

A Two-Timescale Approach to Mobility Management for Multi-Cell Mobile Edge Computing

Zezu Liang, Yuan Liu, Tat-Ming Lok, and Kaibin Huang

Abstract

Mobile edge computing (MEC) is a promising technology for enhancing the computation capacities and features of mobile users by offloading complex computation tasks to the edge servers. However, mobility poses great challenges on delivering reliable MEC service required for latency-critical applications. First, mobility management has to tackle the dynamics of both user's location changes and task arrivals that vary in different timescales. Second, user mobility could induce service migration, leading to reliability loss due to the migration delay. In this paper, we propose a two-timescale mobility management framework by joint control of service migration and transmission power to address the above challenges. Specifically, the service migration operates at a large timescale to support user mobility in the multi-cell network, while the power control is performed at a small timescale for real-time task offloading. Their joint control is formulated as an optimization problem aiming at the long-term mobile energy minimization subject to the reliability requirement of computation offloading. To solve the problem, we propose a Lyapunov-based framework to decompose the problem into different timescales, based on which a low-complexity two-timescale online algorithm is developed by exploiting the problem structure. The proposed online algorithm is shown to be asymptotically optimal via theoretical analysis, and is further developed to accommodate the multiuser management. The simulation results demonstrate that our proposed algorithm can significantly improve the energy and reliability performance.

Index Terms

Mobile-edge computing (MEC), mobility management, service migration, Lyapunov optimization.

Z. Liang and T. M. Lok are with Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: zezuliang@gmail.com; tmlok@ie.cuhk.edu.hk). Y. Liu is with school of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: eeyliu@scut.edu.cn). K. Huang is with Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: huangkb@eee.hku.hk).

I. INTRODUCTION

The rapid development of advanced mobile applications and Internet-of-Things (IoT) calls for high quality of service (QoS), such as ultra-low latency, ultra-high reliability, robust security, enhanced broadband access, and ubiquitous connectivity. It is commonly agreed that these strict requirements cannot be fulfilled by the conventional cloud computing as central cloud is far from real-time data generated by edge users. Mobile (or multi-access) edge computing (MEC) has been proposed as a solution to address the issue by deploying cloud computing functions at the network edges [1]–[5]. Specifically, MEC allows users to offload computation tasks to proximate edge servers [e.g., base stations (BSs) or access points] for execution. This avoids data transportation across backhaul networks and thereby reduces latency and traffic congestion. Given the dense geographical distribution of servers, MEC is envisioned as a promising platform for enabling the emerging computation-intensive and latency-critical applications, such as real-time online gaming and autonomous driving [3]. In this paper, we investigate the mobility management problem in MEC, aiming at supporting the MEC applications under the presence of user mobility.

A. *Related Works*

As mobile users may traverse different cells, one challenge faced by designing multi-cell MEC networks is mobility management to guarantee uninterrupted service [6], [7]. The direct way to support mobility is service migration [8]–[10], namely, continuously migrate the ongoing computing services of mobile users to their dynamically associated servers/BSs along the users' traveling paths. However, the uncertainty of user mobility makes the optimal migration policies difficult to design. Three main approaches have been developed to address this issue. The first one is based on the prediction of the short-term user mobility and service latency to make more informed migration decisions [11], [12]. The second approach involves online migration decision making based on modeling the user movement as a Markovian process and applying the theory of Markov decision process (MDP) to optimize the decisions [13]–[16]. The limitation of such an approach lies in its requirement of statistical information of user mobility, which is not always available in practice.

The last approach, which is closely related to this work, focuses on online migration design without a priori knowledge of future user mobility. Specifically, learning-driven migration schemes are proposed in [17], [18] based on multi-armed bandit theory, and in [19] using the deep reinforcement learning approach, in which the user copes with the lack of prior knowledge

using the trial-and-error method. On the other hand, by utilizing the Lyapunov optimization technique, an online migration strategy is proposed in [20] that balances the service latency, the incurred migration cost, and the long-term user movement. The theory is also applied in [21] to develop a framework of dynamic user-BS association to satisfy the application requirements of latency and reliability under constraints on the task queue lengths.

In view of prior work, two issues have not been addressed. First, only service migration is insufficient for guaranteeing the QoS for many latency-critical applications. In general, latency-critical tasks generated by the application often arrive at a smaller timescale than the service migration that adapts mainly to user movement. For instance, the update time in industrial IoT applications is between $0.5 \sim 500$ ms [22], while the service migration is performed less frequently at the timescale of seconds to minutes in practice [23]. Second, the migration process induces reliability loss due to the migration delay. The BS handover procedure and migration of user's application profiles require a certain amount of time to complete (15 ms delay in a 5G handover scenario [24] for example). This leads to service interruption when tasks arrive during the migration process. To address these two issues motivates this work.

B. Our Contributions

In this paper, we consider a user moving in the multi-cell MEC network and aim to guarantee the reliability of the user's latency-critical application. We propose a novel mobility management framework that features the joint optimization of service migration and power control. Motivated by the fact that the user's location changes slower than the task arrivals, our proposed framework performs service migration at a large timescale to support user mobility and dynamic transmit power control at a small timescale to accommodate the real-time task offloading. In order to quantify the reliability loss caused by the mobile environment as well as the migration delay, we define the event of task failure as one that the offloading time exceeds the latency requirement or a task arrives during the migration process. The mentioned joint optimization aims at minimizing the long-term user's energy consumption while ensuring the reliability requirement that the probability of task failure is below a pre-defined threshold.

The main contributions of this paper are summarized as follows.

- We propose an online two-timescale control algorithm to solve the formulated problem. By invoking the Lyapunov optimization framework, our proposed algorithm can decouple the original two-timescale joint problem into two subproblems with different timescales,

i.e., the service migration subproblems over the large timescale, and the power allocation subproblems over the small timescale. As the core components of the algorithm, we further derive the optimal migration policy and the optimal power strategy for solving these two subproblems, which allows to make online decisions in low complexity and without requiring any future information.

- We prove that the proposed algorithm can achieve asymptotically optimal performance. Furthermore, the optimal power control is proved to have a threshold-based structure. Specifically, in each slot, a task is offloaded with the minimum required power if the power is below the threshold, and the task is dropped otherwise. In addition, it is shown that in each frame the user always migrates its service to the BS with the minimum weighted sum of energy consumption and task-dropping cost.
- We also extend the online algorithm to multiuser management by designing an efficient per-frame migration scheme. The proposed scheme takes into account the load-balance factor in multiuser migration decisions, and it is based on the adjustment of the worst user-BS association to find a near-optimal solution.

The rest of this paper is organized as follows. We introduce the system model and formulate the problem in Section II and Section III, respectively. We design the online algorithm framework in Section IV, and provide the algorithm implementation and performance analysis in Section V. The extension to multiuser management is discussed in Section VI. Simulation results are presented in Section VII, and in Section VIII, we conclude the paper.

II. SYSTEM MODEL

As shown in Fig. 1, we consider that a mobile user moves in a multi-cell network deployed with N based stations (BSs), denoted by set $\mathcal{N} = \{1, 2, \dots, N\}$. The network operates in a time-slotted manner, in which each time slot $t \in \{0, 1, \dots\}$ has slot length τ that is consistent with the coherence time of the wireless channel. Each BS is integrated with an MEC server and can provide computing service. The application of the mobile user is computation-intensive such that all the generated tasks have to be offloaded to the BS (server) for execution. We assume that the computation tasks are homogeneous [25]–[27] and described by $A(L, \xi, \tau_d)$, in which L (in bits) denotes the input data size of the task, ξ denotes the number of CPU cycles required for processing the L -bit input data, and τ_d denotes the task latency requirement. We consider a latency-critical scenario where the task latency requirement does not exceed the slot length,

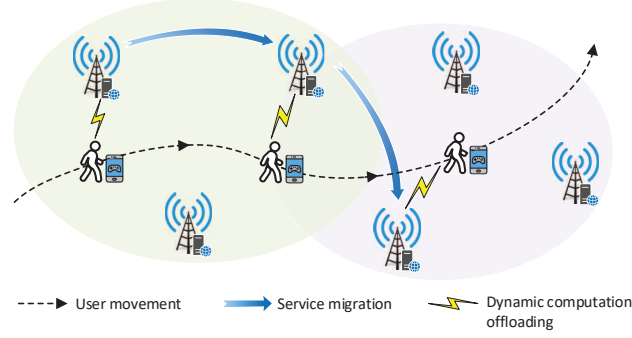


Fig. 1: System Model

i.e., $\tau_d \leq \tau$. The task arrivals across slots are modeled as a Bernoulli process [25], [28], [29]. Specifically, at the beginning of each time slot t , a computation task $A(L, \xi, \tau_d)$ arrives with probability ρ , and with probability $1 - \rho$, there is no task arrival. Therefore, let $a(t) \in \{0, 1\}$ be the task arrival indicator. We have $\Pr(a(t) = 1) = \rho$ and $\Pr(a(t) = 0) = 1 - \rho$.

In order to satisfy the application's latency requirement, joint service migration and transmit power control are considered during the user movement. We assume that service migration occurs when the user changes its BS association from one to another and is conducted by joint communication handover and computation migration between the two BSs [6], [10], [12], [17], [20], [30]. Here, computation migration refers to the migration of the user's application instances (or application state) [6]. Meanwhile, the user performs dynamic transmit-power control for computation offloading as its serving BS and channel change. The corresponding models and assumptions are elaborated as follows.

1) *Two-Timescale Operation Model*: Note that for the case of latency-critical applications, task offloading is often performed more frequently than service migration, due to the different timescales between the task arrivals and the user's location changes. For instance, the tasks generated from road safety are at the timescale of hundred milliseconds [22], while the service migration occurs over the timescale of seconds to minutes [23] since it reacts mainly to the user movement and requires high operational cost. In this regard, we propose a two-timescale mobility management framework for large-timescale service migration and small-timescale power control as shown in Fig. 2. Specifically, we group every consecutive T time slots as a time frame, indexed by $k \in \{0, 1, \dots\}$, and denote the set of time slots in the k -th frame as $\mathcal{T}_k = \{kT, kT + 1, \dots, (k + 1)T - 1\}$. We assume that:

- Large timescale: Service migration is made at the beginning of each frame and remains

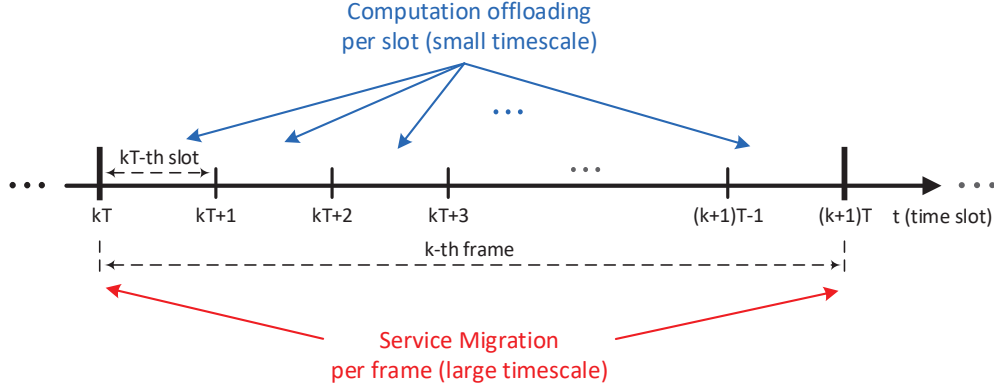


Fig. 2: The two-timescale model of service migration and computation offloading

unchanged during a frame.

- Small timescale: Transmit power control is performed at each time slot for task offloading.

2) *Service Migration Model*: At time slot $t = kT$, i.e., the beginning of a frame, the user determines the migration/association decision for the k -th frame. Let $n(k) \in \mathcal{N}$ denote the user's associated BS. Clearly, service migration is triggered when $n(k) \neq n(k-1)$. We assume that the migration operation can cause C slots of service interruption (i.e., service migration delay) at the beginning of a frame, with $0 \leq C < T$, during which computation offloading is temporarily disrupted. We further denote the set of time slots in frame k for doing migration as \mathcal{T}_k^c . By definition, we have $\mathcal{T}_k^c = \{kT, \dots, kT + C - 1\}$ if $n(k) \neq n(k-1)$, and $\mathcal{T}_k^c = \{\emptyset\}$ otherwise.

3) *Computation Offloading Model*: At each time slot t , if a task arrives, the user adjusts the transmit power $p(t)$ based on the real-time channel condition to support task offloading. We denote the uplink channel power gain from the user to BS n at slot t as $h_n(t)$, which includes path loss (that captures the user's location change) and small-scale fading. We assume that $h_n(t)$ experiences block fading, i.e., $h_n(t)$ remains constant within each time slot but possibly varies over different time slots. Let $f_n(k)$ denote the computation rate (CPU cycles per second) of BS n assigned to the user. We assume that the BS adjusts $f_n(k)$ on a frame basis, since dynamic provisioning of the compute resource (e.g., virtual machines or containers) is often carried out at a larger time interval. Then, given the associated BS $n(k) = n$ and transmit power $p(t)$, the total latency for offloading and computing task $A(L, \xi, \tau_d)$ at slot $t \in \mathcal{T}_k$ can be expressed as

$$D(n, p(t)) = \frac{L}{W \log_2 \left(1 + \frac{p(t)h_n(t)}{\sigma} \right)} + \frac{\xi}{f_n(k)}, \quad (1)$$

where W is the channel bandwidth and is assumed to be homogenous among BSs for simplicity, and σ denotes the noise power. We consider the task offloading and ignore the result downloading

phase because of the relative much smaller sizes of computed results.

Accordingly, the user's energy consumption for offloading a task at slot $t \in \mathcal{T}_k$ is given by

$$E(n, p(t)) = \frac{Lp(t)}{W \log_2 \left(1 + \frac{p(t)h_n(t)}{\sigma}\right)}. \quad (2)$$

III. PROBLEM FORMULATION

Based on the proposed two-timescale mobility management framework, our goal is to design an online service migration and power control algorithm that minimizes the user's energy consumption and meanwhile satisfies the tasks' latency requirements continuously. Nevertheless, due to the service migration delay and the wireless channel fluctuation, some of the arrived computation tasks may not be accomplished within the deadline, leading to task failure. For example, task failure may occur when the user's service is being migrated, or when the wireless channel from user to its associated BS is in a deep fade. To take this aspect into consideration, we denote $X(t) \in \{0, 1\}$ as the task failure indicator, with $X(t) = 1$ indicating the task failure occurs at slot t , and $X(t) = 0$ otherwise. Then, given the BS association decision $n(k) = n$ and transmit power $p(t)$, the task failure event at slot $t \in \mathcal{T}_k$ can be characterized by

$$X(t) = \begin{cases} \mathbb{1}_{\{a(t)=1\}}, & \text{if } t \in \mathcal{T}_k^c, \\ \mathbb{1}_{\{a(t)=1, D(n,p(t))>\tau_d\}}, & \text{if } t \in \mathcal{T}_k \setminus \mathcal{T}_k^c, \end{cases} \quad (3)$$

where $\mathbb{1}_{\{x\}}$ is the indicator function, with $\mathbb{1}_{\{x\}} = 1$ if event x is true and $\mathbb{1}_{\{x\}} = 0$ otherwise. (3) specifies that task failure occurs if there is a task arrival during the service migration process (i.e., $t \in \mathcal{T}_k^c$), or the arrived task can not be completed within the latency requirement.

The task failure events can degrade the service reliability for latency-critical applications. In this regard, we impose the following constraint on the average occurrence rate of task failure:

$$\lim_{K \rightarrow \infty} \frac{1}{KT} \sum_{k=0}^{K-1} \sum_{t \in \mathcal{T}_k} \mathbb{E} \{X(t)\} \leq \epsilon, \quad (4)$$

where $\epsilon \ll 1$ is the maximum threshold of the task-failure rate, which can be seen as the application's reliability requirement. The expectation $\mathbb{E}\{\cdot\}$ is taken over all sources of randomness, including task arrivals and dynamics of channel conditions and BSs' computation rates.

Similarly, combining the factors of task arrival and service interruption during migration, we can express the user's energy consumption at every slot $t \in \mathcal{T}_k$ as

$$\mathcal{E}(t) = \begin{cases} 0, & \text{if } t \in \mathcal{T}_k^c, \\ E(n, p(t)) \cdot \mathbb{1}_{\{a(t)=1, D(n,p(t)) \leq \tau_d\}}, & \text{if } t \in \mathcal{T}_k \setminus \mathcal{T}_k^c, \end{cases} \quad (5)$$

i.e., the user consumes energy only in the case when the arrived task can be accomplished within the latency requirement.

Incorporating the constraint (4), our studied problem is to minimize the user's long-term energy consumption while ensuring the reliability requirement for the latency-critical application, which can be formulated as

$$(P1) \quad \min_{\{n(k)\}, \{p(t)\}} \quad E_{\text{av}} \triangleq \lim_{K \rightarrow \infty} \frac{1}{KT} \sum_{k=0}^{K-1} \sum_{t \in \mathcal{T}_k} \mathbb{E} \{ \mathcal{E}(t) \} \quad (6a)$$

$$\text{s.t.} \quad X_{\text{av}} \triangleq \lim_{K \rightarrow \infty} \frac{1}{KT} \sum_{k=0}^{K-1} \sum_{t \in \mathcal{T}_k} \mathbb{E} \{ X(t) \} \leq \epsilon \quad (6b)$$

$$n(k) \in \mathcal{N}, \quad k = 0, 1, \dots \quad (6c)$$

$$0 \leq p(t) \leq \bar{P}, \quad t = 0, 1, \dots \quad (6d)$$

where (6d) is the peak power constraint of the user.

There are two major challenges in solving Problem (P1). First, optimally solving Problem (P1) requires the complete information of the user trajectory, task arrivals, and network-level conditions over the entire time horizon, which is extremely difficult to acquire in advance. Second, the migration decision $n(k)$ and the power allocation $p(t)$ that change in different timescales, are tightly coupled, e.g., the migration decision $n(k)$ for the k -th frame affects the power allocations $\{p(t)\}$ in slots $t \in \mathcal{T}_k$, and vice versa. To address the above challenges, we develop an online two-timescale control algorithm in the following two sections.

IV. ONLINE TWO-TIMESCALE ALGORITHM DESIGN

In this section, we present the framework design of our online algorithm. First, we transform Problem (P1) into an online optimization problem using the Lyapunov technique. Subsequently, a two-timescale control algorithm is designed to solve the transformed problem optimally.

A. Problem Transformation

In order to take the advantage of Lyapunov optimization, we first convert the reliability constraint (6b) into an equivalent queue stability constraint, which is described as follows. We construct a virtual queue with the queue length evolving according to $X(t)$ and ϵ as

$$Q(t+1) = [Q(t) + X(t) - \epsilon]^+, \quad t = 0, 1, \dots \quad (7)$$

where $[\cdot]^+ \triangleq \max\{\cdot, 0\}$. $Q(t)$ is the queue length at slot t , with $Q(0) = 0$, which indicates how far the current task-failure backlog exceeds the threshold ϵ . According to the Lyapunov optimization theory [31], the long-term time-averaged constraint (6b) is equivalent to the mean-rate stability constraint on the virtual queue, i.e., $\lim_{t \rightarrow \infty} \mathbb{E}\{Q(t)\}/t \rightarrow 0$.

To proceed, we define a T -slot (i.e., frame-based) conditional Lyapunov drift as

$$\Delta_T(Q(t)) \triangleq \mathbb{E}\left\{\frac{1}{2}Q(t+T)^2 - \frac{1}{2}Q(t)^2 \middle| Q(t)\right\}. \quad (8)$$

Given the current queue length $Q(t)$, $\Delta_T(Q(t))$ characterizes the expected change in quadratic function of the queue length after T time slots. Intuitively, minimizing $\Delta_T(Q(t))$ in each T slots can prevent the queue length from unbounded growth and thus stabilize the queue.

Recalling that our problem objective is to minimize the energy consumption defined in (6a), we add the energy consumption (as a penalty function) into (8) to obtain the following drift-plus-penalty term for the k -th frame:

$$\mathcal{D}(Q(kT)) \triangleq \Delta_T(Q(kT)) + V\mathbb{E}\left\{\sum_{t \in \mathcal{T}_k} \mathcal{E}(t) \middle| Q(kT)\right\}. \quad (9)$$

where $V \geq 0$ is a control parameter, indicating an importance weight on how much we emphasize the energy consumption minimization.

The main idea of the Lyapunov optimization-based algorithm is to minimize the upper bound of the drift-plus-penalty term $\mathcal{D}(Q(kT))$ for joint queue stability and energy consumption minimization. To this end, we have the following two lemmas regarding the upper bound of $\mathcal{D}(Q(kT))$ for our two-timescale algorithm design.

Lemma 1: Under any feasible decisions $n(k) \in \mathcal{N}$ and $0 \leq p(t) \leq \bar{P}, \forall t \in \mathcal{T}_k$, $\mathcal{D}(Q(kT))$ is upper bounded by

$$\mathcal{D}(Q(kT)) \leq B_1 T + V\mathbb{E}\left\{\sum_{t \in \mathcal{T}_k} \mathcal{E}(t) \middle| Q(kT)\right\} + \mathbb{E}\left\{\sum_{t \in \mathcal{T}_k} Q(t) [X(t) - \epsilon] \middle| Q(kT)\right\}. \quad (10)$$

Here, $B_1 \triangleq \frac{1}{2}(\rho + \epsilon^2)$ is a constant.

Proof: See Appendix A. ■

The upper bound given in Lemma 1 [i.e., the R.H.S. of (10)] is widely used in the single-timescale control problems [20] (i.e., frame size $T = 1$). However, it is difficult to be applied directly to the two-timescale case since minimizing the R.H.S. of (10) at the beginning of every frame $t = kT$ requires the future information of $\{Q(t)\}$ over $t \in [kT+1, \dots, (k+1)T-1]$, which

is hard to be predicted in practice due to its accumulative nature over time slots. To address this issue, we further relax the R.H.S. of (10) as shown in the following lemma [31]–[33].

Lemma 2: Under any feasible decisions $n(k) \in \mathcal{N}$ and $0 \leq p(t) \leq \bar{P}, \forall t \in \mathcal{T}_k$, we have

$$\mathcal{D}(Q(kT)) \leq B_2 T + \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} V \mathcal{E}(t) + Q(kT) [X(t) - \epsilon] \middle| Q(kT) \right\}. \quad (11)$$

Here, $B_2 \triangleq B_1 + (T - 1)[(1 - \epsilon)\rho + \epsilon^2]/2$ is a constant.

Proof: See Appendix B. ■

The upper bound in Lemma 2 is derived from the R.H.S. of (10) by approximating the future queue length values as the current value at slot kT , i.e., $Q(t) \approx Q(kT)$ for all $t \in [kT + 1, \dots, (k + 1)T - 1]$. This approximation avoids the prediction of future queue lengths, which significantly reduces the complexity and suits more on the two-timescale design. Furthermore, as will be proved in Section V-D, this approximation preserves the asymptotic optimality of our proposed algorithm.

B. Algorithm Design

We now present the online two-timescale algorithm design. The idea of the algorithm is to minimize the drift-plus-penalty upper bound in (11) (i.e., the second term on the R.H.S.), subject to the constraints (6c) and (6d), which can be proved to achieve a good performance for the original Problem (P1). Specifically, our algorithm works in an online manner and takes the following three control actions:

- **(Migration decision per frame)** At time slot $t = kT$, with $k = 0, 1, \dots$, the user observes $Q(kT)$, $n(k - 1)$ and $f_n(k), \forall n$, and decides the optimal BS association $n^*(k)$ by solving the following per-frame problem:

$$\begin{aligned} \min_{n(k), \{p(t)\}} \quad & \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} V \mathcal{E}(t) + Q(kT) X(t) \right\} \\ \text{s.t.} \quad & n(k) \in \mathcal{N}, \quad 0 \leq p(t) \leq \bar{P}, \quad \forall t \in \mathcal{T}_k, \end{aligned} \quad (12)$$

The expectation $\mathbb{E}\{\cdot\}$ here is taken over the task arrival $a(t)$ and the channel randomness $\{h_n(t), \forall n\}$, for all $t \in \mathcal{T}_k$.

- **(Power allocation per slot)** At every slot $t \in \mathcal{T}_k$, given the BS association $n(k)$, the user observes the real-time channel condition $h_{n(k)}(t)$ and task arrival $a(t)$, and decides the power allocation $p^*(t)$ by solving the following per-slot problem:

$$\min_{0 \leq p(t) \leq \bar{P}} V \mathcal{E}(t) + Q(kT) X(t). \quad (13)$$

- **(Queue update)** At each slot $t \in \mathcal{T}_k$, based on the obtained $p^*(t)$, compute $X(t)$ by (3) and update the virtual queue $Q(t)$ according to (7).

We next develop the optimal solutions to the subproblems (12) and (13), respectively, which are the two building blocks for the algorithm implementation.

V. ALGORITHM IMPLEMENTATION AND PERFORMANCE ANALYSIS

In this section, we derive the optimal power strategy and the optimal migration policy for solving the per-slot Problem (13) and the per-frame Problem (12), respectively. We also discuss the optimal migration mechanism for some special cases and analyze the algorithm performance in the end.

A. Real-Time Power Allocation

For the per-slot power allocation Problem (13), first we can easily obtain that $p^*(t) = 0$ in two cases: 1) when $t \in \mathcal{T}_k^c$, i.e., during the service migration slots; and 2) when $a(t) = 0$, i.e., no task arrival at slot t .

For the residual case that tasks arrive at the offloadable slots, i.e., $t \in \mathcal{T}_k \setminus \mathcal{T}_k^c$ with $a(t) = 1$, we can rewrite the corresponding per-slot Problem (13) conditioned on $n(k) = n$ as

$$z_n(t) \triangleq \min_{0 \leq p(t) \leq \bar{P}} VE(n, p(t)) \cdot \mathbb{1}_{\{D(n, p(t)) \leq \tau_d\}} + Q(kT) \cdot \mathbb{1}_{\{D(n, p(t)) > \tau_d\}}, \quad (14)$$

where (14) is derived from (13) by expanding $\mathcal{E}(t)$ and $X(t)$ according to the definitions in (5) and (3).

From (14), we can observe that based on whether the latency requirement is met, the user can choose to consume $E(n, p(t))$ amount of energy to offload the arrived task, or choose not to offload at the expense of $Q(kT)$. The virtual queue length $Q(kT)$ here acts as the price of dropping a task. A higher $Q(kT)$ emphasizes more on reliability, i.e., the arrived tasks should be successfully offloaded as much as possible; while a lower $Q(kT)$ prefers energy saving and tolerates more task failures. Intuitively, through the queue evolution, the performance of energy consumption and task failure can adaptively be coordinated over frames.

Next, we specify the optimal power strategy for Problem (14) as follows.

Proposition 1 (*Optimal Power Strategy for Per-Slot Offloading*): The optimal transmit power for Problem (14) is given by

$$p^*(t) = \begin{cases} p_n^{\min}(t), & \text{if } p_n^{\min}(t) \leq p_n^{\max}(k), \\ 0, & \text{otherwise,} \end{cases} \quad (15)$$

where $p_n^{\min}(t)$ is the minimum transmit power at slot t to meet the task latency requirement, while $p_n^{\max}(k)$ denotes the maximum power allowed for per-slot offloading during k -th frame, which are respectively defined as:

$$p_n^{\min}(t) \triangleq \frac{\sigma}{h_n(t)} \left(2^{W \left[\tau_d - \frac{\xi}{f_n(k)} \right]^+} - 1 \right), \quad (16)$$

$$p_n^{\max}(k) \triangleq \min \left\{ \frac{Q(kT)}{V \left[\tau_d - \frac{\xi}{f_n(k)} \right]^+}, \bar{P} \right\}. \quad (17)$$

Proof: It can be checked from (1) and (2) that $D(n, p(t))$ is monotonically decreasing while $E(n, p(t))$ is monotonically increasing with $p(t)$, $\forall n \in \mathcal{N}$. By letting $D(n, p(t)) = \tau_d$, we obtain $p_n^{\min}(t)$ in (16) as the minimum required power for meeting the latency constraint, and $p^*(t) = p_n^{\min}(t)$ to achieve the minimum energy consumption in each task offloading.

We also note that when $VE(n, p_n^{\min}(t)) > Q(kT)$ in (14), i.e., the minimum energy consumption required for task offloading is higher than the task-dropping price, the task should be dropped for saving energy, thus $p^*(t) = 0$. Let $VE(n, p_n^{\min}(t)) = Q(kT)$ and further incorporate the peak power constraint (6c), we can obtain $p_n^{\max}(k)$ in (17) and the condition $p_n^{\min}(t) < p_n^{\max}(k)$ in (15), which completes the proof. ■

Proposition 1 reveals that the optimal power strategy for Problem (14) follows a *threshold-based policy*. When $p_n^{\min}(t)$ is below the threshold $p_n^{\max}(k)$, the user offloads the arrived task in power $p_n^{\min}(t)$; otherwise, the user should drop the task (i.e., $p^*(t) = 0$) to avoid excessive energy consumption. Notably, $p_n^{\min}(t)$ in (16) changes over each slot, adapting to the real-time channel condition $h_n(t)$, while threshold $p_n^{\max}(k)$ in (17) remains unchanged within a frame but it is adjusted from one frame to another according to the updated $Q(kT)$.

B. Migration Decision Per Frame

In this subsection, we find the optimal $n^*(k)$ by solving the per-frame Problem (12). Recall that Problem (12) is an expectation minimization problem. In order to compute the expectation, we make assumptions [32], [33] that the channel randomness is independent and identically distributed (i.i.d.) over the slots of a frame, and that the user has the statistical knowledge of channels in the current frame (but not the future frames).

According to the optimal power strategy in Proposition 1, we can derive the expected optimal per-slot performance for Problem (14) as follows.

Theorem 1: Suppose that $h_n(t)$ is i.i.d. over the slots of a frame with the probability density function (PDF) denoted by $f_{h_n}(k)$. Then, for all $t \in \mathcal{T}_k \setminus \mathcal{T}_k^c$, the expectation of $z_n(t)$ taken over the channel randomness $h_n(t)$ is obtained as

$$\mathbb{E}\{z_n(t)\} = Ve_n(k) \int_{h_n^{\min}(k)}^{\infty} \frac{1}{h} f_{h_n}(k) dh + Q(kT) \Pr[h_n(t) < h_n^{\min}(k)] \triangleq Z_n(k), \quad (18)$$

where $\Pr[\cdot]$ is the probability function, $e_n(k) \triangleq \sigma[\tau_d - \frac{\xi}{f_n(k)}]^+ \left(2^{\frac{L}{W[\tau_d - \frac{\xi}{f_n(k)}]^+}} - 1\right)$, and

$$h_n^{\min}(k) \triangleq \frac{\sigma}{p_n^{\max}(k)} \left(2^{\frac{L}{W[\tau_d - \frac{\xi}{f_n(k)}]^+}} - 1\right) \quad (19)$$

is the minimum threshold of channel gain to launch task offloading. In other words, task dropping occurs when $h_n(t) < h_n^{\min}(k)$.

Proof: According to Proposition 1 and by comparing $p_n^{\min}(t)$ with $p_n^{\max}(k)$, we can derive

$$z_n(t) = \begin{cases} Vp_n^{\min}(t) \left[\tau_d - \frac{\xi}{f_n(k)}\right]^+ = \frac{Ve_n(k)}{h_n(t)}, & \text{if } h_n(t) > h_n^{\min}(k), \\ Q(kT), & \text{otherwise,} \end{cases} \quad (20)$$

for each slot $t \in \mathcal{T}_k \setminus \mathcal{T}_k^c$. Taking the expectation on $z_n(t)$ over the random variable $h_n(t)$, we can obtain $\mathbb{E}\{z_n(t)\}$ as in (18). Since $h_n(t)$ follows the same distribution among slots $t \in \mathcal{T}_k$, $\mathbb{E}\{z_n(t)\}$'s are identical for all $t \in \mathcal{T}_k \setminus \mathcal{T}_k^c$, which completes the proof. ■

Different from $z_n(t)$ in (14), $Z_n(k)$ in (18) represents the minimum expected execution cost (i.e., weighted sum of energy consumption and task-dropping cost) for each slot $t \in \mathcal{T}_k \setminus \mathcal{T}_k^c$ with task arrival and under a stationary channel environment. Note that $e_n(k)$ and $h_n^{\min}(k)$ in (18) are known constants to the user, since $f_n(k)$ and $Q(kT)$ (that affects $p_n^{\max}(k)$) are known at the beginning of the k -th frame. Therefore, with the statistical knowledge of channels, the user is able to compute $Z_n(k)$ by (18) at the beginning of each frame $t = kT$.

We define $Z_n^{\text{sum}}(k)$ as the optimal objective value of the per-frame Problem (12) under the association $n(k) = n$, i.e.,

$$Z_n^{\text{sum}}(k) \triangleq \min_{\substack{0 \leq p(t) \leq \bar{P} \\ \forall t \in \mathcal{T}_k}} \mathbb{E}\left\{ \sum_{t \in \mathcal{T}_k} V\mathcal{E}(t) + Q(kT)X(t) \mid n(k) = n \right\}, \quad \forall n. \quad (21)$$

As $a(t)$ and $h_n(t)$ are i.i.d. over slots $t \in \mathcal{T}_k$, $Z_n^{\text{sum}}(k)$ in (21) can be decoupled into T independent per-slot problems with expectation minimization, each solved by the optimal power

strategies discussed in the last subsection. Hence, we can further express $Z_n^{sum}(k)$ as follows (see Appendix C):

$$Z_n^{sum}(k) = \begin{cases} \rho(T - C)Z_n(k) + \rho CQ(kT), & \text{if } n \neq n(k-1) \\ \rho T Z_n(k), & \text{if } n = n(k-1) \end{cases}, \quad \forall n, k. \quad (22)$$

Then, the optimal migration decision $n^*(k)$ for the k -th frame can be obtained by

$$n^*(k) = \arg \min_{n \in \mathcal{N}} \{Z_n^{sum}(k)\}. \quad (23)$$

From (22), we can see that the migration operation causes an expected ρC amount of task failure, which is a constant independent of which BS the user chooses to migrate to. Thus, for the BS set $n \in \mathcal{N} \setminus n(k-1)$, we have $\arg \min_{n \in \mathcal{N} \setminus n(k-1)} \{Z_n^{sum}(k)\} = \arg \min_{n \in \mathcal{N} \setminus n(k-1)} \{Z_n(k)\}$. Using this result, we can express the optimal migration decision (23), in the form of the following migration policy:

$$n^*(k) = \begin{cases} n', & \text{if } (1 - \alpha)Z_{n'}(k) + \alpha Q(kT) \leq Z_{n(k-1)}(k), \\ n(k-1), & \text{otherwise,} \end{cases} \quad (24)$$

where $n' \triangleq \arg \min_{n \in \mathcal{N} \setminus n(k-1)} \{Z_n(k)\}$ and $\alpha \triangleq \frac{C}{T}$, with $0 \leq \alpha \leq 1$, denoting the ratio of migration delay to a frame length.

The policy (24) suggests that the user always chooses migrating to the BS with the smallest $Z_n(k)$ whenever it performs a service migration, and the migration occurs only if condition $(1 - \alpha)Z_{n'}(k) + \alpha Q(kT) \leq Z_{n(k-1)}(k)$ is met.

By incorporating the above migration policy and power strategy into the algorithm framework, we summarize the proposed online algorithm in Algorithm 1.

C. Properties of Optimal Migration Policy

In this subsection, we derive additional insights into the migration policy in (24) for a concrete channel model. Specifically, we assume the channel power gain $h_n(t)$, $\forall t \in \mathcal{T}_k$ and $\forall n \in \mathcal{N}$, can be represented by

$$h_n(t) = g_n(t)H_n(k), \quad (25)$$

where $g_n(t)$ accounts for the small-scale fading power component at slot t and $H_n(k)$ represents the large-scale fading power component in the k -th frame. $g_n(t)$ is assumed to be i.i.d. unit mean exponential random variables, i.e., the Rayleigh fading model considered for the fast

Algorithm 1 Online Two-Timescale Algorithm

- 1: Set $V \geq 0$, $\epsilon \in (0, 1)$, and $n(-1)$ as the user's current associated BS.
 - 2: Initialize $t = 0$ and $Q(0) = 0$.
 - 3: **for** each frame $k = 0, 1, \dots, K - 1$ **do**
 - 4: Compute $n(k)$ by (24).
 - 5: Set $\mathcal{T}_k^c = \{kT, kT + 1, \dots, kT + C - 1\}$ if $n(k) \neq n(k - 1)$, and otherwise $\mathcal{T}_k^c = \{\emptyset\}$.
 - 6: **for** each slot $t = kT, kT + 1, \dots, kT + T - 1$ **do**
 - 7: **if** $t \in \mathcal{T}_k \setminus \mathcal{T}_k^c$ and $a(t) = 1$ **then**
 - 8: Compute $p(t)$ by (15).
 - 9: **else**
 - 10: Set $p(t) = 0$.
 - 11: **end if**
 - 12: Compute $X(t)$ and $\mathcal{E}(t)$ by (3) and (5), respectively.
 - 13: Update $Q(t)$ by (7).
 - 14: **end for**
 - 15: **end for**
- Output:** $\{n(k)\}$ and $\{p(t)\}$.
-

fading. $H_n(k)$ captures the path loss and shadowing whose changes matches the timescale of a frame.

Building on the above channel model, we show in the sequel that the migration policy (24) has more straightforward migration mechanism for several special cases.

1) *Homogenous Computation Rates:* Consider the case of $f_n(k) = f(k), \forall n$, where $f(k) > \frac{\xi}{\tau_d}$ for the edge-execution feasibility. Then, $\{e_n(k), h_n^{\min}(k)\}$ are identical for all n (see Theorem 1), and can be re-notated by $\{e(k), h^{\min}(k)\}$. We show that in this case, the migration decision for the k -th frame can be determined by simply comparing the parameter $H_n(k)$ of each BS.

Proposition 2 (Homogenous Computation Rates): Assume that $f_n(k) = f(k) > \frac{\xi}{\tau_d}$, for all n . The following properties hold:

- a) $n' = \operatorname{argmax}_{n \in \mathcal{N} \setminus n(k-1)} \{H_n(k)\}$, i.e., if the user needs a service migration, it will always choose migrating to the BS with the highest $H_n(k)$.

b) The user keeps associating with the serving BS $n(k-1)$ if $H_{n(k-1)}(k)$ satisfies

$$H_{n(k-1)}(k) > \frac{h_n^{\min}(k)}{\ln\left(\frac{1}{1-\alpha}\right)} \triangleq h_\alpha(k). \quad (26)$$

c) When the condition (26) becomes invalid, the user migrates from BS $n(k-1)$ to BS n' if

$$H_{n(k-1)}(k) \leq H_{n'}(k) \cdot \min \left\{ \frac{h_\alpha(k)}{h_\alpha(k) + H_{n'}(k)}, \frac{1}{2} \right\}. \quad (27)$$

Proof: See Appendix D. ■

Remark 1 (Migration Policy): Proposition 2 reveals straightforward migration policies for this case. First, in each frame the user always selects the nearest BS (i.e., the highest $H_n(k)$), if all the BSs have the same computing rate. Second, when the channel gain of the serving BS is above a threshold specified by (26), there is no need of migration. Finally, when the condition (26) becomes invalid, a migration is triggered if the new association can obtain a sufficient channel enhancement as specified by (27).

2) *Heterogenous Computation Rates:* Here we consider heterogenous computation rates by assuming no peak power constraint (6d). Note that the transmit power still is bounded by $p_n^{\max}(k) = \frac{Q(kT)}{V[\tau_d - \frac{\xi}{f_n(k)}]^+}$ in (17). We obtain for this case the migration decision relies on two parameters $H_n(k)$ and $h_n^{\min}(k)$, in which $h_n^{\min}(k)$ [see (19)] is monotonically decreasing with $f_n(k)$.

Proposition 3 (Heterogenous Computation Rates): Assume that $\bar{P} = \infty$ ¹ and $f_n(k) > \frac{\xi}{\tau_d}$, for all n . Let $\nu_n(k) \triangleq \frac{h_n^{\min}(k)}{H_n(k)}$, for all $n \in \mathcal{N}$. The following properties hold:

a) $n' = \operatorname{argmin}_{n \in \mathcal{N} \setminus n(k-1)} \{\nu_n(k)\}$.

b) The user keeps associating with the serving BS $n(k-1)$ if $\nu_n(k)$ itself satisfies

$$\nu_n(k) < \ln\left(\frac{1}{1-\alpha}\right). \quad (28)$$

c) The user migrates the association from BS $n(k-1)$ to BS n' if

$$\nu_{n(k-1)}(k) \geq \max \left\{ \nu_{n'}(k) + \ln\left(\frac{1}{1-\alpha}\right), 2\nu_{n'}(k) \right\}. \quad (29)$$

Proof: See Appendix E. ■

Proposition 3 shows that for this case, the computation rate affects the migration decision through the minimum channel threshold h_n^{\min} [see (19)], with h_n^{\min} being decreasing as $f_n(k)$ increases, and the migration policy is similar to that of the preceding case but works on the basis of $\nu_n(k)$.

¹Proposition 3 also holds for the finite peak power as long as $\bar{P} > \frac{Q(kT)}{V[\tau_d - \frac{\xi}{f_n(k)}]^+}$ is met, $\forall n$.

D. Performance Analysis

In this subsection, we present the performance bounds of the proposed algorithm. For ease of analysis, we assume that the system randomness is i.i.d. over frames and that Problem (P1) is feasible. As such, the feasibility implies that there exists a slack constant $\delta > 0$ and a feasible solution to Problem (P1) such that the following inequality holds for all k :

$$\frac{1}{T} \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} X(t) \right\} < \epsilon - \delta. \quad (30)$$

Based on this, we have the following theorem for theoretically quantifying the performance bounds that the proposed algorithm can achieve.

Theorem 2: Assume that the condition (30) is satisfied for $\exists \delta > 0$, and the initial virtual queue length is zero, i.e., $Q(0) = 0$. Then, for any $V > 0$, we have:

1) The average queue length under the proposed algorithm is upper bounded by

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \{ Q^*(kT) \} \leq \frac{B_2 + V E_{\max}}{\delta}, \quad (31)$$

where $Q^*(t)$ denotes the resultant queue length by the proposed algorithm and $E_{\max} \triangleq \bar{P} \cdot \max_{n \in \mathcal{N}} \left\{ \tau_d - \frac{\xi}{f_n^{\max}} \right\}$.

2) The average energy consumption achieved by the proposed algorithm satisfies

$$\lim_{K \rightarrow \infty} \frac{1}{KT} \sum_{k=0}^{K-1} \sum_{t \in \mathcal{T}_k} \mathbb{E} \{ \mathcal{E}^*(t) \} \leq E_{\text{av}}^{\text{opt}} + \frac{B_2}{V}, \quad (32)$$

where $\mathcal{E}^*(t)$ denotes the resultant energy consumption by the proposed algorithm and $E_{\text{av}}^{\text{opt}}$ denotes the minimum average energy consumption for Problem (P1).

Proof: See Appendix F. ■

Theorem 2 shows that the average energy consumption of the online algorithm can asymptotically achieve the optimum $E_{\text{av}}^{\text{opt}}$ of the original Problem (P1) by increasing the control parameter V . Besides, the average virtual queue length is bounded by $\mathcal{O}(V)$ in (31), indicating the queue is mean rate stable and the reliability constraint (6b) is guaranteed.

VI. EXTENSION TO MULTIUSER MANAGEMENT

In this section, we consider the multiuser mobility management under the proposed two-timescale framework. Multiuser migrations could noticeably change the load of BSs and affect

the computation rates for other users associated at the same BS. Thus, compared with the single-user case, multiuser management requires considering the load balance factor among BSs when making users' migration/association decisions.

Specifically, we consider that M users, denoted by set $\mathcal{M} = \{1, \dots, M\}$, are randomly distributed and move in the multi-cell network. We assume that each user is allocated with a dedicated channel (like in OFDMA) for multiuser offloading [11], [20], [33]. Regarding the multiuser computing, we use the number of associated users $y_n(k)$ to represent the load of BS n in the k -th frame, and model the computation rate for user i as a function of $y_n(k)$ [10], [34], which are respectively given by²

$$\begin{cases} y_n(k) = \sum_{i \in \mathcal{M}} \mathbb{1}_{\{n_i(k)=n\}}, & \forall n \in \mathcal{N}, \\ f_{i,n}(k) = F_{i,n} \alpha_n^{y_n(k)-1}, & \forall n \in \mathcal{N}, \quad \forall i \in \mathcal{M}, \end{cases} \quad (33)$$

where $n_i(k)$ denotes the association decision of user i and $\alpha_n \in (0, 1)$ is the degradation factor that specifies the percentage decrease of user's computation rate as the BS load increases.

The models of task arrival, service migration, and two-timescale operation follow the same settings in the single-user case for each user. Our goal in multiuser management is to minimize the sum of users' time-averaged energy consumption while guaranteeing the reliability requirement of each user, which is formulated as

$$(P2) \quad \min_{\{n_i(k)\}, \{p_i(t)\}} \lim_{K \rightarrow \infty} \frac{1}{KT} \sum_{k=0}^{K-1} \sum_{t \in \mathcal{T}_k} \sum_{i \in \mathcal{M}} \mathbb{E} \{ \mathcal{E}_i(t) \} \quad (34a)$$

$$\text{s.t.} \quad \lim_{K \rightarrow \infty} \frac{1}{KT} \sum_{k=0}^{K-1} \sum_{t \in \mathcal{T}_k} \mathbb{E} \{ X_i(t) \} \leq \epsilon_i, \quad \forall i \in \mathcal{M}, \quad (34b)$$

$$n_i(k) \in \mathcal{N}, \quad \forall i \in \mathcal{M}, \quad k = 0, 1, \dots \quad (34c)$$

$$0 \leq p_i(t) \leq \bar{P}_i, \quad \forall i \in \mathcal{M}, \quad t = 0, 1, \dots \quad (34d)$$

(33) for per-frame computation resource allocation.

Apart from the addition of subscript i to denote the user index, all notations in the above Problem (P2) and their corresponding expressions remain the same as the single-user case.

Similarly, we can develop the Lyapunov-based online algorithm to solve Problem (P2). The algorithm framework is similar to Algorithm 1 (see Section IV-B) and consists of three control

²Other load-aware computation models, such as equal resource allocation among the users at a BS, are also applicable to our proposed management scheme.

actions: multiuser migration decisions per frame, user's power control per slot, and the virtual queue update. The last two actions are executed on each user in parallel and consistent with the results of the single-user case, i.e., each user carries out the power strategy in Proposition 1 and update its queue according to (7) at every slot. Thus, in what follows we focus on the per-frame multiuser migration problem.

A. Per-frame Multiuser Migration Problem

At the beginning of each frame k , based on the observation of $\{Q_i(kT)\}$ and $\{n_i(k-1)\}$, the network operator decides the users' associations $\{n_i(k)\}$ by solving the following per-frame problem:

$$\begin{aligned} \min_{\{n_i(k)\}, \{p_i(t)\}} \quad & \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} \sum_{i \in \mathcal{M}} V \mathcal{E}_i(t) + Q_i(kT) X_i(t) \right\} \\ \text{s.t.} \quad & (33), (34\text{c}), (34\text{d}). \end{aligned} \quad (35)$$

Note that the user's computation rate in Problem (35) is no longer a known constant but a function of $\{n_i(k)\}$ due to the constraints (33). Thus the users' association decisions $\{n_i(k)\}$ are coupled in the per-frame Problem (35).

To facilitate exposition, we introduce a set of binary variables $\mathbf{X} = \{x_{i,n}\}$ to represent the users' association decisions $n_i(k)$, with $x_{i,n} = 1$ indicating $n_i(k) = n$ and $x_{i,n} = 0$ otherwise. Then, by incorporating the power strategy in Proposition 1 and taking the channel assumptions as in the single-user case, i.e., i.i.d. channel randomness over the slots of a frame and the channel statistics in a frame being known, we can refine Problem (35) as the following problem:

$$\min_{\mathbf{X} \in \{0,1\}^{M \times N}} \quad R(\mathbf{X}) = \sum_{i \in \mathcal{M}} \sum_{n \in \mathcal{N}} x_{i,n} Z_{i,n}^{\text{sum}}(y_n) \quad (36\text{a})$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{M}} x_{i,n} = y_n, \quad \forall n \in \mathcal{N}, \quad (36\text{b})$$

$$\sum_{n \in \mathcal{N}} x_{i,n} = 1, \quad \forall i \in \mathcal{M}. \quad (36\text{c})$$

Here we drop the frame index k here for ease of notation. $Z_{i,n}^{\text{sum}}(y_n)$ in (36a) is the minimum expected cost of user i if associating with BS n in the current frame given the BS load y_n (that determines $f_{i,n}$). The expression of $Z_{i,n}^{\text{sum}}(y_n)$ is the same as $Z_n^{\text{sum}}(k)$ of the single-user case and given by (22). $R(\mathbf{X})$ is the sum of users' association cost, which can be regarded as the overall system cost. The constraints (36b) and (36c) are equivalent to (33) and (34c), respectively.

Algorithm 2 Suboptimal Algorithm for Solving Problem (36)

- 1: Initialize $l = 1$, $s = 1$, and set $\mathbf{X}^{(1)}$ as the current user-BS associations.
- 2: Find the worst user-BS association $(i^{(1)}, n^{(1)})$ by (38).
- 3: **repeat**
- 4: Generate $(N - 1)$ new association matrixes $\mathbf{X}_m^{(l)}$, $\forall m \in \mathcal{N} \setminus n^{(l)}$, each by switching the $i^{(l)}$ -th user's association from BS $n^{(l)}$ to BS m .
- 5: Update $\mathbf{X}^{(l+1)}$ by (37).
- 6: Find the worst user-BS association $(i^{(l+1)}, n^{(l+1)})$ by (38).
- 7: **if** $\mathbf{X}^{(l+1)} = \mathbf{X}^{(l)}$ **then**
- 8: Let $i' \leftarrow \text{mod}(s - 1, M) + 1$ and n' be the index of the i' -th user associated BS.
- 9: $(i^{(l+1)}, n^{(l+1)}) \leftarrow (i', n')$.
- 10: $s \leftarrow s + 1$.
- 11: **end if**
- 12: $l \leftarrow l + 1$.
- 13: **until** $\mathbf{X}^{(l)}$ remains unchanged after M consecutive iterations.

Output: $\mathbf{X}^{(l)}$.

Note that due to the binary variables \mathbf{X} , Problem (36) is an integer nonlinear programming problem that is hard to obtain the optimal solution in efficient time-complexity. For this reason, we find a near-optimal solution by developing a low-complexity algorithm in the next subsection.

B. Algorithm Design for Multiuser Migration

We propose an efficient iterative algorithm which converges to a near-optimal solution to the migration Problem (36). The algorithm is based on the intuition that, the user with the worse BS association is more likely to trigger migration to another BS (with a stronger wireless link and/or less compute load), which consequently reduces association cost of the user and the system. Motivated by this, we design an algorithm centering on the worst user-BS association improvement.

The algorithm for solving Problem (36) is presented in Algorithm 2. It starts by initializing the user-BS associations $\mathbf{X}^{(1)}$ and finding the worst user-BS association $(i^{(1)}, n^{(1)})$ from $\mathbf{X}^{(1)}$ as described later via (38). At each iteration l , the algorithm goes through the following two steps:

1) *Association update*: In this step, we adjust the association decision of user $i^{(l)}$, which is equal to update the whole association matrix $\mathbf{X}^{(l)}$ under the entries of other users being fixed. Specifically, we generate $(N - 1)$ new association matrixes $\mathbf{X}_m^{(l)}$, each representing user $i^{(l)}$ is migrated to other BS m , with $m \in \mathcal{N} \setminus n^{(l)}$. Then, among the current and new association matrixes, we choose the one with the minimum system cost $R(\mathbf{X})$ [see (36a)] as the best association matrix for the next iteration:

$$\mathbf{X}^{(l+1)} = \arg \min_{\mathbf{X}^{(l)}, \{\mathbf{X}_m^{(l)} | \forall m \in \mathcal{N} \setminus n^{(l)}\}} R(\mathbf{X}). \quad (37)$$

We can observe from (37) that if $\mathbf{X}^{(l+1)} \neq \mathbf{X}^{(l)}$, the system cost $R(\mathbf{X})$ [i.e., the objective value of Problem (36)] is always decreasing in the association update.

Next, based on the updated $\mathbf{X}^{(l+1)}$, we find the worst user-BS association $(i^{(l)}, n^{(l)})$ for the next iteration by comparing the users' association costs:

$$(i^{(l+1)}, n^{(l+1)}) = \arg \max_{i,n} \left\{ x_{i,n}^{(l+1)} Z_{i,n}^{sum}(y_n^{(l+1)}) \right\}, \quad (38)$$

where $y_n^{(l+1)}$ is computed according to (36b).

2) *User switching*: When $\mathbf{X}^{(l+1)} = \mathbf{X}^{(l)}$, it means that the system cost can not be further reduced by adjusting the association of the worst user $i^{(l)}$. Furthermore, the worst users are equal (i.e., $i^{(l+1)} = i^{(l)}$) in the following iterations, leading to no more changes in the system cost. Clearly, in order to find a potential system-cost reduction, we need to switch another user to adjust its association. To this end, we introduce a user switching step. At each switching step s , let $i' = \text{mod}(s - 1, M) + 1$, and we select user i' instead of the current worst user $i^{(l+1)}$ to perform the association update in the next iteration. Note that the switching process does not affect the non-increasing property of the system cost in the association update.

The iteration process is repeated until $\mathbf{X}^{(l)}$ remains unchanged after M consecutive iterations. The convergence is guaranteed because the system cost always keeps non-increasing in iterations, and all M users have been swept by the switching process and have no association changes when the stopping condition is met.

VII. SIMULATION RESULTS

The simulation settings are as follows. We consider that $N = 25$ BSs are regularly deployed in a 2 km \times 2 km square area. The slot length and the frame size are set to be $\tau = 10$ ms and $T = 500$ slots, respectively. The time horizon is $K = 2500$ frames. We consider a task type with

TABLE I: System Parameters

Parameter	Value
Number of BSs, N	25
Slot length, τ	10 ms
Frame size, T	500 slots
Task arrival probability, ρ	0.5
Peak transmit power, \bar{P}	1 W
BS computation rate, $f_n(k)$	$[1 \times 10^{10}, 2 \times 10^{10}]$ cycles/s
Service migration delay, C	5 slots [24]
Reliability threshold, ϵ	1×10^{-3}
Control parameter, V	5000

$L = 5$ Kbits, $\xi = 2640$ cycles/bit (such as 400 frame video game [20]), $\tau_d = 10$ ms, and $\rho = 0.5$. In terms of the user movement, we assume that the user's locations change over frames and adopt the Random Waypoint Mobility model [35] to generate the user's location for each frame, with the parameters taken as: the static probability and pause time $p_s = t_p = 0$, and the user's velocity $v = v_{\min} = v_{\max} = 5$ m/s. For task offloading, the channel power gains are modeled as $h_n(t) = g_n(t)H_n(k)$ in (25). The large-scale fading $H_n(k)$ is given by $127 + 30 \log_{10}(10^{-3}d_n(k))$, where $d_n(k)$ denotes the distance between the user and BS n in meter at the k -th frame. The small-scale fading $g_n(t)$ follows normalized exponential distribution. Besides, the noise power spectrum density is set as -174 dBm with $W = 10$ MHz channel bandwidth. For service migration, we consider $C = 5$, i.e., migration/handover delay is 50 ms [24]. Unless mentioned otherwise, the main communication and computation parameters used in the simulations are summarized in Table I.

For performance comparison, we also simulate the two traditional handover schemes as the benchmarks:

- 1) *Received signal strength (RSS) only*: For each frame, the user always migrates the association to the BS with the highest large-scale fading $H_n(k)$.
- 2) *RSS plus hysteresis*: Let $n' \triangleq \arg \max_{n \in \mathcal{N} \setminus n(k-1)} \{H_n(k)\}$ denote the target BS with the highest $H_n(k)$ at k -th frame. For each frame, the user migrates to the target BS if $H_{n'}(k) > (1 + \beta_{\text{th}})H_{n(k-1)}(k)$; otherwise it stays at the current BS. Here, β_{th} is a hysteresis margin and set as 2 in the simulations.

Note that the two benchmarks are used to decide the service migration for each frame. For per-slot offloading of the arrived tasks, we consider that both of them use the following power

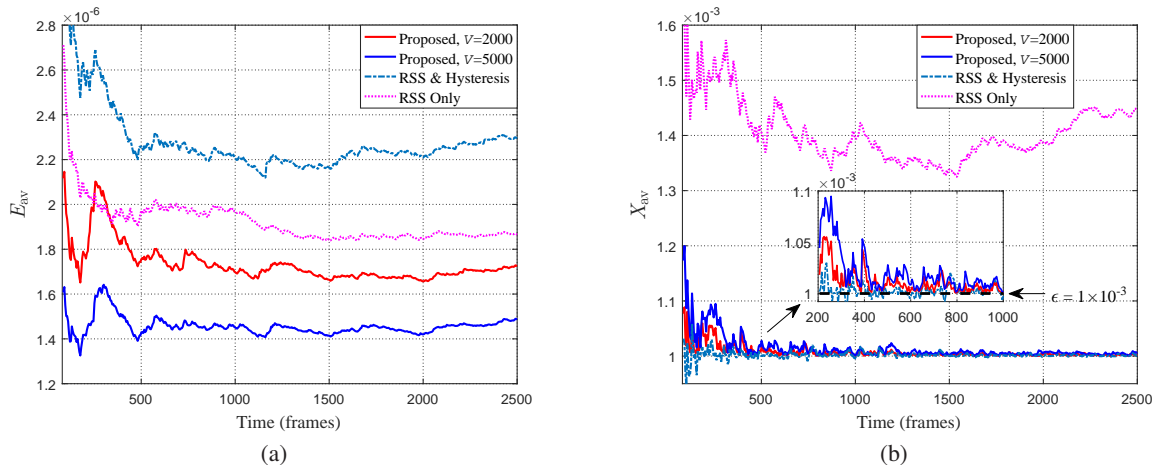


Fig. 3: Time evolution: (a) average energy consumption, and (b) average task-failure rate.

strategy:

$$p(t) = \begin{cases} p_{n(k)}^{\min}(t), & \text{if } t \in \mathcal{T}_k \setminus \mathcal{T}_k^c \text{ and } p_{n(k)}^{\min}(t) \leq p^{\max}(t), \\ 0, & \text{otherwise,} \end{cases} \quad (39)$$

where $p_{n(k)}^{\min}(t)$ is the minimum required power as in (16) and $p^{\max}(t) \triangleq \bar{P} \cdot \mathbb{1}_{\{X_{av}(t) > \epsilon\}} + 0.05\bar{P}$. $\mathbb{1}_{\{X_{av}(t) \leq \epsilon\}}$ is the maximum power threshold at slot t , which is set to be \bar{P} or $0.05\bar{P}$ depending on whether the current average task-failure rate $X_{av}(t)$ exceeds the threshold ϵ or not. Here, an online reliability control is made in (39) by setting two modes on $p^{\max}(t)$: $0.05\bar{P}$ that prefers energy saving, and \bar{P} that emphasizes more on reliability.

A. Single-user Case

Fig. 3 shows the average energy consumption and task-failure rate of the proposed Algorithm 1 and two benchmark schemes over 2500 time frames. First, it can be seen that our proposed algorithm in $V = 2000$ and $V = 5000$ both can achieve lower energy consumption than the two benchmarks while satisfying the reliability constraint. A larger value V in the proposed algorithm can save more energy; however, as shown in the local diagram of Fig. 3(b), its task-failure rate converges more slowly to the reliability threshold ϵ . Among the benchmarks, we can observe that the RSS only scheme has lower energy consumption but does not meet the reliability constraint caused by frequent service migrations; in contrast, the RSS plus hysteresis can reduce excessive migrations to ensure the reliability but at the expense of high energy consumption due to its delayed migration response.

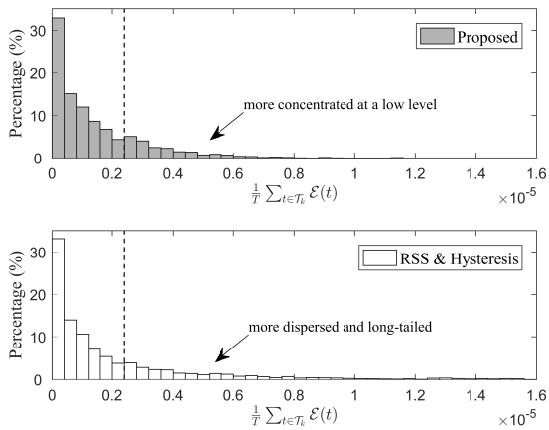


Fig. 4: Distribution of per-frame average energy consumption.

In Fig. 4, we compare the distribution of per-frame average energy consumption between the RSS plus hysteresis scheme and the proposed Algorithm 1 with $V = 5000$. It can be observed that, during the interval $[0.22 \times 10^{-5}, \infty)$, the energy distribution of the proposed algorithm is more centralized at a low level, while the distribution of the RSS plus hysteresis is dispersed and long-tailed at high energy level (e.g., $[0.6 \times 10^{-5}, \infty)$). Note that the service migration mainly serves for energy reduction at the high energy interval corresponding to the user's locations at the cell edge. Therefore, this demonstrates that compared to the RSS plus hysteresis scheme, our proposed algorithm can make more accurate and prompt migration decisions when the user moves across the BSs to reduce energy consumption.

Fig. 5 shows the impact of control parameter V on the average energy consumption, the task-failure rate, and the virtual queue length of our proposed Algorithm 1, where $\epsilon = 10^{-3}$. We can see that the energy consumption decreases inversely proportional to V ; the task-failure rate maintains satisfying the reliability constraint no matter what V is; and the average queue length increases linearly as V increases. These match the results in Theorem 2 that the performance of average energy consumption and queue length follows the $[\mathcal{O}(1/V), \mathcal{O}(V)]$ tradeoff.

Fig. 6 shows the energy-reliability tradeoff of all the algorithms by varying the threshold ϵ . We can observe that the proposed Algorithm 1 always achieves a smaller energy consumption than the two benchmarks under the same reliability requirement. The RSS only scheme performs well when $\epsilon > 3 \times 10^{-3}$. This is because when ϵ is large, the reliability loss in migration is tolerable and migrating the BS with the best channel for each frame helps reduce user's (transmit)

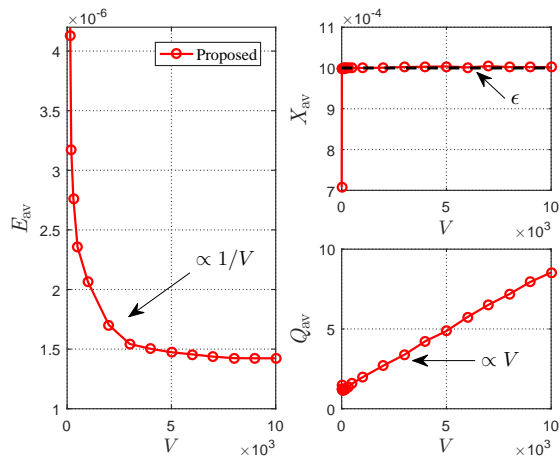
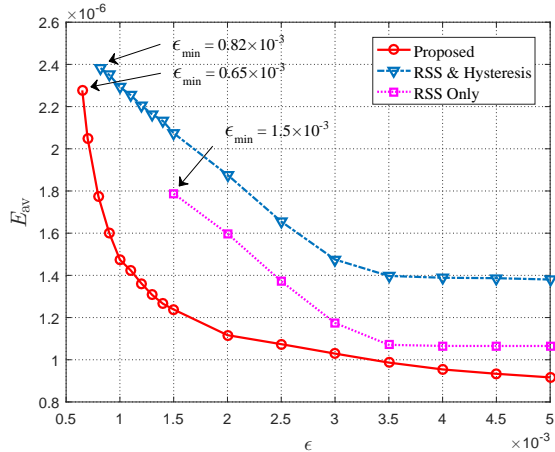
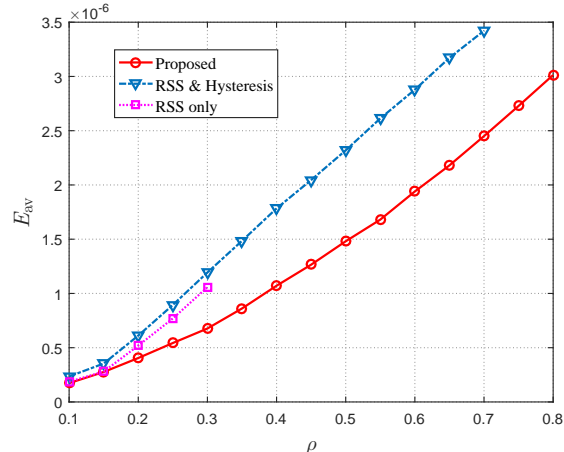
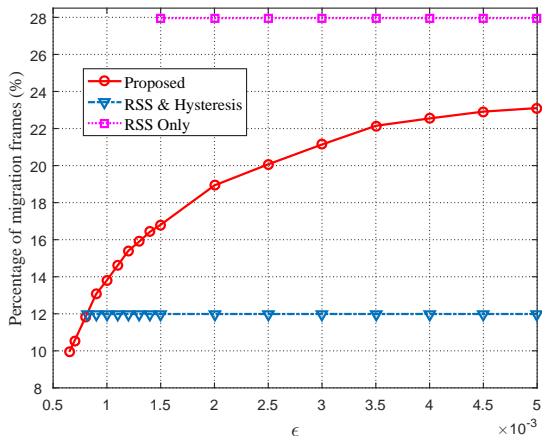
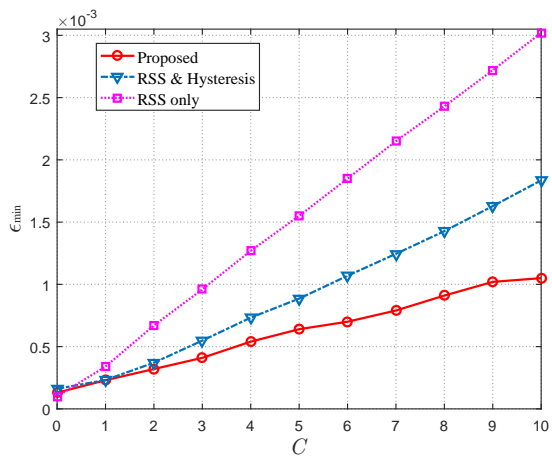


Fig. 5: Impact of control parameter V .

Fig. 6: Average energy consumption vs. ϵ .Fig. 7: Average energy consumption vs. ρ .Fig. 8: Percentage of migration frames vs. ϵ .Fig. 9: Minimum reliability threshold ϵ_{\min} vs. C .

energy consumption. However, it fails to fulfill the stringent reliability requirement due to its aggressive migration strategy. The RSS plus hysteresis can enhance the reliability performance compared to the RSS only, but it suffers high energy consumption. In contrast, our proposed algorithm outperforms the two benchmarks in both reliability and energy performance due to its joint management of service migration and computation offloading.

Fig. 7 shows the average energy consumption versus the task arrival probability ρ . As expected, the proposed Algorithm 1 achieves significant energy reduction compared to the two benchmarks under the same ρ . In addition, our proposed algorithm can accommodate a higher task arrival rate ρ than the two benchmarks to meet the reliability constraint.

In Fig. 8, we plot the percentage of migration frames among the total $K = 2500$ frames under

different reliability thresholds ϵ . As we observe, the RSS only and the RSS plus hysteresis have static migration percentage since their migration policies are only related to channel condition, while the proposed Algorithm 1 can adjust the migration percentage according to the reliability requirement. Combining with the energy behaviors as shown in Fig. 6, these demonstrate that our proposed algorithm performs more flexible migration-frequency control to balance the energy consumption and the reliability performance.

Finally, we evaluate the impact of service migration delay C on the reliability performance in Fig. 9, where the reliability performance is measured by the minimum threshold ϵ_{\min} that can be supported by the algorithms. First, we can see that ϵ_{\min} is increasing with C in all considered algorithms while the ascending rate of our proposed Algorithm 1 is the slowest, indicating that the proposed has the best reliability performance against the migration-delay effect. We also observe that the performance of two benchmarks is close to that of the proposed Algorithm 1 when C is small; however, they dramatically deteriorate as C becomes large. This is because the reliability loss (i.e., task failure) caused by migration is low when C is small, while it becomes a dominant factor and requires effective management when C goes large.

B. Multiuser Case

In this subsection, we verify the performance of our proposed Algorithm 2 in multiuser management. Similarly, we use the Random Waypoint Mobility model to generate the movement of multiple users, where each user moves in a constant speed v_i (m/s), which is randomly chosen from the set $\{2, 4, 6, 8, 10\}$. For multiuser computing, we set $F_{i,n} = 2 \times 10^{10}$ cycles/s and the degradation factor $\alpha_n = 0.926$. Other parameters for each user follow the same settings of the single-user case.

In Fig. 10, we plot the average users' energy consumption versus the number of users in the network M , under the proposed Algorithm 2 and the benchmark scheme of RSS plus hysteresis. We can observe that for both the proposed and the benchmark, the average users' energy consumption is insensitive to the increase of M when $M < 300$ while it begins to increase when $M > 300$. The reason is that, when M is small, each BS is lightly loaded and can provide stable computation rates; when M becomes large, the computation rate suffer severe degradation due to the overloaded BS. Nevertheless, the energy consumption of our proposed algorithm increases at slower rate than that of the benchmark scheme, thanks to its efficient load-aware migration mechanism to balance the workload among BSs.

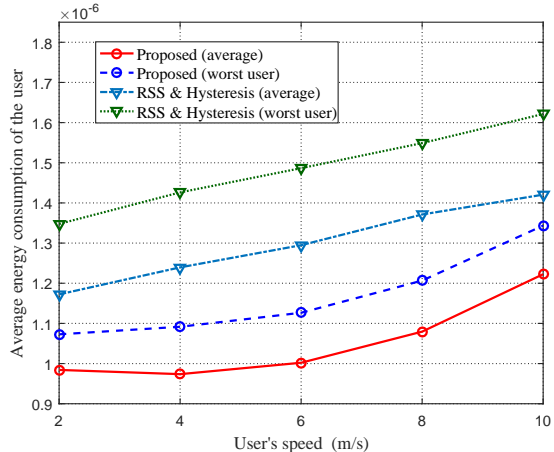
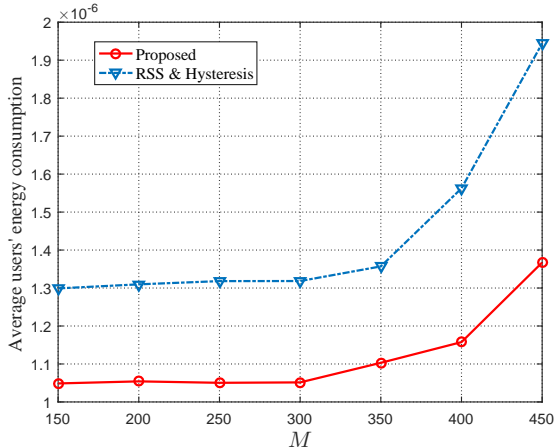


Fig. 10: Average users' energy consumption vs. M . Fig. 11: Average energy consumption of the users with different speeds in both the worst and average cases.

Fig. 10 shows the influence of user's speed on its average energy consumption, where the average performance and the performance of the worst user are considered, and M is set as 250 for this case. We can see that the energy consumption increases with the user's speed in both the propose Algorithm 2 and the RSS plus hysteresis scheme, due to the growth of migration demands. Compared to the RSS plus hysteresis scheme, our proposed algorithm has lower energy consumption and a smaller gap between the average and the worst-user performance. The first one is because our proposed algorithm can make more accurate and prompt migration decisions for every user as discussed in Fig. 10. The second one is because our multiuser migration strategy in the proposed algorithm centres on providing more migration chances to the users with worse BS associations (e.g., cell-edge users) to improve their performance.

VIII. CONCLUSIONS

In this paper, we study the mobility management problem in the multi-cell MEC network, with the goal of minimizing user's energy consumption subject to the reliability constraint for computation offloading. We propose a two-timescale approach with joint optimization of service migration and transmit power control, which is a low-complexity online algorithm and can achieve asymptotical optimality shown by the theoretical analysis. In our approach, the optimal power control for task offloading and the optimal migration policy for BS association, both follow a threshold-based structure. The former uses the threshold to make a binary offloading decision, while the latter uses it to decide whether to migrate from the current BS to the target.

These two thresholds are dynamically adjusted to balance the energy and reliability performance. We also extend our two-timescale approach to multiuser management by designing a load-aware multiuser migration scheme. Simulation results demonstrate the superior performance achieved by our approach, especially when the reliability requirement is stringent. For future investigation, we intend to consider a general case that the short-term mobility information is available to be leveraged, which is expected to achieve more proactive migrations. Another direction is considering the cooperative computing among BSs, as an alternative approach to service migration, to deal with user mobility.

APPENDIX

A. Proof of Lemma 1

According to the queue dynamics (7), we have

$$\begin{aligned}
Q(t+1)^2 - Q(t)^2 &= (\max\{Q(t) + X(t) - \epsilon, 0\})^2 - Q(t)^2 \\
&\stackrel{(a)}{\leq} [Q(t) + X(t) - \epsilon]^2 - Q(t)^2 \\
&\leq X(t)^2 + \epsilon^2 + 2Q(t)[X(t) - \epsilon],
\end{aligned} \tag{40}$$

where (a) is derived by $\max\{x, 0\}^2 \leq x^2$. Summing the above (40) over $t \in \{kT, \dots, (k+1)T-1\}$ and taking conditional expectation given $Q(kT)$, it follows that $\Delta_T(Q(kT))$ defined in (8) is upper bounded by

$$\begin{aligned}
\Delta_T(Q(kT)) &= \mathbb{E} \left\{ \frac{1}{2} Q((k+1)T)^2 - \frac{1}{2} Q(kT)^2 \middle| Q(kT) \right\} \\
&\leq \mathbb{E} \left\{ \frac{1}{2} \sum_{t \in \mathcal{T}_k} X(t)^2 + \frac{1}{2} \epsilon^2 T + \sum_{t \in \mathcal{T}_k} Q(t) [X(t) - \epsilon] \middle| Q(kT) \right\} \\
&\stackrel{(b)}{\leq} \frac{1}{2} \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} a(t) \right\} + \frac{1}{2} \epsilon^2 T + \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} Q(t) [X(t) - \epsilon] \middle| Q(kT) \right\} \\
&\stackrel{(c)}{=} \frac{1}{2} (\rho + \epsilon^2) T + \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} Q(t) [X(t) - \epsilon] \middle| Q(kT) \right\},
\end{aligned} \tag{41}$$

where (b) is using the facts that $X(t)^2 = X(t) \leq a(t)$ and $a(t)$ is independent of $Q(kT)$. Step (c) is because $a(t)$ is i.i.d. over slots with $\mathbb{E}\{a(t)\} = \rho$. Finally, letting $B_1 = (\rho + \epsilon^2)/2$ and adding $V \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} \mathcal{E}(t) \middle| Q(kT) \right\}$ into both sides of (41) yield the result (10).

B. Proof of Lemma 2

Since $X(t) \in \{0, 1\}$, the queue length $Q(t)$ for each slot $t \in \mathcal{T}_k$ is bounded by

$$Q(kT) - (t - kT)\epsilon \leq Q(t) \leq Q(kT) + (t - kT)(1 - \epsilon). \quad (42)$$

Using (42), it can be shown that the term $\sum_{t \in \mathcal{T}_k} Q(t)[X(t) - \epsilon]$ in (10) can be bounded as

$$\begin{aligned} \sum_{t \in \mathcal{T}_k} Q(t)[X(t) - \epsilon] &= \sum_{t \in \mathcal{T}_k} Q(t)X(t) - \sum_{t \in \mathcal{T}_k} \epsilon Q(t) \\ &\leq \sum_{t \in \mathcal{T}_k} [Q(kT) + (t - kT)(1 - \epsilon)] X(t) - \sum_{t \in \mathcal{T}_k} \epsilon [Q(kT) - (t - kT)\epsilon] \\ &\leq \sum_{t \in \mathcal{T}_k} Q(kT) [X(t) - \epsilon] + \sum_{t \in \mathcal{T}_k} (t - kT) [(1 - \epsilon)X(t) + \epsilon^2]. \end{aligned} \quad (43)$$

Taking the conditional expectation on (43) under a given $Q(kT)$, we have

$$\begin{aligned} &\mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} Q(t) [X(t) - \epsilon] \middle| Q(kT) \right\} \\ &\leq \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} Q(kT) [X(t) - \epsilon] \middle| Q(kT) \right\} + \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} (t - kT) [(1 - \epsilon)a(t) + \epsilon^2] \right\} \\ &\leq \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} Q(kT) [X(t) - \epsilon] \middle| Q(kT) \right\} + \frac{T(T-1)[(1-\epsilon)\rho + \epsilon^2]}{2}. \end{aligned} \quad (44)$$

Using the result of (44) and letting $B_2 = B_1 + (T-1)[(1-\epsilon)\rho + \epsilon^2]/2$, we can further relax the inequality (10) into (11), which completes the proof.

C. Proof of (22)

Since the random variables $a(t)$ and $h_n(t)$ are i.i.d. over slots $t \in \mathcal{T}_k$, we have

$$\begin{aligned} Z_n^{sum}(k) &= \min_{\substack{0 \leq p(t) \leq P \\ \forall t \in \mathcal{T}_k}} \sum_{t \in \mathcal{T}_k} \mathbb{E} \left\{ V\mathcal{E}(t) + Q(kT)X(t) \middle| n(k) = n \right\} \\ &= \sum_{t \in \mathcal{T}_k^c} \mathbb{E} \{ V \cdot 0 + Q(kT) \cdot \mathbb{1}_{\{a(t)=1\}} \} + \sum_{t \in \mathcal{T}_k \setminus \mathcal{T}_k^c} \mathbb{E} \{ z_n(t) \cdot \mathbb{1}_{\{a(t)=1\}} \} \\ &= \sum_{t \in \mathcal{T}_k^c} \rho Q(kT) + \sum_{t \in \mathcal{T}_k \setminus \mathcal{T}_k^c} \rho Z_n(k), \end{aligned} \quad (45)$$

where the second term in the last equality is derived according to the mutual independence between channel gain $h_n(t)$ and task arrival $a(t)$. Based on the definition of set \mathcal{T}_k^c , $Z_n^{sum}(k)$ in (45) can be further expanded to (22) for different cases of n , which completes the proof.

D. Proof of Proposition 2

Since $h_n(t) \sim \exp(1/H_n(k))$, i.e., exponential distribution, with some manipulations, $Z_n(k)$ for the channel model (25) can be expressed as

$$Z_n(k) = \frac{Ve_n(k)}{H_n(k)} E_1\left(\frac{h_n^{\min}(k)}{H_n(k)}\right) + Q(kT) \left(1 - e^{-\frac{h_n^{\min}(k)}{H_n(k)}}\right), \quad (46)$$

where $E_1(x) \triangleq \int_x^\infty \frac{e^{-t}}{t} dt$, with $x > 0$, is the exponential integral function.

With $f_n(k) = f(k) > \frac{\xi}{\tau_d}$, $\forall n$, it follows that $\{e_n(k), h_n^{\min}(k)\} = \{e(k), h^{\min}(k)\}$, $\forall n$, and $Z_n(k)$ in (46) can be simplified as a function of $H_n(k)$:

$$Z_n(k) = \frac{Ve(k)}{H_n(k)} E_1\left(\frac{h^{\min}(k)}{H_n(k)}\right) + Q(kT) \left(1 - e^{-\frac{h^{\min}(k)}{H_n(k)}}\right) \triangleq U(H_n(k)), \quad \forall n \in \mathcal{N}. \quad (47)$$

1) *Proof of Property a)*: We need the following preliminary lemma to prove Property a):

Lemma 3: $U(H_n(k))$ is a monotonically decreasing function of $H_n(k) > 0$.

Proof: Using the fact that $\left[\int_{f(x)}^a g(t) dt\right]' = -g(f(x)) \cdot f'(x)$, we have

$$\begin{aligned} U'(H_n(k)) &= -\frac{Ve(k)}{(H_n(k))^2} E_1\left(\frac{h^{\min}(k)}{H_n(k)}\right) + \frac{Ve(k)}{H_n(k)} \left(-\frac{e^{-\frac{h^{\min}(k)}{H_n(k)}}}{\frac{h^{\min}(k)}{H_n(k)}}\right) \cdot -\frac{h^{\min}(k)}{(H_n(k))^2} \\ &\quad - Q(kT) e^{-\frac{h^{\min}(k)}{H_n(k)}} \cdot \frac{h^{\min}(k)}{(H_n(k))^2} \\ &= -\frac{Ve(k)}{(H_n(k))^2} \left[E_1\left(\frac{h^{\min}(k)}{H_n(k)}\right) - e^{-\frac{h^{\min}(k)}{H_n(k)}}\right] - \frac{Q(kT)h^{\min}(k)}{(H_n(k))^2} e^{-\frac{h^{\min}(k)}{H_n(k)}}. \end{aligned} \quad (48)$$

According to the definitions of $e_n(k)$ and $h_n^{\min}(k)$ in Theorem 1, it follows

$$\frac{Ve(k)}{h_n^{\min}(k)} = \min \left\{ Q(kT), V\bar{P} \left[\tau_d - \frac{\xi}{f_n(k)} \right]^+ \right\} \leq Q(kT). \quad (49)$$

Therefore, $Ve(k) \leq Q(kT)h^{\min}(k)$ holds. Plugging it into (48) yields

$$U'(H_n(k)) \leq -\frac{Q(kT)h^{\min}(k)}{(H_n(k))^2} E_1\left(\frac{h^{\min}(k)}{H_n(k)}\right) \leq 0, \quad (50)$$

which completes the proof of Lemma 3. ■

Since $Z_n(k) = U(H_n(k))$, $\forall n \in \mathcal{N}$, and $U(H_n(k))$ monotonically decreases with $H_n(k)$ by Lemma 3, we have $n' \triangleq \arg \max_{n \in \mathcal{N} \setminus n(k-1)} \{Z_n(k)\} = \arg \min_{n \in \mathcal{N} \setminus n(k-1)} \{H_n(k)\}$, completing the proof.

2) *Proof of Property b*): Recall that in the migration policy (24), the user migrates from its current associated BS $n(k-1)$ to BS n' if $(1-\alpha)Z_{n'}(k) + \alpha Q(kT) < Z_{n(k-1)}(k)$ is met. Plugging (47) into the condition and simplifying, we obtain

$$(1-\alpha) \left[\frac{Ve(k)}{H_{n'}(k)} \mathbb{E}_1 \left(\frac{h^{\min}(k)}{H_{n'}(k)} \right) - Q(kT) e^{-\frac{h^{\min}(k)}{H_{n'}(k)}} \right] < \frac{Ve(k)}{H_{n(k-1)}(k)} \mathbb{E}_1 \left(\frac{h^{\min}(k)}{H_{n(k-1)}(k)} \right) - Q(kT) e^{-\frac{h^{\min}(k)}{H_{n(k-1)}(k)}}. \quad (51)$$

Note that for $x > 0$, $\frac{1}{2}e^{-x} \ln(1 + \frac{x}{2}) < \mathbb{E}_1(x) < e^{-x} \ln(1 + \frac{1}{x})$. Utilizing this property, (51) can be re-written as the following sufficient condition:

$$\underbrace{e^{-\frac{h^{\min}(k)}{H_{n'}(k)} + \ln(1-\alpha)}}_{A_1} \underbrace{\left[\frac{Ve(k)}{H_{n'}(k)} \ln \left(1 + \frac{H_{n'}(k)}{h^{\min}(k)} \right) - Q(kT) \right]}_{B_1} < \underbrace{e^{-\frac{h^{\min}(k)}{H_{n(k-1)}(k)}}}_{A_2} \underbrace{\left[\frac{Ve(k)}{2H_{n(k-1)}(k)} \ln \left(1 + \frac{2H_{n(k-1)}(k)}{h^{\min}(k)} \right) - Q(kT) \right]}_{B_2}. \quad (52)$$

Here, $\frac{Ve(k)}{H_{n'}(k)} \ln \left(1 + \frac{H_{n'}(k)}{h^{\min}(k)} \right) < \frac{Ve(k)}{H_{n'}(k)} \frac{H_{n'}(k)}{h^{\min}(k)} = \frac{Ve(k)}{h^{\min}(k)} \leq Q(kT)$, where the last equality is according to (49). Similarly, $\frac{Ve(k)}{2H_{n(k-1)}(k)} \ln \left(1 + \frac{2H_{n(k-1)}(k)}{h^{\min}(k)} \right) < Q(kT)$. Thus, we have $B_1, B_2 < 0$. With $A_1, A_2 > 0$, (52) can be written as $A_1 \cdot (-B_1) > A_2 \cdot (-B_2)$, and decomposed into the following two conditions by letting $A_1 > A_2$ and $-B_1 > -B_2$, respectively:

$$-\frac{h^{\min}(k)}{H_{n'}(k)} + \ln(1-\alpha) > -\frac{h^{\min}(k)}{H_{n(k-1)}(k)}, \quad (53)$$

$$\frac{1}{H_{n'}(k)} \ln \left(1 + \frac{H_{n'}(k)}{h^{\min}(k)} \right) < \frac{1}{2H_{n(k-1)}(k)} \ln \left(1 + \frac{2H_{n(k-1)}(k)}{h^{\min}(k)} \right). \quad (54)$$

It can be checked that $\frac{1}{x} \ln(1+x)$ is monotonically decreasing with x ; thus (54) is equivalent to $H_{n'}(k) > 2H_{n(k-1)}(k)$. By letting $H_{\text{th}} = \frac{h^{\min}(k)}{\ln(\frac{1}{1-\alpha})}$, (53) is equivalent to $H_{n'}(k) > \frac{H_{\text{th}}(k)H_{n(k-1)}(k)}{H_{\text{th}}(k) - H_{n(k-1)}(k)}$, with $H_{n(k-1)}(k) < H_{\text{th}}(k)$. Summarizing above conditions yields the results of Property b).

E. Proof of Proposition 3

For notational simplicity, we use variable $\nu_n(k)$ to replace $\frac{h_n^{\min}(k)}{H_n(k)}$ for all $n \in \mathcal{N}$ in this proof.

Recall that $Z_n(k)$ for the channel model (25) is given by (46). By plugging $\nu_n(k) = \frac{h_n^{\min}(k)}{H_n(k)}$ into (46), we have

$$Z_n(k) = \frac{Ve_n(k)}{h_n^{\min}(k)} \nu_n(k) \mathbb{E}_1(\nu_n(k)) + Q(kT) (1 - e^{-\nu_n(k)}). \quad (55)$$

Assume that $\bar{P} > \frac{Q}{V[\tau_d - \frac{\xi}{f_n(k)}]^+}$, for all n . Then, according to (49), $\frac{V e_n(k)}{h_n^{\min}(k)} = Q(kT)$ and $Z_n(k)$ can be further expressed as

$$Z_n(k) = Q(kT) + Q(kT) [\nu_n(k) E_1(\nu_n(k)) - e^{-\nu_n(k)}] \triangleq Y(\nu_n(k)). \quad (56)$$

1) *Proof of Property a):* Similar to the proof of Proposition 2, $n' = \operatorname{argmin}_{n \in \mathcal{N} \setminus n(k-1)} \{\nu_n(k)\}$ is sufficient to verifying $Y(\nu_n(k))$ is a monotonic increasing function with $\nu_n(k)$. Since $E_1'(x) = -e^x/x$, we have

$$Y'(\nu_n(k)) = Q(kT) \left[E_1(\nu_n(k)) + \nu_n(k) \frac{-e^{\nu_n(k)}}{\nu_n(k)} + e^{\nu_n(k)} \right] = Q(kT) E_1(\nu_n(k)) \quad (57)$$

Since $Z_n(k) = Y(\nu_n(k))$, $\forall n \in \mathcal{N}$, and $Y(\nu_n(k))$ is monotonically increasing with $\nu_n(k)$, we have $n' = \operatorname{argmin}_{n \in \mathcal{N} \setminus n(k-1)} \{Z_n(k)\} = \operatorname{argmin}_{n \in \mathcal{N} \setminus n(k-1)} \{\nu_n(k)\}$, which completes the proof.

2) *Proof of Property b):* Letting $(1 - \alpha)Z_{n'}(k) + \alpha Q(kT) < Z_{n(k-1)}(k)$ and simplifying, we have the migration condition for this case:

$$(1 - \alpha) [\nu_{n'}(k) E_1(\nu_{n'}(k)) - e^{-\nu_{n'}(k)}] < \nu_{n(k-1)}(k) E_1(\nu_{n(k-1)}(k)) - e^{-\nu_{n(k-1)}(k)} \quad (58)$$

Similar to the proof in Proposition 2, using the facts that $\frac{1}{2}e^{-x} \ln(1 + \frac{2}{x}) < E_1(x) < e^{-x} \ln(1 + \frac{1}{x})$ and $x \ln(1 + \frac{1}{x}) < 1$ hold for $x > 0$, we have

$$\underbrace{e^{-\nu_{n'}(k) + \ln(1-\alpha)}}_{C_1} \underbrace{\left[1 - \nu_{n'}(k) \ln\left(1 + \frac{1}{\nu_{n'}(k)}\right) \right]}_{D_1} > \underbrace{e^{-\nu_{n(k-1)}(k)}}_{C_2} \underbrace{\left[1 - \frac{\nu_{n(k-1)}(k)}{2} \ln\left(1 + \frac{2}{\nu_{n(k-1)}(k)}\right) \right]}_{D_2} \quad (59)$$

where components C_1, C_2, D_1, D_2 are all positive. Letting $C_1 > C_2$ and $D_1 > D_2$, it follows

$$\nu_{n'}(k) + \ln\left(\frac{1}{1-\alpha}\right) < \nu_{n(k-1)}(k) \quad (60)$$

$$\nu_{n'}(k) \ln\left(1 + \frac{1}{\nu_{n'}(k)}\right) < \frac{\nu_{n(k-1)}(k)}{2} \ln\left(1 + \frac{2}{\nu_{n(k-1)}(k)}\right) \quad (61)$$

Note that $x \ln(1 + \frac{1}{x})$ is monotonically increasing with x . Thus, (61) is equivalent to $2\nu_{n'}(k) < \nu_{n(k-1)}$. Combining this with (60) gives the result of Property b).

F. Proof of Theorem 2

Using Lemma 2 and the fact that the proposed algorithm is developed through minimizing the R.H.S. of the inequality (11), we have

$$\begin{aligned} & \Delta_T(Q^*(kT)) + V \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} \mathcal{E}^*(t) \middle| Q^*(kT) \right\} \\ & \leq B_2 T + \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} V \mathcal{E}^*(t) + Q^*(kT) [X(t)^* - \epsilon] \middle| Q^*(kT) \right\} \end{aligned} \quad (62)$$

$$\begin{aligned}
&\stackrel{(d)}{\leq} B_2T + \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} V \widehat{\mathcal{E}}(t) + Q^*(kT) \left[\widehat{X}(t) - \epsilon \right] \middle| Q^*(kT) \right\} \\
&\stackrel{(e)}{\leq} B_2T + \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} V \widehat{\mathcal{E}}(t) \middle| Q^*(kT) \right\} - Q^*(kT) \delta T.
\end{aligned} \tag{63}$$

Here, $\widehat{\mathcal{E}}(t)$ and $\widehat{X}(t)$ denote the energy consumption and task failure achieved by the policy satisfying the conditions (30), respectively. Step (d) is because the right term of (62) obtained by the proposed algorithm is not more than that of any other feasible policy including the policy satisfying the conditions (30). Step (e) is derived by the conditions (30).

Rearranging the terms and noting that $|\mathcal{E}^*(t) - \widehat{\mathcal{E}}(t)| \leq \overline{P} \cdot \max_{n \in \mathcal{N}} \left\{ \tau_d - \frac{\xi}{f_n^{\max}} \right\} \triangleq E_{\max}$, we have

$$\Delta_T(Q^*(kT)) \leq B_2T + VTE_{\max} - Q^*(kT)\delta T. \tag{64}$$

Taking expectation of the above and summing it over $k = 0, 1, \dots, K-1$ yield

$$\frac{1}{2} \mathbb{E} \{ Q^*(kT)^2 \} - \frac{1}{2} \mathbb{E} \{ Q^*(0)^2 \} \leq K [B_2T + VTE_{\max}] - \delta T \sum_{k=0}^{K-1} \mathbb{E} \{ Q^*(kT) \}. \tag{65}$$

Rearranging terms in the above, dividing both sides of $K\delta T$, and taking limit as $K \rightarrow \infty$ yield

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \{ Q^*(kT) \} \leq \frac{B_2 + VE_{\max}}{\delta} + \frac{\mathbb{E} \{ Q^*(0)^2 \} - \mathbb{E} \{ Q^*(kT)^2 \}}{2K\delta T} \leq \frac{B_2 + VE_{\max}}{\delta} \tag{66}$$

This proves (31) in Theorem 2.

According to [31, Theorem 4.5], if the problem is feasible, there exists a stationary optimal ω -only policy, in which decisions $n(k)$ and $\{p(t)\}$ are made independent of the queue length, achieving the minimum energy consumption $E_{\text{av}}^{\text{opt}}$ while meeting the queue stability constraint. Therefore, we have

$$\Delta_T(Q^*(kT)) + V \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} \mathcal{E}^*(t) \middle| Q^*(kT) \right\} \leq B_2T + V \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} \mathcal{E}^{\text{opt}}(t) \right\} \tag{67}$$

where the term $Q^*(kT) \mathbb{E} \{ \sum_{t \in \mathcal{T}_k} [X^{\text{opt}}(t) - \epsilon] \}$ is neglected in the R.H.S. of (67) since it is non-positive due to satisfying the queue stability constraint.

Taking expectations of the above inequality and summing it over $k = 0, 1, \dots, K-1$ yield

$$\begin{aligned}
&\frac{1}{2} \mathbb{E} \{ Q^*(kT)^2 \} - \frac{1}{2} \mathbb{E} \{ Q^*(0)^2 \} + V \sum_{k=0}^{K-1} \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} \mathcal{E}^*(t) \right\} \\
&\leq KB_2T + V \sum_{k=0}^{K-1} \mathbb{E} \left\{ \sum_{t \in \mathcal{T}_k} \mathcal{E}^{\text{opt}}(t) \right\}.
\end{aligned} \tag{68}$$

Dividing both sides by VKT , taking the limit as $K \rightarrow \infty$, and noting that $\{h_n(t), a(t)\}$ are i.i.d. over slots within a frame, we have

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{t \in \mathcal{T}_k} \mathbb{E} \{ \mathcal{E}^*(t) \} \leq \frac{B_2}{V} + \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{t \in \mathcal{T}_k} \mathbb{E} \{ \mathcal{E}^{opt}(t) \} = \frac{B_2}{V} + E_{av}^{opt}. \quad (69)$$

This yields (32) in Theorem 2.

REFERENCES

- [1] European Telecommunications Standards Institute (ETSI), “Mobile-edge computing - Introductory technical white paper,” Sept. 2014.
- [2] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, “On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 2017.
- [3] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, Fourthquarter 2017.
- [4] Z. Liang, Y. Liu, T. Lok, and K. Huang, “Multiuser computation offloading and downloading for edge computing with virtualization,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4298–4311, Sept. 2019.
- [5] M. Liu and Y. Liu, “Price-based distributed offloading for mobile-edge computing with computation capacity constraints,” *IEEE Commun. Lett.*, vol. 7, no. 3, pp. 420–423, Jun. 2018.
- [6] European Telecommunications Standards Institute (ETSI), “Mobile edge computing (MEC); End to end mobility aspects,” ETSI GR MEC 018 V1.1.1, Oct. 2017.
- [7] M. Li, J. Gao, L. Zhao, and X. Shen, “Deep reinforcement learning for collaborative edge computing in vehicular networks,” *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 4, pp. 1122–1135, Dec. 2020.
- [8] S. Wang, J. Xu, N. Zhang, and Y. Liu, “A survey on service migration in mobile edge computing,” *IEEE Access*, vol. 6, pp. 23 511–23 528, 2018.
- [9] Z. Rejiba, X. Masip-Bruin, and E. Marín-Tordera, “A survey on mobility-induced service migration in the fog, edge, and related computing paradigms,” *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–33, Sept. 2019.
- [10] Z. Liang, Y. Liu, T.-M. Lok, and K. Huang, “Multi-cell mobile edge computing: Joint service migration and resource allocation,” *IEEE Trans. Wireless Commun.*, Early Access, Apr. 2021.
- [11] H. Ma, Z. Zhou, and X. Chen, “Leveraging the power of prediction: Predictive service placement for latency-sensitive mobile edge computing,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6454–6468, Oct. 2020.
- [12] S. Ge, M. Cheng, and X. Zhou, “Interference aware service migration in vehicular fog computing,” *IEEE Access*, vol. 8, pp. 84 272–84 281, 2020.
- [13] A. Ksentini, T. Taleb, and M. Chen, “A markov decision process-based service migration procedure for follow me cloud,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 10-14, 2014, pp. 1350–1354.
- [14] S. Wang, R. Uргаonkar, M. Zafer, T. He, K. Chan, and K. K. Leung, “Dynamic service migration in mobile edge computing based on markov decision process,” *IEEE/ACM Trans. Netw.*, vol. 27, no. 3, pp. 1272–1288, Jun. 2019.
- [15] Q. Cui, J. Zhang, X. Zhang, K. Chen, X. Tao, and P. Zhang, “Online anticipatory proactive network association in mobile edge computing for IoT,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4519–4534, Jul. 2020.
- [16] C. Liu, F. Tang, Y. Hu, K. Li, Z. Tang, and K. Li, “Distributed task migration optimization in MEC by extending multi-agent deep reinforcement learning approach,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 7, pp. 1603–1614, Jul. 2021.

- [17] Y. Sun, S. Zhou, and J. Xu, "EMM: Energy-aware mobility management for mobile edge computing in ultra dense networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2637–2646, Nov. 2017.
- [18] T. Ouyang, R. Li, X. Chen, Z. Zhou, and X. Tang, "Adaptive user-managed service placement for mobile edge computing: An online learning approach," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Paris, France, Apr. 29-May 2, 2019, pp. 1468–1476.
- [19] J. Wang, J. Hu, and G. Min, "Online service migration in edge computing with incomplete information: A deep recurrent actor-critic method." [Online]. Available: <https://arxiv.org/pdf/2012.08679.pdf>
- [20] T. Ouyang, Z. Zhou, and X. Chen, "Follow me at the edge: Mobility-aware dynamic service placement for mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2333–2345, Oct. 2018.
- [21] C. Liu, M. Bennis, M. Debbah, and H. V. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4132–4150, Jun. 2019.
- [22] P. Schulz *et al.*, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.
- [23] A. Sang, X. Wang, M. Madhian, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," *Wireless Netw.*, vol. 14, pp. 103–120, Jan. 2008.
- [24] M. Erel-Özçevik and B. Canberk, "Road to 5G reduced-latency: A software defined handover model for eMBB services," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8133–8144, Aug. 2019.
- [25] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.
- [26] Z. Chang, Z. Zhou, T. Ristaniemi, and Z. Niu, "Energy efficient optimization for computation offloading in fog computing system," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 4-8, 2017, pp. 1–6.
- [27] L. Chen, J. Xu, S. Ren, and P. Zhou, "Spatio-temporal edge service placement: A bandit learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8388–8401, Dec. 2018.
- [28] X. Yang, Z. Chen, K. Li, Y. Sun, N. Liu, W. Xie, and Y. Zhao, "Communication-constrained mobile edge computing systems for wireless virtual reality: Scheduling and tradeoff," *IEEE Access*, vol. 6, pp. 16 665–16 677, 2018.
- [29] C. She, Y. Duan, G. Zhao, T. Q. S. Quek, Y. Li, and B. Vucetic, "Cross-layer design for mission-critical IoT in mobile edge computing systems," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9360–9374, Dec. 2019.
- [30] K. Guo and T. Q. S. Quek, "Dynamic computation offloading in multi-server MEC systems: An online learning approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 7-11, 2020, pp. 1–6.
- [31] M. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool, 2010.
- [32] Y. Yao, L. Huang, A. B. Sharma, L. Golubchik, and M. J. Neely, "Power cost reduction in distributed data centers: A two-time-scale approach for delay tolerant workloads," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 200–211, Jan. 2014.
- [33] H. Yu, M. H. Cheung, L. Huang, and J. Huang, "Power-delay tradeoff with predictive scheduling in integrated cellular and wi-fi networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 735–742, Apr. 2016.
- [34] D. Bruneo, "A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 3, pp. 560–569, Mar. 2014.
- [35] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 2, no. 3, pp. 257–269, Jul.-Sept. 2003.