mostly considered optimization objectives under the infinite backlog traffic model.
2) We formulate a queuing model and provide the underlying Markov chain for the flow-level model with time-varying service rates, which facilitate the derivation of the distributed solutions for resource control. Both the statistical information in the space and time domains are exploited.
3) We propose a two-timescale, distributed and joint mode selection and resource allocation control under the flow-level model, where the mode selection control is only adaptive to the flow dynamics but not the channel variation due to fast fading. The tradeoff between performance, complexity and signaling overhead is fully considered in the proposed solution.

The organization of the paper is as follows. Section II reviews the related work. The system model of D2D overlaid cellular network with flow-level dynamics is described in Section III. In Section IV, we formulate an infinite horizon average reward MDP model for the optimization of mode selection and resource allocation control. In Section V we derive a low complexity learning algorithm, which updates the per-queue Q-factors based on real-time observations of channel state information (CSI) and queue state information (QSI), as well as a distributive mode selection algorithm and a resource allocation algorithm with auction mechanism. In Section V, we discuss the performance simulations. Finally, we summarize the main results in Section VI.

II. RELATED WORK

Mode selection in D2D communications has been extensively studied by many works in literature [10], almost all of which focus on the packet-level model with fixed number of users - mostly adopt an infinite backlog traffic model [11]–[16], while a few of our recent works assume the dynamic packet arrival model [8], [9], [17]. Mode selection can be performed at different time scales, either statically where the selected mode remains fixed for a pair of D2D UEs, or dynamically per time slot. Dynamic mode selection can capture and utilize the fast fading effects of wireless channels opportunistically, while static mode selection has the advantage of saving computation and communication overhead. The design of mode selection function has to take into account the other two closely related resource control functions, i.e., resource allocation and power control, and the three functions can be either separately [4], [11], [12], [15], [17] or jointly [8], [9], [13], [14], [18] optimized.

Research on mode selection uses various methods to deal with the problem. The mode that can achieve the best system performance in terms of spectral efficiency [11], energy efficiency [12], or delay [17] may be selected, where the performance under every mode is estimated assuming the power control and resource allocation algorithms are either given or optimized. For this category of methods, the main difference between static and dynamic mode selection lies in whether the long-term time-average performance or instantaneous performance per time slot is used as the selection criteria. Another category of methods perform distance-based mode selection using stochastic geometry analysis [4], [15], where [4] only considers the distances between D2D UEs and derives an optimal distance threshold, while [15] additionally considers the distances between D2D UEs and the BS and an optimal bias factor is determined. This category of methods exploits the statistical information in the space domain (i.e., the distributions of the UE locations) and belongs to the static mode selection. Finally, when mode selection is performed dynamically at each time slot, the problem becomes deciding whether to schedule the D2D link or the cellular link for a pair of D2D UEs. Therefore, a third category of methods implicitly solve the mode selection problem by the resource allocation algorithms [8], [9], [13], [14].

Mode selection can be either centrally performed by the BS or distributively determined by the UEs themselves. Most existing works in literature focus on the centralized scheme. In [16], a dynamic Stackelberg game is presented in which the BS and the D2D UEs act as the leader and the followers, respectively, to enable distributed user-controlled mode selection.

III. SYSTEM MODEL

Consider a D2D overlaid OFDM cellular network, where there are multiple D2D UE (DUE) pairs and cellular UEs (CUEs) in a cell. A DUE pair consists of a source D2D UE (src. DUE) and a destination D2D UE (dest. DUE) within direct over-the-air communication range. The whole uplink spectrum is divided into N_{F} equal size subchannels, where ηN_{F} subchannels are allocated to cellular communications and the rest of the $(1-\eta)N_{\mathrm{F}}$ subchannels are allocated to D2D communications. The data transmission is performed on a slot-by-slot basis, where all time slots have equal length.

A. TRAFFIC CHARACTERISTICS

We define a dynamic flow model for elastic traffic, where new DUE pairs and new CUEs arrive at the system with finite-size file transmission tasks from the src. DUEs to the dest. DUEs and the CUEs to the BS, and leave the system after they finish transmissions. Therefore, we will also use the terms "D2D flow" and "cellular flow" to refer to "DUE pair" and "CUE" in the rest of the paper. Under the dynamic flow model, the number of UEs or flows in the system varies over time due to the arrival and departure processes. Denote by $n_{\mathrm{C},t}$ and $n_{\mathrm{D},t}$ the number of cellular flows and D2D flows at time slot t, respectively.

The cellular (resp. D2D) flow sizes are assumed to be independently and exponentially distributed with mean measure σ_{C} (resp. σ_{D}). Consider cellular (resp. D2D) flows arrive at the system as a homogeneous spatial Poisson process with mean arrival rate λ_{C} (resp. λ_{D}), so that the rate at which CUEs (resp. src. DUEs) arrive into an area $\mathcal{X}$ is $\lambda_{\mathrm{C}}\frac{\mathcal{X}}{\pi R^2}$ (resp. $\lambda_{\mathrm{D}}\frac{\mathcal{X}}{\pi R^2}$) with R denoting the radius of the circular cell. We consider the transceiver distance d_{D} of a typical DUE pair is Rayleigh distributed with maximum value of R [2]:

$$f_{d_{\mathrm{D}}}(x)=\frac{2\pi\xi x e^{-\xi\pi x^2}}{1-e^{-\xi\pi R^2}},\quad 0\leq x\leq R \tag{1}$$

Note that (1) implies that a dest. DUE is randomly distributed on the circle centered at its src. DUE according to a two-dimensional Gaussian distribution. The following analysis can be extended to other distributions as well.

We consider that at most n_{Cmax} cellular flows and n_{Dmax} D2D flows can be admitted into a cell simultaneously. Cellular (resp. D2D) flows which arrive when there are already n_{Cmax} (resp. n_{Dmax}) transfers in progress in the cell are denied access and abandon. Let $i\in\{1,\ldots,n_{\mathrm{Cmax}}\}$ and $j\in\{1,\ldots,n_{\mathrm{Dmax}}\}$ denote the cellular flow index and D2D flow index, respectively. A flow is assigned an index that is not used by other flows once it is admitted into the system, and the index is released and can be assigned to new flows after the considered flow finishes transmission and leaves the system. The flow index and positions of the UEs remain fixed for the duration of the transfer.

B. RESOURCE CONTROL POLICY

Mode selection needs to be performed for any new D2D flow upon arrival. Let $y_{j,t}, j\in\{1,\ldots,n_{\mathrm{Dmax}}\}$ denote the selected mode of D2D flow j at time slot t, where $y_{j,t}=1$ if D2D mode is selected and $y_{j,t}=0$ otherwise. If D2D flow j does not exist at time slot t, let $y_{j,t}=0$. Denote the total number of D2D flows in D2D mode and cellular mode at time slot t by $n_{\mathrm{Dd},t}=\sum_{j=1}^{n_{\mathrm{Dmax}}}y_{j,t}$ and $n_{\mathrm{Dc},t}=n_{\mathrm{D},t}-\sum_{j=1}^{n_{\mathrm{Dmax}}}y_{j,t}$, respectively.

The data of a cellular flow is always transmitted over the corresponding cellular uplink, while the data of a D2D flow can be either transmitted over a D2D link or a cellular uplink, depending on whether the D2D mode or cellular mode is selected. Let $l_{\mathrm{C},i0}$, $i \in \{1, \ldots, n_{\mathrm{Cmax}}\}$ denote the cellular uplink corresponding to the communication channel from CUE i to the BS when cellular flow i exists. Let $l_{\mathrm{D},j0}$ and $l_{\mathrm{D},jj}$, $j \in \{1, \ldots, n_{\mathrm{Dmax}}\}$ denote the cellular uplink corresponding to the communication channel from src. DUE j to the BS, and the D2D link corresponding to the communication channel from src. DUE j to dest. DUE j, respectively, when D2D flow j exists. Since there are two types of links in our model, i.e., D2D links and cellular uplinks, we use the term "link" to refer to any type of links in the rest of paper.

When the selected mode is determined for all the D2D flows at time slot t, the sets of cellular uplinks and D2D links for D2D flows are restricted to $\{l_{\mathrm{D},j0} | y_{j,t} = 0\}$ and $\{l_{\mathrm{D},jj} | y_{j,t} = 1\}$, respectively. We consider each of the ηN_{F} (resp. $(1-\eta)N_{\mathrm{F}}$) subchannels is assigned to one of the cellular uplinks (resp. D2D links) at each time slot. Denote the index of a subchannel by $m \in \{1, \ldots, N_F\}$. Let $x^{(m)}_{\mathrm{C},i0,t}$(resp. $x^{(m)}_{\mathrm{D},j0,t}, x^{(m)}_{\mathrm{D},jj,t}$) $\in \{0, 1\}$ denote the subchannel allocation for link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}$, $l_{\mathrm{D},jj}$) at time slot t, where $x^{(m)}_{\mathrm{C},i0,t}$(resp. $x^{(m)}_{\mathrm{D},j0,t}, x^{(m)}_{\mathrm{D},jj,t}$) $= 1$ if subchannel m is allocated to link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}$, $l_{\mathrm{D},jj}$), and $x^{(m)}_{\mathrm{C},i0,t}$(resp. $x^{(m)}_{\mathrm{D},j0,t}, x^{(m)}_{\mathrm{D},jj,t}$) $= 0$ otherwise. Moreover, let $x^{(m)}_{\mathrm{C},i0,t}$(resp. $x^{(m)}_{\mathrm{D},j0,t}, x^{(m)}_{\mathrm{D},jj,t}$) $= 0$ if cellular flow i (resp. D2D flow j) does not exist. Therefore, for any $m \in \{1, \ldots, \eta N_{\mathrm{F}}\}$ we have $\sum_{i=1}^{n_{\mathrm{Cmax}}} x^{(m)}_{\mathrm{C},i0,t} + \sum_{j=1}^{n_{\mathrm{Dmax}}} x^{(m)}_{\mathrm{D},j0,t}(1 - y_{j,t}) = 1$; while for any $m \in \{\eta N_{\mathrm{F}} + 1, \ldots, N_{\mathrm{F}}\}$ we have $\sum_{j=1}^{n_{\mathrm{Dmax}}} x^{(m)}_{\mathrm{D},jj,t} y_{j,t} = 1$.

We consider a cellular uplink or D2D link uses channel inversion for power control, i.e., $P = d^{\alpha}$, where P and d denote the transmission power and link length, respectively, and $\alpha > 2$ denotes the path-loss exponent. This means that the channel state processes of all UEs are symmetric in the sense that they are subject to the same slow fading (that does not change over the time period of interest), with the fast fading being statistically identical for all users. We use the term "D2D transmitter" to refer to a src. DUE in D2D mode, while "cellular transmitter" to refer to either a CUE or a src. DUE in cellular mode. Let $P_{\mathrm{Ct},i,t}$ and $P_{\mathrm{Ct},j,t}$ denote the transmission power of a cellular transmitter corresponding to cellular flow i or D2D flow j in cellular mode, respectively. Let $P_{\mathrm{Dd},j,t}$ denote the transmission power of a D2D transmitter corresponding to D2D flow j in D2D mode. Since the transmission power of cellular transmitters are dependent on their distance with the BS, we divide the circular area of the considered cell into K disjunct zones by $K - 1$ concentric circles around the BS, where the zone $k \in \{1, \ldots, K\}$ is the region between two concentric circles with radius d_{k-1} and d_k. Obviously, $d_0 = 0$ and $d_K = R$, respectively. The transmission power of all the cellular transmitters in zone k can be considered as approximately the same, i.e., $P_{\mathrm{Ct},i,t} = P_{\mathrm{Ct},j,t} = P^{(k)} = (\frac{d_{k-1}+d_k}{2})^{\alpha}$, if CUE i and src. DUE j are in zone k. Similarly, since the transmission power of D2D transmitters are dependent on the distance with their respective D2D receivers, we consider that the transmission power of all D2D transmitters with $d_{\mathrm{D}} \in (d_{k-1}, d_k]$ are approximately the same, i.e, $P_{\mathrm{Dd},j,t} = P^{(k)}$ if the distance between DUE pair j falls within $(d_{k-1}, d_k]$. Let $P_{\mathrm{Ct},i,t} = 0$ (resp. $P_{\mathrm{Ct},j,t} = P_{\mathrm{Dd},j,t} = 0$) if cellular i (resp. cellular j) does not exist at time slot t.

Remark 1 (Motivation of Two Time-Scale Control Policy): We consider that the mode selection and resource allocation policies are performed over two different time scales. While the resource allocation decision is made on a per-slot basis, the mode selection decision is updated at the much slower time scale of flow dynamics. This is because although the channel state variation due to fast fading does have some impact on the mode selection decision, the most essential factors that determine the optimal decision are the number of flows and the spatial distribution of the UEs, which only change with the flow arrival and departure.

C. INSTANTANEOUS DATA RATE

The path loss of the wireless channels is compensated by the power control mechanism. We consider that the instantaneous channel gain comprising only the fast fading effect of link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}$, $l_{\mathrm{D},jj}$) on any subchannel m remains constant within a time slot and i.i.d. between time slots, the value of which is denoted by $G^{(m)}_{\mathrm{C},i0,t}$ (resp. $G^{(m)}_{\mathrm{D},j0,t}, G^{(m)}_{\mathrm{D},jj,t}$). The Signal to Noise Ratio (SNR) of a link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}$ and $l_{\mathrm{D},jj}$) on a subchannel m at time slot t can be derived as $\gamma^{(m)}_{\mathrm{C},i0,t} = \frac{G^{(m)}_{\mathrm{C},i0,t}}{N^{(m)}_{\mathrm{C},i0,t}}$ (resp. $\gamma^{(m)}_{\mathrm{D},j0,t} = \frac{G^{(m)}_{\mathrm{D},j0,t}}{N^{(m)}_{\mathrm{D},j0,t}}, \gamma^{(m)}_{\mathrm{D},jj,t} = \frac{G^{(m)}_{\mathrm{D},jj,t}}{N^{(m)}_{\mathrm{D},jj,t}}$), where $N^{(m)}_{\mathrm{C},i0,t}$ (resp. $N^{(m)}_{\mathrm{D},j0,t}, N^{(m)}_{\mathrm{D},jj,t}$) denotes the noise power on subchannel m at time slot t.

We assume that Adaptive Modulation and Coding (AMC) is used in the physical layer, where the SNR values are divided into V non-overlapping consecutive regions. For any $v \in \{1, \ldots, V\}$, if the SNR value $\gamma^{(m)}_{\mathrm{C},i0,t}$ (resp. $\gamma^{(m)}_{\mathrm{D},j0,t}, \gamma^{(m)}_{\mathrm{D},jj,t}$) of link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}$, $l_{\mathrm{D},jj}$) falls within the v-th region $[\Gamma_{v-1}, \Gamma_v)$, the corresponding data rate $r^{(m)}_{\mathrm{C},i0,t}$ (resp. $r^{(m)}_{\mathrm{D},j0,t}$, $r^{(m)}_{\mathrm{D},jj,t}$) is a fixed value R_v according to the selected modulation and coding scheme in this state. Obviously, $\Gamma_0 = 0$ and $\Gamma_V = \infty$. Also, we have $R_1 = 0$, i.e., no packet is transmitted in channel state 1 to avoid the high transmission error probability. Define the channel state information (CSI) of link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}$, $l_{\mathrm{D},jj}$) as $\mathbf{H}_{\mathrm{C},i0,t} := \{H^{(m)}_{\mathrm{C},i0,t}\}_{m=1}^{\eta N_{\mathrm{F}}}$ (resp. $\mathbf{H}_{\mathrm{D},j0,t} := \{H^{(m)}_{\mathrm{D},j0,t}\}_{m=1}^{\eta N_{\mathrm{F}}}$, $\mathbf{H}_{\mathrm{D},jj,t} := \{H^{(m)}_{\mathrm{D},jj,t}\}_{m=\eta N_{\mathrm{F}}+1}^{N_{\mathrm{F}}}$), where $H^{(m)}_{\mathrm{C},i0,t}$ (resp. $H^{(m)}_{\mathrm{D},j0,t}$, $H^{(m)}_{\mathrm{D},jj,t}$) denotes its channel state on subchannel m, which equals v if $\gamma^{(m)}_{\mathrm{C},i0,t}$ (resp. $\gamma^{(m)}_{\mathrm{D},j0,t}, \gamma^{(m)}_{\mathrm{D},jj,t}$) is between $[\Gamma_{v-1}, \Gamma_v)$. Let $H^{(m)}_{\mathrm{C},i0,t} = 1$ (resp. $H^{(m)}_{\mathrm{D},j0,t} = 1$, $H^{(m)}_{\mathrm{D},jj,t} = 1$) if cellular flow i (resp. D2D flow j) does not exist at time slot t.

Let $r_{\mathrm{C},i0,t}$ (resp. $r_{\mathrm{D},j0,t}$,$r_{\mathrm{D},jj,t}$) be the instantaneous data rate of link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}$, $l_{\mathrm{D},jj}$) during time slot t, which is equal to the sum of the instantaneous data rate $r^{(m)}_{\mathrm{C},i0,t}$ (resp. $r^{(m)}_{\mathrm{D},j0,t}$, $r^{(m)}_{\mathrm{D},jj,t}$) of link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}$, $l_{\mathrm{D},jj}$) on all the subchannels that are assigned to it at time slot t, i.e., $r_{\mathrm{C},i0,t} = \sum_{m=1}^{\eta N_F} r_{\mathrm{C},i0,t}^{(m)} x_{\mathrm{C},i0,t}^{(m)}$, $r_{\mathrm{D},j0,t} = \sum_{m=1}^{\eta N_F} r_{\mathrm{D},j0,t}^{(m)} x_{\mathrm{D},j0,t}^{(m)}(1-y_{j,t})$ and $r_{\mathrm{D},jj,t} = \sum_{m=\eta N_F+1}^{N_F} r_{\mathrm{D},jj,t}^{(m)} x_{\mathrm{D},jj,t}^{(m)} y_{j,t}$.

IV. PROBLEM FORMATION

In this paper, we study the design of the mode selection and resource allocation mechanisms in order to optimize the flow-level performance. In this section, we shall formulate the problem into an infinite-horizon average reward MDP model, which consists of four elements: states, actions, state transition probabilities, and rewards.

A. SYSTEM STATE

Before defining the system state of the MDP model, we first formulate a queuing model based on the system model described above. We consider there are $n_{\mathrm{Cmax}} + n_{\mathrm{Dmax}}$ virtual queues in a single cell, where $\{q_{\mathrm{C},i}\}_{i=1}^{n_{\mathrm{Cmax}}}$ denotes the set of queues for cellular flows, while $\{q_{\mathrm{D},j}\}_{j=1}^{n_{\mathrm{Dmax}}}$ denotes the set of queues for D2D flows. A cellular (resp. D2D) flow that arrives at the system is immediately assigned by a virtual flow dispatcher to an empty queue $q_{\mathrm{C},i}$, $i \in \{1,\dots,n_{\mathrm{Cmax}}\}$ (resp. $q_{\mathrm{D},j}, j \in \{1,\dots,n_{\mathrm{Dmax}}\}$) once admitted into the system. On the other hand, an occupied queue becomes empty after the corresponding flow finishes transmission and departs. Let $Q_{\mathrm{C},i,t}$ (resp. $Q_{\mathrm{D},j,t}$) denote the length of queue $q_{\mathrm{C},i}$ (resp. $q_{\mathrm{D},j}$) at time slot t. Thus, $Q_{\mathrm{C},i,t}, Q_{\mathrm{D},j,t} \in \{0,1\}, i \in \{1,\dots,n_{\mathrm{Cmax}}\}$, $j \in \{1,\dots,n_{\mathrm{Dmax}}\}$ indicate whether the i-th queue for cellular flows and j-th queue for D2D flows are occupied or not, respectively. Since at most n_{Cmax} cellular flows and n_{Dmax} D2D flows can be admitted in a cell, $\sum_{i=1}^{n_{\mathrm{Cmax}}} Q_{\mathrm{C},i,t} = n_{\mathrm{C},t}$ and $\sum_{j=1}^{n_{\mathrm{Dmax}}} Q_{\mathrm{D},j,t} = n_{\mathrm{D},t}$ represent the number of cellular flows and D2D flows at time slot t, respectively.

Based on the above queuing model, the global system state at time slot t can be characterized by the global queue state information (QSI) and CSI, i.e., $\mathbf{S}_t = (\mathbf{Q}_t, \mathbf{H}_t)$. The global QSI is defined as $\mathbf{Q}_t = (\mathbf{Q}_{\mathrm{C},t}, \mathbf{Q}_{\mathrm{D},t})$, where $\mathbf{Q}_{\mathrm{C},t} = \{Q_{\mathrm{C},i,t}, P_{\mathrm{Ct},i,t}\}_{i=1}^{n_{\mathrm{Cmax}}}$, $\mathbf{Q}_{\mathrm{D},t} = \{Q_{\mathrm{D},j,t}, P_{\mathrm{Ct},j,t}, P_{\mathrm{Dd},j,t}\}_{j=1}^{n_{\mathrm{Dmax}}}$. The global CSI is defined as $\mathbf{H}_t = (\mathbf{H}_{\mathrm{C},t}, \mathbf{H}_{\mathrm{D},t})$ with $\mathbf{H}_{\mathrm{C},t} = \{\mathbf{H}_{\mathrm{C},i0,t}\}_{i=1}^{n_{\mathrm{Cmax}}}$ and $\mathbf{H}_{\mathrm{D},t} = \{\mathbf{H}_{\mathrm{D},j0,t}, \mathbf{H}_{\mathrm{D},jj,t}\}_{j=1}^{n_{\mathrm{Dmax}}}$. Recall that $\mathbf{H}_{\mathrm{C},i0,t}$ (resp. $\mathbf{H}_{\mathrm{D},j0,t}, \mathbf{H}_{\mathrm{D},jj,t}$) is the CSI of link $l_{\mathrm{C},i0}$ (resp. $l_{\mathrm{D},j0}, l_{\mathrm{D},jj}$) at time slot t as given in section III.

For the system state space $\mathcal{S}$, we have $\mathcal{S} = \mathcal{Q} \times \mathcal{H}$, where $\mathcal{Q}$ represents the queue state space with $|\mathcal{Q}| = (K+1)^{n_{\mathrm{Cmax}}} + (K^2+1)^{n_{\mathrm{Dmax}}}$ and $\mathcal{H}$ represents the channel state space with $|\mathcal{H}| = V^{(n_{\mathrm{Cmax}} + 2n_{\mathrm{Dmax}})N_{\mathrm{F}}}$.

B. CONTROL POLICY

At each time slot t, based on the current state $\mathbf{S}_t$, an action $\mathbf{a}_t = \{\mathbf{y}_t, \mathbf{x}_{\mathrm{C},t}, \mathbf{x}_{\mathrm{D},t}\}$ is taken from the set of allowable actions in the action space $\mathcal{A}$, which is discrete and finite. The action is composed of the mode selection action $\mathbf{y}_t = \{y_{j,t}\}_{j\in\{1,\dots,n_{\mathrm{Dmax}}\}}$, subchannel allocation action of cellular uplinks $\mathbf{x}_{\mathrm{C},t} = \{\{x_{\mathrm{C},i0,t}^{(m)}\}_{i\in\{1,\dots,n_{\mathrm{Cmax}}\}} \cup \{x_{\mathrm{D},j0,t}^{(m)}\}_{j\in\{1,\dots,n_{\mathrm{Dmax}}\}} | m = 1,\dots,\eta N_{\mathrm{F}}\}$, as well as subchannel allocation action of D2D links $\mathbf{x}_{\mathrm{D},t} = \{\{x_{\mathrm{D},jj,t}^{(m)}\}_{j\in\{1,\dots,n_{\mathrm{Dmax}}\}} | m = \eta N_{\mathrm{F}}+1,\dots,N_{\mathrm{F}}\}$. Note that $\mathcal{A} = \mathcal{A}_{\mathrm{y}} \times \mathcal{A}_{\mathrm{xC}} \times \mathcal{A}_{\mathrm{xD}}$, where $\mathbf{y} \in \mathcal{A}_{\mathrm{y}}$, $\mathbf{x}_{\mathrm{C}} \in \mathcal{A}_{\mathrm{xC}}$, and $\mathbf{x}_{\mathrm{D}} \in \mathcal{A}_{\mathrm{xD}}$. There are $2^{n_{\mathrm{Dmax}}}$ actions in the set $\mathcal{A}_{\mathrm{y}}$, $(n_{\mathrm{Cmax}} + n_{\mathrm{Dmax}} + 1)^{\eta N_{\mathrm{F}}}$ actions in the set $\mathcal{A}_{\mathrm{xC}}$ and $(n_{\mathrm{Dmax}}+1)^{(1-\eta)N_{\mathrm{F}}}$ actions in the set $\mathcal{A}_{\mathrm{xD}}$.

A control policy prescribes a procedure for action selection in each state at all decision epoches t. We consider stationary Markovian control policies. A deterministic control policy Ω is a mapping $\mathcal{S} \to \mathcal{A}$ from the state space to the action space, which is given by $\Omega(\mathbf{S}) = \mathbf{a} \in \mathcal{A}$, $\forall \mathbf{S} \in \mathcal{S}$. In this paper, the policy Ω is composed of the mode selection policy Ω_{y} and resource allocation policy Ω_{x}, where $\Omega(\mathbf{S}) = (\Omega_{\mathrm{y}}(\mathbf{Q}), \Omega_{\mathrm{x}}(\mathbf{Q},\mathbf{H}))$.

C. STATE TRANSITION PROBABILITY

The induced random process can be represented by the discrete-time Markov chain (DTMC) $\{\mathbf{S}_t\}_{t=0,1,\dots}$. Given a system state $\mathbf{S}_t$ and an action $\mathbf{a}_t$ at time slot t, the state transition probability of the DTMC is given by

$$\begin{aligned}\Pr.\{\mathbf{S}_{t+1}|\mathbf{S}_t,\mathbf{a}_t\} &= \Pr.\{\mathbf{H}_{t+1}|\mathbf{Q}_{t+1}\}\Pr.\{\mathbf{Q}_{t+1}|\mathbf{Q}_t,\mathbf{H}_t,\mathbf{a}_t\}\\ &= \Pr.\{\mathbf{H}_{t+1}|\mathbf{Q}_{t+1}\}\prod_{i=1}^{n_{\mathrm{Cmax}}}\Pr.\{Q_{i,t+1},P_{\mathrm{Ct},i,t+1}|\mathbf{Q}_t,\mathbf{H}_t,\mathbf{a}_t\}\\ &\quad\times\prod_{j=1}^{n_{\mathrm{Dmax}}}\Pr.\{Q_{j,t+1},P_{\mathrm{Ct},j,t+1},P_{\mathrm{Dd},j,t+1}|\mathbf{Q}_t,\mathbf{H}_t,\mathbf{a}_t\}. \quad (2)\end{aligned}$$

We assume that the time slot duration τ is substantially smaller than the average flow inter-arrival time as well as the average flow service time. There is a flow departure at the t-th slot if the remaining service time of a flow is less than the current slot duration. By the memoryless property of the exponential distribution, the remaining flow length at any slot t is also exponentially distributed. Thus, a Bernoulli random variable $\underline{B}_{\mathrm{C},i,t}$ denotes the departure of cellular flow $i \in \{1,\dots,n_{\mathrm{Cmax}}\}$ at time slot t, with

$$\underline{B}_{\mathrm{C},i,t} = \begin{cases} 1, & \text{w.p. } 1-\exp(-\frac{r_{\mathrm{C},i0,t}}{\sigma_C}) \\ 0, & \text{w.p. } \exp(-\frac{r_{\mathrm{C},i0,t}}{\sigma_C}). \end{cases}$$

A Bernoulli random variable $\underline{B}_{\mathrm{D},j,t}$ denotes the departure of D2D flow $j \in \{1,\dots,n_{\mathrm{Dmax}}\}$ at time slot t with

$$\underline{B}_{\mathrm{D},j,t} = \begin{cases} 1, & \text{w.p. } 1-\exp(-\frac{r_{\mathrm{D},j0,t}}{\sigma_D}) \text{ if } y_j = 0 \\ & \text{w.p. } 1-\exp(-\frac{r_{\mathrm{D},jj,t}}{\sigma_D}) \text{ if } y_j = 1 \\ 0, & \text{w.p. } \exp(-\frac{r_{\mathrm{D},j0,t}}{\sigma_D}) \text{ if } y_j = 0 \\ & \text{w.p. } \exp(-\frac{r_{\mathrm{D},jj,t}}{\sigma_D}) \text{ if } y_j = 1. \end{cases}$$

Since a newly arrived flow is randomly dispatched to an empty virtual queue, the probability of a cellular flow (resp. a D2D flow) arrival event at an empty virtual queue can be denoted by a Bernoulli random variable $\overline{B}_{\mathrm{C},i,t}$ (resp. $\overline{B}_{\mathrm{D},j,t}$)

with

$$\overline{B}_{\mathrm{C},i,t} = \begin{cases} 1, & \text{w.p. } \dfrac{\lambda_{\mathrm{C}}}{\sum_{i=1}^{n_{\mathrm{Cmax}}}(1 - Q_{\mathrm{C},i,t})} \\ 0, & \text{w.p. } 1 - \dfrac{\lambda_{\mathrm{C}}}{\sum_{i=1}^{n_{\mathrm{Cmax}}}(1 - Q_{\mathrm{C},i,t})}, \end{cases}$$

$$\overline{B}_{\mathrm{D},j,t} = \begin{cases} 1, & \text{w.p. } \dfrac{\lambda_{\mathrm{D}}}{\sum_{j=1}^{n_{\mathrm{Dmax}}}(1 - Q_{\mathrm{D},j,t})} \\ 0, & \text{w.p. } 1 - \dfrac{\lambda_{\mathrm{D}}}{\sum_{j=1}^{n_{\mathrm{Dmax}}}(1 - Q_{\mathrm{D},j,t})}. \end{cases}$$

Based on the above discussion, the queue state transition probability in (2) can be derived as

$$\begin{aligned} &\Pr.\{Q_{i,t+1}, P_{\mathrm{Ct},i,t+1} | \mathbf{H}_t, \mathbf{Q}_t, \mathbf{a}_t\} \\ &= \begin{cases} \Pr.\{\underline{B}_{\mathrm{C},i,t} = 1\}, \\ \text{if } Q_{\mathrm{C},i,t} = 1, P_{\mathrm{Ct},i,t} \in \{P^{(k)}\}_{k=1}^{K} \\ \ \text{and } Q_{i,t+1} = 0, P_{\mathrm{Ct},i,t+1} = 0 \\ \Pr.\{\overline{B}_{\mathrm{C},i,t} = 1\} p_{\mathrm{Ct},k}, \\ \text{if } Q_{\mathrm{C},i,t} = 0, P_{\mathrm{Ct},i,t} = 0 \\ \ \text{and } Q_{i,t+1} = 1, P_{\mathrm{Ct},i,t+1} = P^{(k)}, \forall k \in \{1, \ldots, K\} \\ \Pr.\{\underline{B}_{\mathrm{C},i,t} = 0\}, \\ \text{if } Q_{\mathrm{C},i,t} = Q_{i,t+1} = 1, P_{\mathrm{Ct},i,t} = P_{\mathrm{Ct},i,t+1} \\ \Pr.\{\overline{B}_{\mathrm{C},i,t} = 0\}, \\ \text{if } Q_{\mathrm{C},i,t} = Q_{i,t+1} = 0, P_{\mathrm{Ct},i,t} = P_{\mathrm{Ct},i,t+1} = 0. \end{cases} \end{aligned} \tag{3}$$

$$\begin{aligned} &\Pr.\{Q_{j,t+1}, P_{\mathrm{Ct},j,t+1}, P_{\mathrm{Dd},j,t+1} | \mathbf{H}_t, \mathbf{Q}_t, \mathbf{a}_t\} \\ &= \begin{cases} \Pr.\{\underline{B}_{\mathrm{D},j,t} = 1\}, \\ \text{if } Q_{\mathrm{D},j,t} = 1, P_{\mathrm{Ct},j,t}, P_{\mathrm{Dd},j,t} \in \{P^{(k)}\}_{k=1}^{K} \\ \ \text{and } Q_{i,t+1} = 0, P_{\mathrm{Ct},i,t+1} = 0 \\ \Pr.\{\overline{B}_{\mathrm{D},j,t} = 1\} p_{\mathrm{Ct},k} p_{\mathrm{Dd},k'}, \\ \text{if } Q_{\mathrm{D},j,t} = 0, P_{\mathrm{Ct},j,t} = P_{\mathrm{Dd},j,t} = 0 \\ \ \text{and } Q_{j,t+1} = 1, P_{\mathrm{Ct},j,t+1} = P^{(k)} \\ P_{\mathrm{Dd},j,t+1} = P^{(k')}, \forall k, k' \in \{1, \ldots, K\} \\ \Pr.\{\underline{B}_{\mathrm{D},j,t} = 0\}, \\ \text{if } Q_{\mathrm{D},j,t} = Q_{j,t+1} = 1, P_{\mathrm{Ct},j,t} = P_{\mathrm{Ct},j,t+1}, \\ P_{\mathrm{Dd},j,t} = P_{\mathrm{Dd},j,t+1} \\ \Pr.\{\overline{B}_{\mathrm{D},j,t} = 0\}, \\ \text{if } Q_{\mathrm{D},j,t} = Q_{j,t+1} = 0, P_{\mathrm{Ct},j,t} = P_{\mathrm{Ct},j,t+1}, \\ P_{\mathrm{Dd},j,t} = P_{\mathrm{Dd},j,t+1} = 0. \end{cases} \end{aligned} \tag{4}$$

where $p_{\mathrm{Ct},k}$ and $p_{\mathrm{Dd},k}$ are the probabilities that a cellular transmitter and a D2D transmitter have transmission power $P^{(k)}$, respectively, where $p_{\mathrm{Ct},k} = \frac{d_k^2 - d_{k-1}^2}{R^2}$ and $p_{\mathrm{Dd},k} = \int_{d_{k-1}}^{d_k} f_{d_{\mathrm{D}}}(x)$.

D. REWARD FUNCTION

Under the flow-level model, we define a new performance metric, i.e., the mean energy consumption for the transmission of a flow, which equals the product of the mean delay for the transmission of a flow and the transmission power. The mode selection policy will impact the mean delay through its impact to the traffic load of cellular and D2D communications, while it will also impact the transmission power depending on the UE locations. The mode selection policy should consider both factors and also be jointly optimized with the subchannel allocation policy, which also impacts the mean delay. We consider the low load regime where the blocking probability is negligible.

We are interested in the weighted sum of the mean energy consumption for the transmission of a cellular flow and a D2D flow, which is given by

$$\begin{aligned} \bar{U} &= \mathbf{E}^{\pi(\Omega)}\left[g(\mathbf{S}, \Omega(\mathbf{S}))\right], \\ g(\mathbf{S}, \Omega(\mathbf{S})) &= \omega_{\mathrm{C}} \sum_{i=1}^{n_{\mathrm{Cmax}}} \frac{Q_{\mathrm{C},i} P_{\mathrm{Ct},i}}{\lambda_{\mathrm{C}}} \\ &\quad + \omega_{\mathrm{D}} \sum_{j=1}^{n_{\mathrm{Dmax}}} \frac{(Q_{\mathrm{D},j} - y_j) P_{\mathrm{Ct},j} + y_j P_{\mathrm{Dd},j}}{\lambda_{\mathrm{D}}}, \end{aligned} \tag{5}$$

where ω_{C} and ω_{D} represent the relative importance of the cellular flows and D2D flows, and $\mathbf{E}^{\pi(\Omega)}[x]$ denotes the expectation operation taken w.r.t. the unique steady-state distribution induced by the given policy Ω. The proof of the derivation of $\bar{U}$ is given in Appendix A.

Our objective is to design of the optimal mode selection and resource allocation control policy to minimize $\bar{U}$. Using the MDP formalism, the optimization problem is formulated as the infinite horizon MDP problem

$$\begin{aligned} \min \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{E}^{\Omega}[g(\mathbf{S}_t, \Omega(\mathbf{S}_t))] \\ = \min \mathbf{E}^{\pi(\Omega)}[g(\mathbf{S}, \Omega(\mathbf{S}))]. \end{aligned} \tag{6}$$

where the equality holds under any unichain policy. Therefore, the reward function $g(\mathbf{S}_t, \Omega(\mathbf{S}_t))$ can be derived directly from $g(\mathbf{S}, \Omega(\mathbf{S}))$ given in (5). Note that the reward function is only dependent on the QSI $\mathbf{Q}_t$ and the mode selection action $\Omega_{\mathrm{y}}(\mathbf{Q}_t)$. Therefore, we will use the notation $g(\mathbf{S}_t, \Omega(\mathbf{S}_t))$ and $g(\mathbf{Q}_t, \Omega_{\mathrm{y}}(\mathbf{Q}_t))$ interchangeably in the rest of the paper.

V. OPTIMAL SOLUTION

The formulated problem is a classical infinite horizon average reward MDP problem, which can be solved by the Bellman's equation.

$$\begin{aligned} \theta + V(\mathbf{S}) = \min_{\Omega(\mathbf{S})} \Bigg\{ g(\mathbf{S}, \Omega(\mathbf{S})) + \sum_{\mathbf{S}' \in \mathcal{S}} \Pr.[\mathbf{S}' | \mathbf{S}, \\ \Omega(\mathbf{S})] V(\mathbf{S}') \Bigg\}, \quad \forall \mathbf{S} \in \mathcal{S}, \end{aligned} \tag{7}$$

where $V(\mathbf{S})$ is the value function representing the average reward obtained following policy Ω from each state $\mathbf{S}$, while θ represents the optimal average reward per-period for a system in steady-state.

As a remark, note that the Bellman's equation (7) represents a series of fixed-point equations, where the number of equations are determined by the number of value functions $V(\mathbf{S})$, which is $|\mathbf{S}|$. Theoretically, the BS can use the brute force value iteration method to offline solve (7) and derive the optimal control policy, in which $|\mathbf{S}|$ value functions need to be stored and the computation complexity is $O(|\mathcal{S}|^2|\mathcal{A}|)$ in one iteration. Therefore, the offline value iteration algorithm is too complicated to compute due to curse of dimensionality.

A. EQUIVALENT BELLMAN's EQUATION

In order to reduce the state space of the above MDP, we construct an equivalent Bellman's equation. We first define the partitioned actions of a policy Ω as follows.

Definition 1 (Definition of Partitioned Actions): Given a control policy Ω, we define

$$\Omega(\mathbf{Q}) = \{\Omega(\mathbf{Q},\mathbf{H})|\forall\mathbf{H}\} \subseteq \mathcal{A}$$

as the collection of $|\mathcal{H}|$ actions, where every action is mapped by policy Ω from a system state with given QSI $\mathbf{Q}$, and a different realization of CSI $\mathbf{H} \in \mathcal{H}$.

Lemma 1: The control policy obtained by solving the original Bellman's equation (7) *is equivalent to the control policy obtained by solving the reduced-state Bellman's equation* (8)

$$\theta + V(\mathbf{Q}) = \min_{\Omega(\mathbf{Q})}\Bigg\{g(\mathbf{Q},\Omega_{\mathrm{y}}(\mathbf{Q})) + \sum_{\mathbf{Q}'\in\mathcal{Q}}\Pr.[\mathbf{Q}'|\mathbf{Q}, \Omega(\mathbf{Q})]V(\mathbf{Q}')\Bigg\}, \quad \forall\mathbf{Q}\in\mathcal{Q}, \tag{8}$$

where $V(\mathbf{Q}) = \mathbf{E}_{\mathbf{H}}\Big[V(\mathbf{Q},\mathbf{H})|\mathbf{Q}\Big] = \sum_{\mathbf{H}\in\mathcal{H}}\Pr.[\mathbf{H}|\mathbf{Q}]V(\mathbf{Q},\mathbf{H})$ *is the conditional expectation of value function $V(\mathbf{S})$ taken over the channel state space $\mathcal{H}$, while* $g(\mathbf{Q},\Omega_{\mathrm{y}}(\mathbf{Q})) = \mathbf{E}_{\mathbf{H}}\Big[g(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H}))|\mathbf{Q}\Big]$ *and* $\Pr.[\mathbf{Q}'|\mathbf{Q},\Omega(\mathbf{Q})] = \mathbf{E}_{\mathbf{H}}\Big[\Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})|\mathbf{Q}]\Big]$ *are the conditional expectations of reward function* $g(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H}))$ *and transition probability* $\Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})]$ *taken over the channel state space $\mathcal{H}$.*

Proof: Please refer to Appendix B. ■

As a remark, note that the equivalent Bellman's equation (8) represents a series of fixed-point equations, where the numbers of equations are determined by the possible values of value functions $V(\mathbf{Q})$, which is $|\mathcal{Q}|$. Therefore, we only need to solve $|\mathcal{Q}|$ instead of $|\mathcal{H}| \times |\mathcal{Q}|$ fixed-point equations with the reduced-state Bellman's equation (8). In order to solve one such fixed-point equation using value iteration, the R.H.S. of (8) has to be minimized with given value functions $V(\mathbf{Q}')$. For this purpose, the R.H.S. of (8) can be written as

$$\min_{\Omega(\mathbf{Q})}\sum_{\mathbf{H}\in\mathcal{H}}\Pr.[\mathbf{H}|\mathbf{Q}]f(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})), \tag{9}$$

where

$$f(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})) = g(\mathbf{Q},\Omega_{\mathrm{y}}(\mathbf{Q})) + \sum_{\mathbf{Q}'\in\mathcal{Q}}\Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})]V(\mathbf{Q}'). \tag{10}$$

Since (9) is a decoupled objective function w.r.t. different CSI realizations $\mathbf{H}$ with a given QSI $\mathbf{Q}$, we need to obtain $|\mathcal{H}|$ optimal actions in order to achieve the minimization objective in the R.H.S. of (8), where every optimal action is w.r.t. a system state $(\mathbf{H},\mathbf{Q})$ with given $\mathbf{Q}$ and a different CSI realization $\mathbf{H} \in \mathcal{H}$ that minimizes the value of $f(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H}))$. This means that the control policy obtained by solving (8) is based on the system state $\mathbf{S}$ instead of only the QSI $\mathbf{Q}$.

B. Q-FACTOR

To facilitate the mode selection control which is only adaptive to $\mathbf{Q}$, we introduce the mode selection control Q-factor $\mathbb{Q}(\mathbf{Q},\mathbf{y})$ as

$$\theta + \mathbb{Q}(\mathbf{Q},\mathbf{y}) = \min_{\Omega_{\mathrm{x}}(\mathbf{Q})}\Bigg\{g(\mathbf{Q},\mathbf{y}) + \sum_{\mathbf{Q}'\in\mathcal{Q}}\Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{y}, \Omega_{\mathrm{x}}(\mathbf{Q})]V(\mathbf{Q}')\Bigg\}, \quad \forall\mathbf{Q}\in\mathcal{Q}, \forall\mathbf{y}\in\mathcal{A}_{\mathrm{y}}. \tag{11}$$

where $\mathbf{y}$ is an arbitrary action in state space $\mathcal{A}_{\mathrm{y}}$. Therefore, according to (8), we have

$$V(\mathbf{Q}) = \min_{\mathbf{y}\in\mathcal{A}_{\mathrm{y}}}\mathbb{Q}(\mathbf{Q},\mathbf{y}), \quad \forall\mathbf{Q}\in\mathcal{Q} \tag{12}$$

and $\mathbb{Q}(\mathbf{Q},\mathbf{y})$ satisfies the following "Q-factor form" of the Bellman's equation

$$\theta + \mathbb{Q}(\mathbf{Q},\mathbf{y}) = \min_{\Omega_{\mathrm{x}}(\mathbf{Q})}\Bigg\{g(\mathbf{Q},\mathbf{y}) + \sum_{\mathbf{Q}'\in\mathcal{Q}}\Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{y}, \Omega_{\mathrm{x}}(\mathbf{Q})]\min_{\mathbf{y}'\in\mathcal{A}_{\mathrm{y}}}\mathbb{Q}(\mathbf{Q}',\mathbf{y}')\Bigg\}, \quad \forall\mathbf{Q}\in\mathcal{Q}, \forall\mathbf{y}\in\mathcal{A}_{\mathrm{y}}. \tag{13}$$

Moreover, the optimal mode selection control is given by

$$\Omega^*_{\mathrm{y}}(\mathbf{Q}) = \arg\min_{\mathbf{y}\in\mathcal{A}_{\mathrm{y}}}\mathbb{Q}(\mathbf{Q},\mathbf{y})\ \forall\mathbf{Q}\in\mathcal{Q}. \tag{14}$$

Given $\Omega^*_{\mathrm{y}}(\mathbf{Q})$, the optimal subchannel allocation control $\Omega^*_{\mathrm{x}}(\mathbf{Q},\mathbf{H})$ can be derived as

$$\Omega^*_{\mathrm{x}}(\mathbf{Q},\mathbf{H}) = \arg\min_{\mathbf{x}\in\mathcal{A}_{\mathrm{x}}} f(\mathbf{Q},\mathbf{H},\Omega^*_{\mathrm{y}}(\mathbf{Q}),\mathbf{x}), \tag{15}$$

where $f(\mathbf{Q},\mathbf{H},\Omega^*_{\mathrm{y}}(\mathbf{Q}),\mathbf{x})$ is obtained by taking $\Omega^*_{\mathrm{y}}(\mathbf{Q})$ into the R.H.S of (10), and replacing $V(\mathbf{Q}')$ with $\mathbb{Q}(\mathbf{Q},\Omega^*_{\mathrm{y}}(\mathbf{Q}))$.

As a remark, by the two-timescale requirement, the mode selection control policy is defined on the partial system state $\mathbf{Q}$, while the resource allocation control policy is defined on the complete system state $\mathbf{S} = (\mathbf{Q},\mathbf{H})$. We need to solve $|\mathcal{Q}| \times |\mathcal{A}_{\mathrm{y}}|$ fixed-point equations with the Q-factor structure, which still faces the curse of dimensionality problem. Thus, we will develop a solution with reduced complexity using linear value approximation and online stochastic learning in the following sections.

C. LINEAR APPROXIMATION ON MODE SELECTION Q-FACTOR

In this section, we use linear approximation method to further reduce the state space and facilitate distributed implementation. We first introduce a randomized base policy, under which the Q-factor satisfying (13) can be decomposed into the additive form, i.e.,

$$\mathbb{Q}(\mathbf{Q},\mathbf{y}) \approx \sum_{i=1}^{n_{\mathrm{Cmax}}} \mathbb{Q}_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) + \sum_{j=1}^{n_{\mathrm{Dmax}}} \mathbb{Q}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j). \tag{16}$$

Definition 2 (Randomized Base Policy): A randomized base policy is denoted as $\hat{\Omega} = (\hat{\Omega}_{\mathrm{y}}, \hat{\Omega}_{\mathrm{x}})$. A randomized base policy for mode selection $\hat{\Omega}_{\mathrm{y}}$ is given by a distribution on the action space of $\mathbf{y}$, i.e., $\mathcal{A}_{\mathrm{y}}$. A randomized base policy for subchannel allocation policy is given by a mapping from the CSI $\mathbf{H}$ to a probability distribution $\hat{\Omega}_{\mathrm{x}}(\mathbf{H})$ on the action space of $\mathbf{x}$, i.e., $\mathcal{A}_{\mathrm{x}}$.

Under a randomized base policy $\hat{\Omega}$, the corresponding Q-factors have the following decomposition structure.

Lemma 2: Given any randomized base policy $\hat{\Omega}$, the mode selection Q-factor $\hat{\mathbb{Q}}(\mathbf{Q},\mathbf{y})$ can be decomposed into the sum of per-queue mode selection Q-factors $\hat{\mathbb{Q}}_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})$ and $\hat{\mathbb{Q}}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)$ as given in (16), where the per-queue Q-factors satisfy the following fixed point equations for each queue $i \in \{1, \ldots, n_{\mathrm{Cmax}}\}$ and $j \in \{1, \ldots, n_{\mathrm{Dmax}}\}$, respectively:

$$\begin{aligned}
&\theta_{\mathrm{C},i} + \hat{\mathbb{Q}}_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) \\
&= g_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) + \sum_{Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i}} \hat{\Pr} \\
&\quad \cdot [Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i} | Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}] \hat{V}_{\mathrm{C},i}(Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i}),
\end{aligned} \tag{17}$$

$$\begin{aligned}
&\theta_{\mathrm{D},j} + \hat{\mathbb{Q}}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j) \\
&= g_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j) \\
&\quad + \sum_{Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j}} \hat{\Pr}.[Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j} | Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j] \\
&\quad \hat{V}_{\mathrm{D},j}(Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j}),
\end{aligned} \tag{18}$$

where

$$g_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) = \omega_{\mathrm{C}} \frac{Q_{\mathrm{C},i} P_{\mathrm{Ct},i}}{\lambda_{\mathrm{C}}},$$

$$g_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j) = \omega_{\mathrm{D}} \frac{(Q_{\mathrm{D},j} - y_j) P_{\mathrm{Ct},j} + y_j P_{\mathrm{Dd},j}}{\lambda_{\mathrm{D}}},$$

$$\begin{aligned}
&\hat{\Pr}.[Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i} | Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}] = \mathbf{E}^{\hat{\Omega}_{\mathrm{y}}} \\
&\quad \times [\mathbf{E}_{\mathbf{H}}[\mathbf{E}^{\hat{\Omega}_{\mathrm{x}}}[\Pr.[Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i} | Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}, \mathbf{H}, \mathbf{y}, \mathbf{x}] | \mathbf{H}, \mathbf{y}]]],
\end{aligned}$$

$$\begin{aligned}
&\hat{\Pr}.[Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j} | y_j] = \mathbf{E}^{\hat{\Omega}_{\mathrm{y}}}[\mathbf{E}_{\mathbf{H}}[\mathbf{E}^{\hat{\Omega}_{\mathrm{x}}}[\Pr. \\
&[Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j} | Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, \mathbf{H}, \mathbf{y}, \mathbf{x}] | \mathbf{H}, \mathbf{y}]]],
\end{aligned}$$

$$\hat{V}_{\mathrm{C},i}(Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i}) = \hat{\mathbb{Q}}_{\mathrm{C},i}(Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i}),$$

$$\hat{V}_{\mathrm{D},j}(Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j}) = \mathbf{E}^{\hat{\Omega}_{\mathrm{y}}}[\hat{\mathbb{Q}}_{\mathrm{D},j}(Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j}, y'_j)].$$

Proof: Please refer to Appendix C. ■

Note that under the randomized base policy $\hat{\Omega}_{\mathrm{y}}$, the value function can also be written in the decomposed form, i.e.,

$$\begin{aligned}
\hat{V}(\mathbf{Q}) = \hat{\mathbb{Q}}(\mathbf{Q}, \hat{\Omega}_{\mathrm{y}}(\mathbf{Q})) &\approx \sum_{i=1}^{n_{\mathrm{Cmax}}} \hat{V}_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) \\
&+ \sum_{j=1}^{n_{\mathrm{Dmax}}} \hat{V}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}).
\end{aligned} \tag{19}$$

where the approximate equality is by (16) and the definition of $\hat{V}_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})$ and $\hat{V}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j})$ are given in Lemma 2.

Based on the randomized base policy $\hat{\Omega}$, we shall obtain a low complexity deterministic policy $\hat{\Omega}^*$ by Q-factor approximation (16). In this section, we first assume we could obtain the per-queue Q-factors $\{\hat{\mathbb{Q}}_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})\}_{i=1}^{n_{\mathrm{Cmax}}}$ and $\{\hat{\mathbb{Q}}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}_{i=1}^{n_{\mathrm{Dmax}}}$ via some means (e.g., via offline value iteration) and focus on deriving the optimal action under every system state. The solution is elaborated below.

1) MODE SELECTION POLICY

According to (14), the mode selection control is given by

$$\begin{aligned}
\hat{\Omega}^*_{\mathrm{y}}(\mathbf{Q}) = \arg\min_{\mathbf{y} \in \mathcal{A}_{\mathrm{y}}} \Bigg[&\sum_{i=1}^{n_{\mathrm{Cmax}}} \hat{\mathbb{Q}}_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) \\
&+ \sum_{j=1}^{n_{\mathrm{Dmax}}} \hat{\mathbb{Q}}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j) \Bigg],
\end{aligned} \tag{20}$$

which is equivalent to

$$\begin{aligned}
y_j^* &= \arg\min_{y_j \in \{0,1\}} \hat{\mathbb{Q}}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j), \\
&\quad \forall j \text{ with } Q_{\mathrm{D},j} = 1.
\end{aligned} \tag{21}$$

2) SUBCHANNEL ALLOCATION POLICY

Since the duration of the time slot is relatively short with respect to the size of the flows (e.g., the minimum scheduling time unit is 1ms in 3G LTE, while it usually takes at least several seconds to transmit a flow), we have $\frac{r_{\mathrm{C},i0}}{\sigma_{\mathrm{C}}} \ll 1$, $\frac{r_{\mathrm{D},j0}}{\sigma_{\mathrm{D}}} \ll 1$, $\frac{r_{\mathrm{D},jj}}{\sigma_{\mathrm{D}}} \ll 1$. Therefore, the probability of a cellular flow i (resp. a D2D flow j in D2D mode or cellular mode) departs in time slot t approximately equals $1 - \exp(-\frac{r_{\mathrm{C},i0,t}}{\sigma_C}) \approx \frac{r_{\mathrm{C},i0,t}}{\sigma_C}$ (resp. $1 - \exp(-\frac{r_{\mathrm{D},j0,t}}{\sigma_D}) \approx \frac{r_{\mathrm{D},j0,t}}{\sigma_D}$ or $1 - \exp(-\frac{r_{\mathrm{D},jj,t}}{\sigma_D}) \approx \frac{r_{\mathrm{D},jj,t}}{\sigma_D}$). Under the mode selection action $\mathbf{y}^*$ obtained by (21), the subchannel allocation policy can be obtained by the R.H.S. of (15), and replaces $V(\mathbf{Q}')$ by the decomposed form of $\hat{V}(\mathbf{Q}')$ as given in (19).

$$\begin{aligned}
&\hat{\Omega}^*_{\mathrm{x}}(\mathbf{Q}, \mathbf{H}) \\
&= \min_{\Omega_{\mathrm{x}}(\mathbf{Q},\mathbf{H})} \Bigg\{ g(\mathbf{Q}, \mathbf{y}^*) + \sum_{\mathbf{Q}' \in \mathcal{Q}} \Pr.[\mathbf{Q}' | \mathbf{Q}, \mathbf{H}, \mathbf{y}^*, \Omega_{\mathrm{x}}(\mathbf{Q}, \mathbf{H})] \\
&\quad \times \Bigg[\sum_{i=1}^{n_{\mathrm{Cmax}}} \hat{V}_{\mathrm{C},i}(Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i}) + \sum_{j=1}^{n_{\mathrm{Dmax}}} \hat{V}_{\mathrm{D},j}(Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j}) \Bigg] \Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&= \arg\min_{\Omega_{\mathrm{x}}(\mathbf{Q},\mathbf{H})} \\
&\quad\times \Bigg\{ \sum_{i=1}^{n_{\mathrm{Cmax}}} \sum_{Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i}} \Pr.[Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i} | \mathbf{Q}, \mathbf{H}, \mathbf{y}^*, \Omega_{\mathrm{x}}(\mathbf{Q}, \mathbf{H})] \\
&\quad\times \hat{V}_{\mathrm{C},i}(Q'_{\mathrm{C},i}, P'_{\mathrm{Ct},i}) + \sum_{j=1}^{n_{\mathrm{Dmax}}} \sum_{Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j}} \Pr. \\
&\quad\times [Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j} | \mathbf{Q}, \mathbf{H}, \mathbf{y}^*, \Omega_{\mathrm{x}}(\mathbf{Q}, \mathbf{H})] \\
&\quad\times \hat{V}_{\mathrm{D},j}(Q'_{\mathrm{D},j}, P'_{\mathrm{Ct},j}, P'_{\mathrm{Dd},j}) \Bigg\} \\
&= \arg\max_{\mathbf{x}\in\mathcal{A}_{\mathrm{x}}} \Bigg\{ \sum_{m=1}^{\eta N_{\mathrm{F}}} \Bigg[\sum_{i=1}^{n_{\mathrm{Cmax}}} \underbrace{\frac{r_{\mathrm{C},i0}^{(m)} x_{\mathrm{C},i0}^{(m)}}{\sigma_{\mathrm{C}}} \Delta\hat{V}_{\mathrm{C},i}(1, P_{\mathrm{Ct},i})}_{W_{\mathrm{C},i0}^{(m)}} \\
&\quad + \sum_{j=1}^{n_{\mathrm{Dmax}}} \underbrace{\frac{r_{\mathrm{D},j0}^{(m)} x_{\mathrm{D},j0}^{(m)} (1 - y_j^*)}{\sigma_{\mathrm{D}}} \Delta\hat{V}_{\mathrm{D},j}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j})}_{W_{\mathrm{D},j0}^{(m)}} \Bigg] \\
&\quad + \sum_{n=\eta N_{\mathrm{F}}+1}^{N_{\mathrm{F}}} \Bigg[\sum_{j=1}^{n_{\mathrm{Dmax}}} \underbrace{\frac{r_{\mathrm{D},jj}^{(m)} x_{\mathrm{D},jj}^{(m)} y_j^*}{\sigma_{\mathrm{D}}} \Delta\hat{V}_{\mathrm{D},j}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j})}_{W_{\mathrm{D},jj}^{(m)}} \Bigg] \Bigg\}.
\end{aligned}
\tag{22}
$$

where $\Delta\hat{V}_{\mathrm{C},i}(1, P_{\mathrm{Ct},i}) = \hat{V}_{\mathrm{C},i}(1, P_{\mathrm{Ct},i}) - \hat{V}_{\mathrm{C},i}(0, 0)$ and $\Delta\hat{V}_{\mathrm{D},j}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}) = \hat{V}_{\mathrm{D},j}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}) - \hat{V}_{\mathrm{D},j}(0, 0, 0)$. Recall that for any $m \in \{1, \ldots, \eta N_{\mathrm{F}}\}$ we have $\sum_{i=1}^{n_{\mathrm{Cmax}}} x_{\mathrm{C},i0,t}^{(m)} + \sum_{j=1}^{n_{\mathrm{Dmax}}} x_{\mathrm{D},j0,t}^{(m)}(1 - y_{j,t}) = 1$; while for any $m \in \{\eta N_{\mathrm{F}} + 1, \ldots, N_{\mathrm{F}}\}$ we have $\sum_{j=1}^{n_{\mathrm{Dmax}}} x_{\mathrm{D},jj,t}^{(m)} y_{j,t} = 1$. Therefore, the problem becomes determining the largest element within set $\{W_{\mathrm{C},i0}^{(m)}\}_{i\in\{1,\ldots,n_{\mathrm{Cmax}}\}} \bigcup \{W_{\mathrm{D},j0}^{(m)}\}_{j\in\{1,\ldots,n_{\mathrm{Dmax}}\}}$ on every subchannel for cellular communications and the largest element within set $\{W_{\mathrm{D},jj}^{(m)}\}_{j\in\{1,\ldots,n_{\mathrm{Dmax}}\}}$ on every subchannel for D2D communications, i.e.,

$$
x_{\mathrm{C},i0}^{(m)} = \begin{cases} 1, & \text{if } i = \arg\max_{i'} W_{\mathrm{C},i'0}^{(m)} \\ & \text{and } \max_{i'} W_{\mathrm{C},i'0}^{(m)} \geq \max_{j'} W_{\mathrm{D},j'0}^{(m)}, \\ 0, & \text{otherwise} \end{cases}
$$
$$
\forall i = 1, \ldots, n_{\mathrm{Cmax}}, m = 1, \ldots, \eta N_{\mathrm{F}}. \tag{23}
$$

$$
x_{\mathrm{D},j0}^{(m)} = \begin{cases} 1, & \text{if } j = \arg\max_{j'} W_{\mathrm{D},j'0}^{(m)} \\ & \text{and } \max_{j'} W_{\mathrm{D},j'0}^{(m)} \geq \max_{i'} W_{\mathrm{C},i'0}^{(m)}, \\ 0, & \text{otherwise} \end{cases}
$$
$$
\forall j = 1, \ldots, n_{\mathrm{Dmax}}, m = 1, \ldots, \eta N_{\mathrm{F}}. \tag{24}
$$

$$
x_{\mathrm{D},jj}^{(m)} = \begin{cases} 1, & \text{if } j = \arg\max_{j'} W_{\mathrm{D},j'j'}^{(m)} \\ 0, & \text{otherwise} \end{cases}
$$
$$
\forall j = 1, \ldots, n_{\mathrm{Dmax}}, m = \eta N_{\mathrm{F}} + 1, \ldots, N_{\mathrm{F}}. \tag{25}
$$

The proposed control policy $\hat{\Omega}_{\mathrm{y}}^*(\mathbf{Q})$ and $\hat{\Omega}_{\mathrm{x}}^*(\mathbf{Q}, \mathbf{H})$ can be implemented in a distributed fashion as follows:

- Step i (**Storage and transfer of per-queue Q-factors**): The BS stores all the per-queue Q-factors. When a new cellular flow $i \in \{1, \ldots, n_{\mathrm{Cmax}}\}$ (resp. D2D flow $j \in \{1, \ldots, n_{\mathrm{Dmax}}\}$) arrives at the system, the BS sends the per-queue Q-factors $\hat{\mathbb{Q}}_{\mathrm{C},i}(1, P_{\mathrm{Ct},i})$, (resp. $\{\hat{\mathbb{Q}}_{\mathrm{D},j}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}$ to CUE i (resp. src. DUE j), where $P_{\mathrm{Ct},i}$ (resp. $P_{\mathrm{Ct},j}$, $P_{\mathrm{Dd},j}$) can be determined by the UE according to its measured Reference Signal Received Power (RSRP) in LTE system and reported to the BS. Note that the transmission power remains unchanged during the transmission of the flow.
- Step ii (**Mode selection policy**): The mode selection is performed distributively. Each src. DUE only needs to determine its mode by choosing the larger per-queue Q-factor between two candidates based on its local QSI according to (21).
- Step iii (**Subchannel allocation policy**): Each CUE i (resp. src. DUE j (based on its mode selection action)) distributively calculates its bid $\{W_{\mathrm{C},i0}^{(m)}\}_{m=1}^{\eta N_{\mathrm{F}}}$ (resp. $\{W_{\mathrm{D},j0}^{(m)}\}_{m=1}^{\eta N_{\mathrm{F}}}$ if $y_j = 0$ and $\{W_{\mathrm{D},jj}^{(m)}\}_{m=\eta N_{\mathrm{F}}+1}^{N_{\mathrm{F}}}$ if $y_j = 1$) according to (22), and submits the bid to the BS. The BS determines the optimal subchannel allocation action according to (23), (24), and (25).

Remark 2 (Complexity Reduction Due to the Linear Approximation): We can represent the $|\mathcal{Q}| \times |\mathcal{A}_{\mathrm{y}}|$ global Q-factors with $n_{\mathrm{Cmax}}(K + 1) + n_{\mathrm{Dmax}}(2K^2 + 1)$ per-queue Q-factors by the linear approximation architecture, which greatly reduce the storage capacity at the BS. Moreover, each CUE only needs to store one local per-queue Q-factor, while each src. DUE only need to store 2 local per-queue Q-factors. The complexity of determining the mode selection action at each src. DUE is 1. The complexity of determining the subchannel allocation action at the BS based on the submitted bid of the UEs is $\mathcal{O}((n_{\mathrm{Cmax}} + n_{\mathrm{Dmax}})N_{\mathrm{F}})$ instead of $|\mathcal{A}_{\mathrm{xC}}| \times |\mathcal{A}_{\mathrm{xD}}|$ without the linear approximation.

The following theorem shows that the proposed distributed policy always achieves better performance than the randomized base policy.

Theorem 1 (Performance Improvement): If $\Pr.[\mathbf{Q}'|\mathbf{Q}, \mathbf{y}, \mathbf{x}_{\mathrm{C}}, \mathbf{x}_{\mathrm{D}}] \neq \Pr.[\mathbf{Q}'|\mathbf{Q}, \mathbf{y}', \mathbf{x}'_{\mathrm{C}}, \mathbf{x}'_{\mathrm{D}}]$ *for any* $(\mathbf{y}, \mathbf{x}_{\mathrm{C}}, \mathbf{x}_{\mathrm{D}}) \neq (\mathbf{y}', \mathbf{x}'_{\mathrm{C}}, \mathbf{x}'_{\mathrm{D}})$ *and* $\mathbf{Q} \in \mathcal{Q}$. *Since the minimum R.H.S of* (15) *can be achieved with proposed solution, we have* $\theta^*(\mathbf{Q}) < \theta$, $\forall \mathbf{Q} \in \mathcal{Q}$, *where* $\theta^*(\mathbf{Q})$ *is the average reward under the proposed solution starting from state* $\mathbf{Q}$ *and* θ *is the average reward under any randomized base policy, respectively.*

D. ONLINE STOCHASTIC LEARNING

1) ONLINE STOCHASTIC LEARNING

The above discussion in Section V.C assumes that the per-queue Q-factors are already known and derive a distributed control policy. However, we still have to determine the per-queue Q-factors. For this purpose, instead of solving the fixed point equation using offline value iteration, we will estimate per-queue Q-factors $\{\hat{\mathbb{Q}}_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})\}_{i=1}^{n_{\mathrm{Cmax}}}$ and $\{\hat{\mathbb{Q}}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}_{j=1}^{n_{\mathrm{Dmax}}}$ using online stochastic learning algorithm on the instantaneous observation.

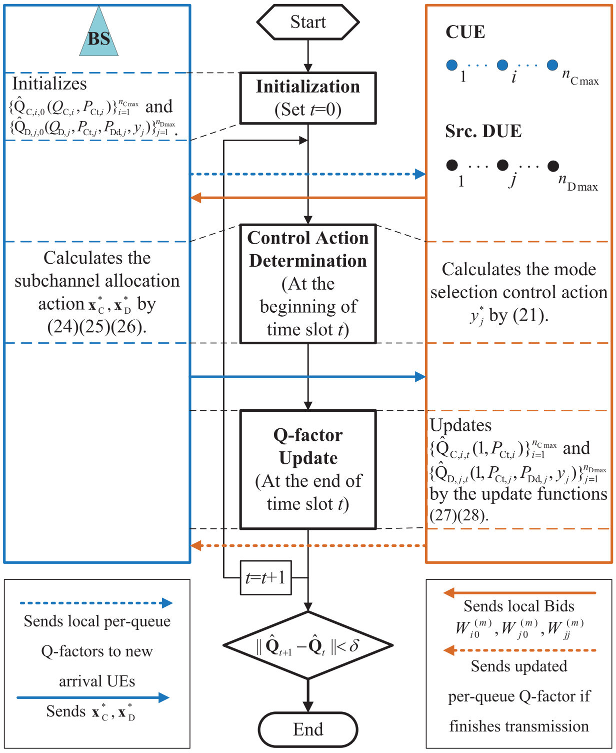
BS
Start
CUE
1
i
n_{Cmax}
Src. DUE
1
j
n_{Dmax}
Initializes $\{\hat{\mathrm{Q}}_{\mathrm{C},i,0}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})\}_{i=1}^{n_{\mathrm{Cmax}}}$ and $\{\hat{\mathrm{Q}}_{\mathrm{D},j,0}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}_{j=1}^{n_{\mathrm{Dmax}}}$.
Initialization (Set t=0)
Calculates the subchannel allocation action $\mathbf{x}_{\mathrm{C}}^*, \mathbf{x}_{\mathrm{D}}^*$ by (24)(25)(26).
Control Action Determination (At the beginning of time slot t)
Calculates the mode selection control action y_j^* by (21).
Q-factor Update (At the end of time slot t)
Updates $\{\hat{\mathrm{Q}}_{\mathrm{C},i,t}(1, P_{\mathrm{Ct},i})\}_{i=1}^{n_{\mathrm{Cmax}}}$ and $\{\hat{\mathrm{Q}}_{\mathrm{D},j,t}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}_{j=1}^{n_{\mathrm{Dmax}}}$ by the update functions (27)(28).
t=t+1
$\| \hat{\mathbf{Q}}_{t+1} - \hat{\mathbf{Q}}_t \| < \delta$
End
Sends local per-queue Q-factors to new arrival UEs
Sends $\mathbf{x}_{\mathrm{C}}^*, \mathbf{x}_{\mathrm{D}}^*$
Sends local Bids $W_{i0}^{(m)}, W_{j0}^{(m)}, W_{jj}^{(m)}$
Sends updated per-queue Q-factor if finishes transmission

FIGURE 1. **The implementation flow of the proposed solution.**

Fig. 1 illustrates the implementation flow of the overall solution with detailed steps as follows:

- Step 1 (**Initialization**): Set t=0. The $\{\hat{\mathbb{Q}}_{\mathrm{C},i,0}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})\}_{i=1}^{n_{\mathrm{Cmax}}}$ and $\{\hat{\mathbb{Q}}_{\mathrm{D},j,0}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}\ , y_j)\}_{j=1}^{n_{\mathrm{Dmax}}}$ are initialized at the BS. The second subscript denotes the index of time slot.
- Step 2 (**Control Action Determination**): At the beginning of time slot t, for any new cellular flow i (resp. D2D flow j) that arrived during the previous time slot $t-1$, the BS sends its local per-queue Q-factors as specified in the Step i of Section V.C. Based on its local per-queue Q-factors, every src. DUE j first calculates the mode selection control action $y_{j,t}^*$ as specified in the Step ii of Section V.C, and then every CUE and src. DUE calculates and submits its bid to the BS, which determines the subchannel allocation action $\mathbf{x}_{\mathrm{C},t}^*$ and $\mathbf{x}_{\mathrm{D},t}^*$ and notifies the corresponding UEs, as specified in the Step iii of Section V.C.
- Step 3 (**Q-factor Update**): At the end of time slot t, the per-queue Q-factors $\{\hat{\mathbb{Q}}_{\mathrm{C},i,t}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})\}_{i=1}^{n_{\mathrm{Cmax}}}$ and $\{\hat{\mathbb{Q}}_{\mathrm{D},j,t}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}_{j=1}^{n_{\mathrm{Dmax}}}$ can be updated to the per-queue Q-factors $\{\hat{\mathbb{Q}}_{i,t+1}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})\}_{i=1}^{n_{\mathrm{Cmax}}}$ and $\{\hat{\mathbb{Q}}_{j,t+1}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}_{j=1}^{n_{\mathrm{Dmax}}}$ using the update functions (26) and (27). Specifically, every CUE (resp. src. DUE) updates its respective local per-queue Q-factor(s) $\hat{\mathbb{Q}}_{\mathrm{C},i}(1, P_{\mathrm{Ct},i})$ (resp. $\{\hat{\mathbb{Q}}_{\mathrm{D},j}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}_{y_j \in \{0,1\}}$) based on the randomized policy. Remarkably, the per-queue Q-factors $\hat{\mathbb{Q}}_{\mathrm{C},i}(0,0,0) = 0$ and $\hat{\mathbb{Q}}_{\mathrm{D},j}(0,0,0,0) = 0$ according to the update functions (26) and (27). If a flow finishes transmission at the end of time slot t, it sends its updated per-queue Q-factors $\hat{\mathbb{Q}}_{\mathrm{C},i}(1, P_{\mathrm{Ct},i,t})$ or $\{\hat{\mathbb{Q}}_{\mathrm{D},j}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\}_{y_j \in \{0,1\}}$ to the BS.

$$\hat{\mathbb{Q}}_{\mathrm{C},i,t+1}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) = \hat{\mathbb{Q}}_{\mathrm{C},i,t}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) + \epsilon_{\tau_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}, t)} \times \Delta\hat{\mathbb{Q}}_{\mathrm{C},i,t}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}), \quad (26)$$

$$\begin{aligned}&\hat{\mathbb{Q}}_{\mathrm{D},j,t+1}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\\ &\quad = \hat{\mathbb{Q}}_{\mathrm{D},j,t}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\\ &\quad\quad + \epsilon_{\tau_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j, t)}\\ &\quad\quad \times \Delta\hat{\mathbb{Q}}_{\mathrm{D},j,t}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j), \end{aligned} \quad (27)$$

where

$$\begin{aligned}&\Delta\hat{\mathbb{Q}}_{\mathrm{C},i,t}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}) = \omega_{\mathrm{C}} \frac{Q_{\mathrm{C},i} P_{\mathrm{Ct},i}}{\lambda_{\mathrm{C}}}\\ &\quad + \mathbf{E}^{\hat{\Omega}_{\mathrm{y}}}[\mathbf{E}_{\mathbf{H}}[\mathbf{E}^{\hat{\Omega}_{\mathrm{x}}}[(1 - \frac{\sum_{m=1}^{\eta N_F} r_{\mathrm{C},i0}^{(m)} x_{\mathrm{C},i0}^{(m)}}{\sigma_{\mathrm{C}}})|\mathbf{H}, \mathbf{y}]]]\\ &\quad \hat{V}_{\mathrm{C},i,t}(1, P_{\mathrm{Ct},i}) - \sum_{k=1}^{K} \frac{\lambda_{\mathrm{C}} p_{\mathrm{Ct},k}}{n_{\mathrm{Cmax}} - n_{\mathrm{C},\bar{t}}} \hat{V}_{\mathrm{C},i,t}(1, P^{(k)})\\ &\quad - \hat{\mathbb{Q}}_{\mathrm{C},i,t}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}),\end{aligned}$$

$$\begin{aligned}&\Delta\hat{\mathbb{Q}}_{\mathrm{D},j,t}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)\\ &\quad = \omega_{\mathrm{D}} \frac{(Q_{\mathrm{D},j} - y_j) P_{\mathrm{Ct},j} + y_j P_{\mathrm{Dd},j}}{\lambda_{\mathrm{D}}}\\ &\quad\quad + \mathbf{E}^{\hat{\Omega}_{\mathrm{y}}}[\mathbf{E}_{\mathbf{H}}[\mathbf{E}^{\hat{\Omega}_{\mathrm{x}}}[(1 - (\frac{\sum_{m=1}^{\eta N_F} r_{\mathrm{D},j0}^{(m)} x_{\mathrm{D},j0}^{(m)}(1 - y_j)}{\sigma_{\mathrm{D}}}\\ &\quad\quad + \frac{\sum_{m=\eta N_F+1}^{N_F} r_{\mathrm{D},jj}^{(m)} x_{\mathrm{D},jj}^{(m)} y_j}{\sigma_{\mathrm{D}}}))|\mathbf{H}, \mathbf{y}]]]\\ &\hat{V}_{\mathrm{D},j,t}(1, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j})\\ &\quad\quad - \sum_{k=1}^{K} \sum_{k'=1}^{K} \frac{\lambda_{\mathrm{D}} p_{\mathrm{Ct},k} p_{\mathrm{Dd},k'}}{n_{\mathrm{Dmax}} - n_{\mathrm{D},\bar{t}}} \hat{V}_{\mathrm{D},j,t}(1, P^{(k)}, P^{(k')})\\ &\quad\quad - \hat{\mathbb{Q}}_{\mathrm{D},j,t}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j),\end{aligned}$$

$$\begin{aligned}&\epsilon_{\tau_{\mathrm{C},i}(Q_{\mathrm{C},i}, P_{\mathrm{Ct},i}, t)}\\ &\quad = \sum_{t'=0}^{t} \mathbf{I}[(Q_{\mathrm{C},i,t'}, P_{\mathrm{Ct},i,t'}) = (Q_{\mathrm{C},i}, P_{\mathrm{Ct},i})],\end{aligned}$$

$$\begin{aligned}&\epsilon_{\tau_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j, t)}\\ &\quad = \sum_{t'=0}^{t} \mathbf{I}[(Q_{\mathrm{D},j,t'}, P_{\mathrm{Ct},j,t'}, P_{\mathrm{Dd},j,t'}, y_{j,t'})\\ &\quad = (Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j}, y_j)],\\ &\quad \bar{t} \triangleq \sup\{t : (Q_{\mathrm{C},i,t}, P_{\mathrm{Ct},i,t}) = (0,0)\}\end{aligned}$$

or

$$\bar{t} \triangleq \sup\{t : (Q_{\mathrm{D},j,t}, P_{\mathrm{Ct},j,t}, P_{\mathrm{Dd},j,t}) = (0,0,0)\}.$$

In the above equations, $\{\epsilon_t\}$ are the sequences of step sizes, which satisfy:

$$\epsilon_t \geq 0, \sum_{t=1}^{\infty} \epsilon_t = \infty, \sum_{t=1}^{\infty} \epsilon_t^2 < \infty. \tag{28}$$

- Step 4 (**Termination**):If $\|\hat{\mathbf{Q}}_{t+1} - \hat{\mathbf{Q}}_t\| < \delta$, stop; otherwise, set $t := t + 1$ and go to Step 2.

Remark 3 (Signaling Overhead): The signaling overhead involved in the proposed solution mainly consists of two parts. The first part is related to the per-queue Q-factors transfer between the BS and UEs. After a new cellular flow (resp. D2D flow) arrives and before a cellular flow (resp. D2D flow) leaves, the BS and CUE (resp. src. DUE) needs to exchange 1 *(resp.* 2*) real numbers. The second part is related to the determination of the subchannel allocation action, where every CUE and src. DUE in cellular mode needs to submit* ηN_{F} *bid, while every src. DUE in D2D mode needs to submit* $(1-\eta)N_{\mathrm{F}}$ *bid. Moreover, every src. DUE needs to send a one bit flag to the BS if its mode selection action changes. Compared to the Best-CSI subchannel allocation algorithm which allocates a subchannel to the UE with best channel state at any time slot, the additional signaling overhead per time slot is* $(n_{\mathrm{C},t} + n_{\mathrm{Dc},t})\eta N_{\mathrm{F}}$*, since the src. DUEs in D2D mode has to submit its CSI to the BS in the Best-CSI algorithm and the signaling overhead is the same with the proposed solution.*

TABLE 1. Simulation parameters.

Cell radius R	500m
D2D distance parameter ξ	$10 \times (\pi 200^2)^{-1}\mathrm{m}^{-2}$
Number of subchannels N_{F}	100
Path loss exponent α	3
Mean flow size σ_{C}, σ_{D}	20Kbits
Maximum flow number n_{Cmax}, n_{Dmax}	30
Number of zones K	5
Radius of zone k, d_k $(k = 1, \ldots, K-1)$	$(100 \times K)$m
Number of channel states V	7

VI. SIMULATION RESULTS

In this section, we evaluate the performance of our proposed mode selection and resource allocation algorithm with approximate MDP and online stochastic learning via simulation. We develop discrete event system-level simulator for D2D communications system with dynamic flow arrivals and our simulation parameters are summarized in Table 1. The carrier frequency and the time slot duration are set to 2GHz and 1ms, respectively. The system bandwidth is 10MHz and a subchannel corresponds to a resource block in 3G LTE system with 180KHz bandwidth. We select $L = 6$ MCSs from the 32 MCSs in 3G LTE specification [19] and the parameters are given in Table 2.

We compare the proposed algorithm against the following baseline algorithms in terms of the average power consumption per flow. These baseline algorithms are simple yet classical algorithms that can be used to clearly reveal the performance improvement achieved by considering the flow-level model.

- **Mode Selection**:
  - **RN**: Each DUE pair randomly determines its mode with 50% probability for each mode.
  - **DS**: Each DUE pair chooses the mode with the minimum distance between the transmitter and receiver [20].
- **Resource Allocation**:
  - **MaxR**: A subchannel is allocated to the link with the maximum achievable rate in a time slot [11].
  - **MaxRPR**: A subchannel is allocated to the link with the maximum ratio between its achievable rate and its transmit power as in [18] to achieve the best power-efficiency in a time slot.

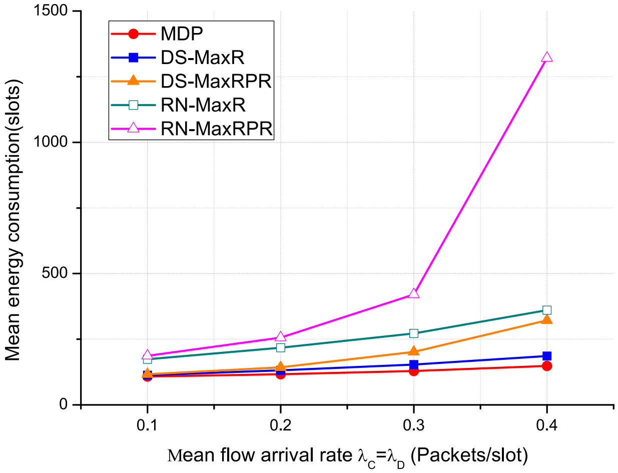

FIGURE 2. The mean energy consumption over all UEs versus the mean arrival rate of cellular flows λ_C, where λ_C is equal to the mean arrival rate of D2D flows λ_D.

Fig. 2 shows that the mean energy consumption over all UEs versus the mean arrival rate of cellular flows λ_C, where λ_C is equal to the mean arrival rate of D2D flows λ_D. It can be observed that the performance of the proposed MDP algorithm is better than that of DS-MaxR algorithm, DS-MaxRPR algorithm, RN-MaxR algorithm and RN-MaxRPR algorithm. The algorithms with DS have lower mean energy consumption than RN algorithms. When λ_C (resp. λ_D) grows, the mean energy consumption gradually increases, but the increasing rate of the proposed MDP algorithm is the slowest.

Fig. 3 shows that the average delay over all UEs versus the mean arrival rate of cellular flows λ_C, where λ_C is equal to the mean arrival rate of D2D flows λ_D. From Fig. 2, it can be observed that when λ_C (resp. λ_D) increases from 0.1 to 0.4 packets/slot, the proposed MDP algorithm has the lowest mean energy consumption, but it can not achieve the lowest average delay. In Fig. 3, it is obvious that DS algorithms have lower average delay than RN algorithms, and the average delay of the proposed MDP algorithm is close

TABLE 2. **AMC parameters for LTE.**

	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6
Modulation order	2	2	4	4	6	6
Rate R_v(bits/ms/180KHz)	56	120	208	280	408	552
Γ_v(dB)	-0.37	3.09	5.63	8.31	11.23	15.31

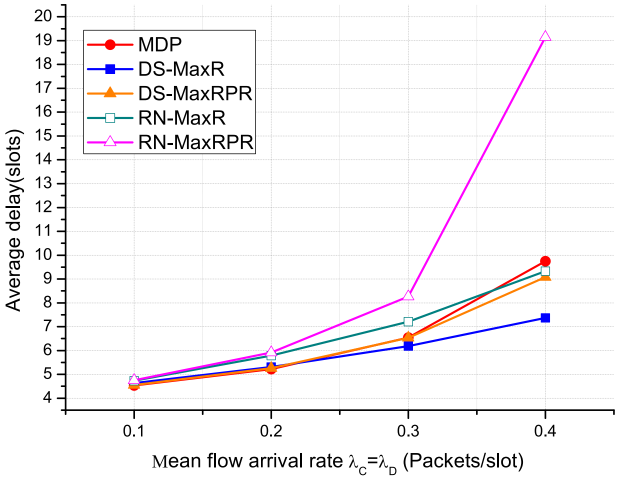

FIGURE 3. **The average delay over all UEs versus the mean arrival rate of cellular flows λ_C, where λ_C is equal to the mean arrival rate of D2D flows λ_D.**

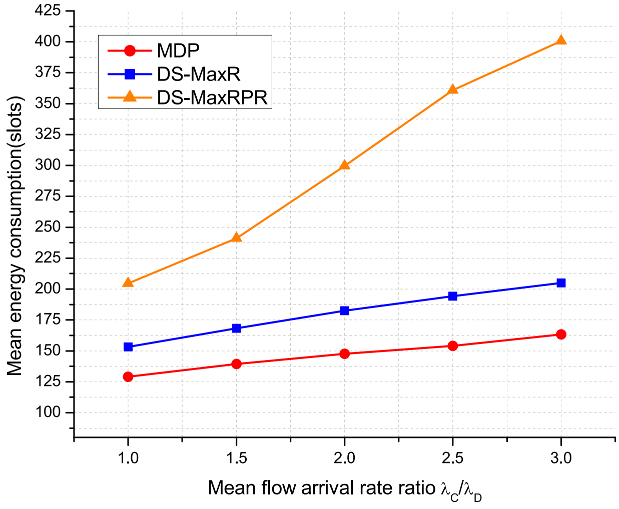

FIGURE 4. **The mean energy consumption over all UEs versus the ratio of the mean arrival rate of cellular flows λ_C to D2D flows λ_D, where $\lambda_C + \lambda_D = 0.6$.**

to those of DS algorithms when λ_C (resp. λ_D) is smaller than 0.3 packets/slot. When $0.3 < \lambda_C$ (resp. λ_D)$\leq$ 0.4, the proposed MDP algorithm has larger average delay than the two DS algorithms. Therefore, the proposed MDP algorithm achieves better performance for delay when λ_C (resp. λ_D) is small.

Fig. 4 shows the mean energy consumption over all UEs versus the ratio of the mean arrival rate of cellular flows λ_C to D2D flows λ_D, where $\lambda_C + \lambda_D = 0.6$. The ratio denotes the degree of traffic load imbalance between cellular flows and D2D flows. Because the performance of the proposed MDP algorithm is far better than the RN algorithms, we just show the comparison results of MDP algorithm with DS algorithms. It can be observed that with the increase of ratio between λ_C and λ_D, the mean energy consumption also increases, but the increasing rate of the mean energy consumption in the proposed MDP algorithm is the slowest among the three algorithms. The proposed MDP algorithm achieves a better performance when the gap between λ_C and λ_D is greater, which means MDP algorithm can effectively cope with the traffic load imbalance between cellular flows and D2D flows.

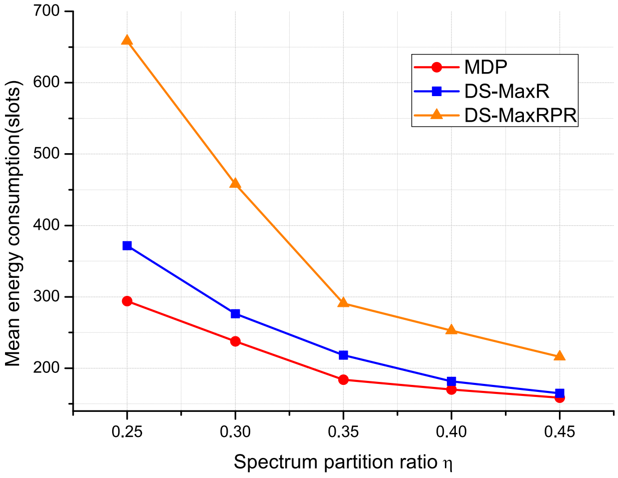

FIGURE 5. **The mean energy consumption over all UEs versus the spectrum partition ratio η.**

Fig. 5 shows that the mean energy consumption over all UEs versus the spectrum partition ratio η. For example, when $\eta = 0.25$, the number of subchannels allocated to cellular flows is 25, while the number of subchannels for D2D flows is 75. We increase η from 0.25 to 0.45, where smaller η leads to greater gap between the number of subchannels allocated to cellular flows and D2D flows. It is obvious that as η decreases, but increasing rate of the mean energy consumption for the DS-MaxR algorithm and DS-MaxRPR algorithm are greater than the proposed MDP algorithm. This is because the mode selection Q-factors take η into consideration in proposed MDP algorithm.

VII. CONCLUSION

In this paper, we have proposed mode selection and resource allocation algorithm for D2D communications with dynamic flow arrivals, which minimizes the average energy consumption of flow transmission. The optimal resource control problem is cast into an infinite horizon average reward MDP. We addressed the issue of exponential memory requirement and computational complexity by using Q-factor approximation techniques to reduce the state space. Moreover, online stochastic learning algorithm was adopted to update the Q-factors based on the real-time observations of CSI. The obtained solution has a simple structure with limited signaling overhead. Simulation results show that the proposed approach outperforms various existing baseline algorithms, which can be applied in the future 5G wireless system to support the massive number of connected devices.

APPENDIX

A. DERIVATION OF $\bar{U}$

First, according to the definition of $\bar{U}$, we have $\bar{U} = \omega_{\mathrm{C}}\bar{U}_{\mathrm{C}} + \omega_{\mathrm{D}}\bar{U}_{\mathrm{D}}$, where $\bar{U}_{\mathrm{C}}$ and $\bar{U}_{\mathrm{D}}$ are the mean energy consumption for the transmission of a cellular flow and a D2D flow, respectively. According to Little's law, the mean delay of a cellular flow equals $\frac{n_{\mathrm{C},t}}{\lambda_{\mathrm{C}}}$, and we have $\bar{U}_{\mathrm{C}} = \sum_{i=1}^{n_{\mathrm{Cmax}}} \frac{Q_{\mathrm{C},i}P_{\mathrm{Ct},i}}{\lambda_{\mathrm{C}}}$ by the definition of mean energy consumption of a flow. On the other hand, $\bar{U}_{\mathrm{D}} = p_{\mathrm{Dd}}\bar{U}_{\mathrm{Dd}} + p_{\mathrm{Dc}}\bar{U}_{\mathrm{Dc}}$, where p_{Dd} (resp. p_{Dc}) is the probability that a D2D flow selects D2D mode (resp. cellular mode), and $\bar{U}_{\mathrm{Dd}}$ (resp. $\bar{U}_{\mathrm{Dc}}$) is the mean energy consumption of a D2D flow in D2D mode (resp. cellular mode). According to Little's law, the mean delay of a D2D flow in D2D mode (resp. cellular flow) equals $\frac{n_{\mathrm{Dd},t}}{\lambda_{\mathrm{D}}p_{\mathrm{Dd}}}$ (resp. $\frac{n_{\mathrm{Dc},t}}{\lambda_{\mathrm{D}}p_{\mathrm{Dc}}}$), and we have $\bar{U}_{\mathrm{D}} = \sum_{j=1}^{n_{\mathrm{Dmax}}} \frac{(Q_{\mathrm{D},j}-y_j)P_{\mathrm{Ct},j}+y_jP_{\mathrm{Dd},j}}{\lambda_{\mathrm{D}}}$ by the definition of mean energy consumption. This completes the proof.

B. PROOF OF LEMMA 1

$$
\begin{aligned}
\theta + V(\mathbf{Q},\mathbf{H}) = & \min_{\Omega(\mathbf{Q},\mathbf{H})} \Big\{ g(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})) \\
& + \sum_{\mathbf{Q}',\mathbf{H}'} \Pr.[\mathbf{Q}',\mathbf{H}'|\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})] V(\mathbf{Q}',\mathbf{H}') \Big\} \\
\overset{(a)}{=} & \min_{\Omega(\mathbf{Q},\mathbf{H})} \Big\{ g(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})) \\
& + \sum_{\mathbf{Q}'} \Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})] \Big(\sum_{\mathbf{H}'} \Pr.(\mathbf{H}'|\mathbf{Q}') V(\mathbf{Q}',\mathbf{H}') \Big) \Big\} \\
\overset{(b)}{=} & \min_{\Omega(\mathbf{Q},\mathbf{H})} \Big\{ g(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})) \\
& + \sum_{\mathbf{Q}'} \Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})] V(\mathbf{Q}') \Big\}, \quad \forall \mathbf{Q} \in \mathcal{Q},\ \mathbf{H} \in \mathcal{H},
\end{aligned}
$$

where (a) is due to (2) by the i.i.d. assumption of CSI over time slots, (b) is due to the definition $V(\mathbf{Q})$ given in Section IV.A.

Taking the conditional expectation (conditioned on $\mathbf{Q}$) on both sides of the equation above, we have $\forall \mathbf{Q} \in \mathcal{Q}$

$$
\begin{aligned}
\theta + V(\mathbf{Q}) = & \mathbf{E}_{\mathbf{H}} \Big[\min_{\Omega(\mathbf{Q},\mathbf{H})} \Big\{ g(\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})) \\
& + \sum_{\mathbf{Q}'} \Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{H},\Omega(\mathbf{Q},\mathbf{H})] V(\mathbf{Q}') \Big\} \Big] \\
\overset{(c)}{=} & \min_{\Omega(\mathbf{Q})} \Big\{ g(\mathbf{Q},\Omega_{\mathrm{y}}(\mathbf{Q})) + \sum_{\mathbf{Q}'} \Pr.[\mathbf{Q}'|\mathbf{Q},\Omega(\mathbf{Q})] V(\mathbf{Q}') \Big\},
\end{aligned}
$$

where (c) is due to the definition of "conditional reward" $g(\mathbf{Q},\Omega_{\mathrm{y}}(\mathbf{Q}))$ and "conditional transition probability" $\Pr.[\mathbf{Q}'|\mathbf{Q},\Omega(\mathbf{Q})]$ given in Section IV.A.

C. PROOF OF LEMMA 2

Under the randomized base policy $\hat{\Omega}$, we have the following Bellman's equation:

$$
\theta + \mathbb{Q}(\mathbf{Q},\mathbf{y}) = g(\mathbf{Q},\mathbf{y}) + \sum_{\mathbf{Q}' \in \mathcal{Q}} \hat{\Pr}.[\mathbf{Q}'|\mathbf{Q},\mathbf{y}] \underbrace{\mathbf{E}^{\hat{\Omega}_{\mathrm{y}}}[\mathbb{Q}(\mathbf{Q}',\mathbf{y}')]}_{\hat{V}(\mathbf{Q}')}. \tag{29}
$$

where $\hat{\Pr}.[\mathbf{Q}'|\mathbf{Q},\mathbf{y}] = \mathbf{E}_{\mathbf{H}}[\mathbf{E}^{\hat{\Omega}_{\mathrm{x}}}[\Pr.[\mathbf{Q}'|\mathbf{Q},\mathbf{H},\mathbf{y},\mathbf{x}]|\mathbf{H},\mathbf{y}]]$.

First, assume the additive property w.r.t. the Q-factor and value function hold under the randomized base policy $\hat{\Omega}$. Next, we have

$$
\begin{aligned}
g(\mathbf{Q},\mathbf{y}) = & \sum_{i=1}^{n_{\mathrm{Cmax}}} g_{\mathrm{C},i}(Q_{\mathrm{C},i},P_{\mathrm{Ct},i}) \\
& + \sum_{j=1}^{n_{\mathrm{Dmax}}} g_{\mathrm{D},j}(Q_{\mathrm{D},j},P_{\mathrm{Ct},j},P_{\mathrm{Dd},j},y_j)
\end{aligned} \tag{30}
$$

where $g_{\mathrm{C},i}(Q_{\mathrm{C},i},P_{\mathrm{Ct},i})$ and $g_{\mathrm{D},j}(Q_{\mathrm{D},j},P_{\mathrm{Ct},j},P_{\mathrm{Dd},j},y_j)$ are given in Lemma 2. Thus, from (29) and (30) we have

$$
\begin{aligned}
& \sum_{i=1}^{n_{\mathrm{Cmax}}} \theta_{\mathrm{C},i} + \sum_{j=1}^{n_{\mathrm{Dmax}}} \theta_{\mathrm{D},j} + \sum_{i=1}^{n_{\mathrm{Cmax}}} \hat{\mathbb{Q}}_{\mathrm{C},i}(Q_{\mathrm{C},i},P_{\mathrm{Ct},i}) \\
& \quad + \sum_{j=1}^{n_{\mathrm{Dmax}}} \hat{\mathbb{Q}}_{\mathrm{D},j}(Q_{\mathrm{D},j},P_{\mathrm{Ct},j},P_{\mathrm{Dd},j}) \\
& = \sum_{i=1}^{n_{\mathrm{Cmax}}} g_{\mathrm{C},i}(Q_{\mathrm{C},i},P_{\mathrm{Ct},i}) + \sum_{j=1}^{n_{\mathrm{Dmax}}} g_{\mathrm{D},j}(Q_{\mathrm{D},j},P_{\mathrm{Ct},j},P_{\mathrm{Dd},j},y_j) \\
& \quad + \sum_{i=1}^{n_{\mathrm{Cmax}}} \sum_{Q'_{\mathrm{C},i},P'_{\mathrm{Ct},i}} \hat{\Pr}.[Q'_{\mathrm{C},i},P'_{\mathrm{Ct},i}|Q_{\mathrm{C},i},P_{\mathrm{Ct},i}] \\
& \quad \hat{V}_{\mathrm{C},i}(Q'_{\mathrm{C},i},P'_{\mathrm{Ct},i}) + \sum_{j=1}^{n_{\mathrm{Dmax}}} \sum_{Q'_{\mathrm{D},j},P'_{\mathrm{Ct},j},P'_{\mathrm{Dd},j}} \\
& \quad \hat{\Pr}.[Q'_{\mathrm{D},j},P'_{\mathrm{Ct},j},P'_{\mathrm{Dd},j}|Q_{\mathrm{D},j},P_{\mathrm{Ct},j},P_{\mathrm{Dd},j},y_j] \\
& \quad \hat{V}_{\mathrm{D},j}(Q'_{\mathrm{D},j},P'_{\mathrm{Ct},j},P'_{\mathrm{Dd},j})
\end{aligned}
$$

The structure in (31) is decoupled under the additive assumption, $\theta_{\mathrm{C},i}$ and $\theta_{\mathrm{D},j}$ are unique, as well as $\{\hat{V}_{\mathrm{C},i}(Q_{\mathrm{C},i},P_{\mathrm{Ct},i})\}$

and $\{\hat{V}_{\mathrm{D},j}(Q_{\mathrm{D},j}, P_{\mathrm{Ct},j}, P_{\mathrm{Dd},j})\}$ are unique up to an additive constant [21]. Therefore, the per-queue Bellman's equation (17) and (18) hold.

REFERENCES

[1] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Commun.*, vol. 19, no. 3, pp. 96–104, Jun. 2012.

[2] X. Lin, J. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.

[3] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: Challenges, solutions, and future directions," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 86–92, May 2014.

[4] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.

[5] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *IEEE/ACM Trans. Netw.*, vol. 13, no. 3, pp. 636–647, Jun. 2005.

[6] R. Prakash and V. V. Veeravalli, "Centralized wireless data networks with user arrivals and departures," *IEEE Trans. Inf. Theory*, vol. 53, no. 2, pp. 695–713, Feb. 2007.

[7] U. Ayesta, M. Erausquin, and P. Jacko, "A modeling framework for optimizing the flow-level scheduling with time-varying channels," *Perform. Eval.*, vol. 67, no. 11, pp. 1014–1029, 2010.

[8] L. Lei, Y. Kuang, N. Cheng, X. Shen, D. Zhong, and C. Lin, "Delay-optimal dynamic mode selection and resource allocation in device-to-device communications—Part I: Optimal policy," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3474–3490, May 2016.

[9] L. Lei, Y. Kuang, N. Cheng, X. Shen, D. Zhong, and C. Lin, "Delay-optimal dynamic mode selection and resource allocation in device-to-device communications—Part II: Practical algorithm," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3491–3505, May 2016.

[10] P. Mach, Z. Becvar, and T. Vanek, "In-band device-to-device communication in OFDMA cellular networks: A survey and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1885–1922, 4th Quart., 2015.

[11] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[12] D. Feng *et al.*, "Mode switching for energy-efficient device-to-device communications in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6993–7003, Dec. 2015.

[13] G. Yu, L. Xu, D. Feng, R. Yin, G. Y. Li, and Y. Jiang, "Joint mode selection and resource allocation for device-to-device communications," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3814–3824, Nov. 2014.

[14] A. Ghazanfari, A. Tölli, and J. Kaleva, "Joint power loading and mode selection for network-assisted device-to-device communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 2548–2553.

[15] H. ElSawy, E. Hossain, and M. S. Alouini, "Analytical modeling of mode selection and power control for underlay D2D communication in cellular networks," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4147–4161, Nov. 2014.

[16] K. Zhu and E. Hossain, "Joint mode selection and spectrum partitioning for device-to-device communication: A dynamic Stackelberg game," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1406–1420, Mar. 2015.

[17] L. Lei, X. Shen, M. Dohler, C. Lin, and Z. Zhong, "Queuing models with applications to mode selection in device-to-device communications underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6697–6715, Dec. 2014.

[18] C. Gao, X. Sheng, J. Tang, W. Zhang, S. Zou, and M. Guizani, "Joint mode selection, channel allocation and power assignment for green device-to-device communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 178–183.

[19] "Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation," Technical Specification Group Radio Access Network, 3GPP, Tech. Rep. 36.211, Jun. 2008.

[20] S. Hakola, T. Chen, J. Lehtomaki, and T. Koskela, "Device-to-device (D2D) communication in cellular network—Performance analysis of optimum and practical communication mode selection," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2010, pp. 1–6.

[21] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Belmont, MA, USA: Athena Scientific, 2007.

LEI LEI (M'13) received the B.S. and Ph.D. degrees in telecommunications engineering from the Beijing University of Posts and Telecommunications, China, in 2001 and 2006, respectively. From 2006 to 2008, she was a Post-Doctoral Fellow with the Computer Science Department, Tsinghua University, Beijing, China. She was with the Wireless Communications Department, China Mobile Research Institute, from 2008 to 2011. She has been a Professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, since 2011. Her current research interests include performance evaluation, quality-of-service, and radio resource management in wireless communication networks.

QINGYUN HAO received the B.S. degree from Beijing Jiaotong University, China, in 2014. She is currently pursuing the master's degree with Beijing Jiaotong University. Her research interests lie in wireless communications, especially radio resource optimization and device-to-device communications.

ZHANGDUI ZHONG is currently a Professor and an Advisor of the Ph.D. candidates with Beijing Jiaotong University. He is the Dean of the School of Computer and Information Technology and a Chief Scientist of the State Key Laboratory of Rail Traffic Control and Safety with Beijing Jiaotong University. He is also the Director of the Innovative Research Team of Ministry of Education, and a Chief Scientist of the Ministry of Railways, China. He is an Executive Council Member of the Radio Association of China, and a Deputy Director of the Radio Association of Beijing. He has authored or co-authored seven books, five invention patents, and over 200 scientific research papers in his research area. His interests are wireless communications for railways, control theory and techniques for railways, and GSM-R system. His research has been widely used in the railway engineering, such as Qinghai-Xizang railway, Datong-Qinhuangdao Heavy Haul railway, and many high-speed railway lines of China. He received the MaoYiSheng Scientific Award of China, the Zhan-TianYou Railway Honorary Award of China, and the Top 10 Science/Technology Achievements Award of Chinese Universities.