Towards Mobility-Aware Dynamic Service Migration in Mobile Edge Computing

Fangzheng Liu, Bofeng Lv, Jiwei Huang(✉), and Sikandar Ali

Beijing Key Laboratory Petroleum Data Mining,
China University of Petroleum-Beijing, Beijing 102249, China
2019310704@student.cup.edu.cn, lvbofeng@foxmail.com,
{huangjw,sikandar}@cup.edu.cn

Abstract. Mobile edge computing is beneficial to reduce service response time by pushing cloud functionalities to the network edge. However, it is necessary to consider whether to conduct service migration to ensure the quality of service as users migrate to new locations. It is challenging to make migration decisions optimally due to the mobility of the users. To address this issue, we propose a mobility-aware dynamic service migration scheme for mobile edge computing. In order to predict a mobile user's movement behavior in terms of boundary crossing probability, we use a new approach for modeling user mobility and formulate the service migration problem as a Markov Decision Process (MDP). This policy can effectively weigh the relationship between delay and migration costs. Our methods capture general cost models and provide a mathematical framework to design optimal service migration policies. Experimental evaluations based on real-world mobility traces of Beijing taxis show superior performance of the proposed solution.

Keywords: Mobile edge computing · Service migration · Markov Decision Process (MDP) · User mobility

1 Introduction

With the prevalence of mobile terminals and the Internet of Things (IoT), Mobile Edge Computing (MEC) has emerged as a novel architecture where cloud computing services are extended to the edge of networks leveraging mobile base stations [1–3]. It integrates the techniques of cloud computing and mobile computing and pushes part of the applications, data and services away from centralized cloud data centers to the logical extremes of a network where edge servers are deployed. Since the local edge servers are located closer to the users and IoT devices than the centralized cloud data centers, the quality of service (QoS, e.g. response time and throughput) and privacy can be improved, and the overhead can be reduced as well [4]. Therefore, MEC has become increasingly popular for supporting a variety of innovative applications and services in mobile environments.

H. Gao et al. (Eds.): CollaborateCom 2020, LNICST 349, pp. 115–131, 2021.
https://doi.org/10.1007/978-3-030-67537-0_8

In most of MEC scenarios, the locations of users and devices are time-varied and dispersed in a wide area. Devices and users on the edge site can only access to the services within the signal coverage of edge base stations (or MEC servers). When they move out, they can choose to continue to let the service process at the original edge node, and ensure the continuity of service through the data transmission between the edge nodes; however, too long network distance may increase the delay of data transmission between the user and the edge server that hosts the service, and affect the service quality perceived by users [5]. To address the issue brought by the user mobility, dynamic service migration techniques have been put forward for improving user experience under MEC [6].

The basic idea of dynamic service migration in edge computing is to migrate the services from one edge server to another edge server according to the movements of the users being invoking the services. Figure 1 illustrates a typical scenario. In service migration, we have to solve the following two problems. The first one is whether or not to migrate a service at a certain time point, and if yes the second one is where to migrate the service. Migrating a service may cause service interruption and bring in network overhead, whereas not migrating a service may increase the data transmission delay between the user and the edge server that hosts the service when the user moves away from its original location. It is quite challenging to make an optimal migration decision due to the uncertainty of user mobility as well as the complex trade-off between the costs related to migration and distant data transmission.

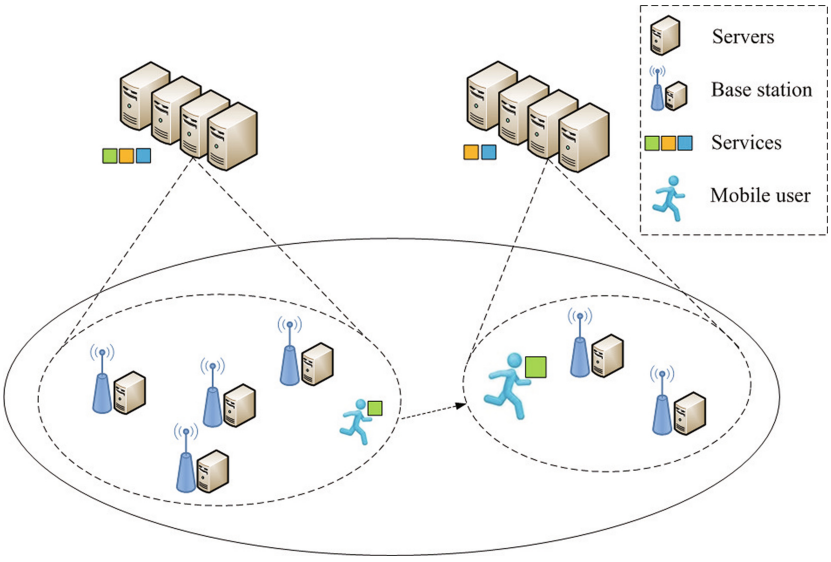


Fig. 1. Service migration in mobile edge computing

For a smooth service migration in MEC, user mobility as one of the most important factors that should be taken into account. There have been existing works on mobility-driven service migration dedicating to model or predict user mobility patterns. The performance of MEC in the presence of user mobility is first studied in [7] using a Markovian mobility model, but decisions on whether and where to migrate the service are not considered. A preliminary work on mobility-driven service migration based on Markov Decision Processes (MDPs) is given in [8,9], which mainly considers one-dimensional (1-D) mobility patterns and takes the uniform random walk migration model as modeling hypothesis. But in fact individual users do not necessarily follow a uniform random walk. To the best of our knowledge, real user mobility has not been considered in the literature, which is a much more realistic scenario compared to the uniform random walk model and we consider in this article.

To address this problem, this paper presents a mobility-aware dynamic service migration scheme for mobile edge computing. Based on the analysis of the trajectory data from users, we propose a geometry-based user mobility model for predicting the probability of a user to move out from the coverage of the current edge server to another one. Considering the trade-off between QoS and cost, the service migration is formulated by a Markov decision problem, and a mobility-aware dynamic migration (MODEM) algorithm is designed. Finally, with trajectory data set from real-life applications, extensive simulation experiments are conducted to validate the effectiveness of our MODEM algorithm.

The remainder of the paper is organized as follows. Section 2 reviews the representative research efforts relevant to our work. Section 3 presents user mobility model to predict the next moving area of the user. Section 4 introduces general cost models and provides a mathematical framework to design optimal service migration policies. Section 5 reports detailed experimental results. Finally, Sect. 6 concludes this work and discusses future research directions.

2 Related Work

Mobile edge computing is an extension of cloud computing, with the benefits of reduced delay. Due to the edge nodes coverage is small, the user's mobility will have a great impact on service quality. Choosing a reasonable service migration strategy based on the predicted results of user mobility is crucial to ensuring service quality. There has been extensive work devoted to user mobility prediction and service migration.

Research directions related to service migration mainly focuses on the load balancing of distributed data center. Ouyang *et al.* [10] has proposed an Lyapunov optimization technique to incorporate the long-term budget into a series of real-time optimization problems, which achieve a desirable balance between time-averaged user-perceived latency and migration cost. Similarly, Chowdhury *et al.* [11] has proposed to use the load information of each node in a period of time in the data center to predict the load and change trend at the next time point. On this basis, the allocation of services is determined to avoid unnecessary

overhead caused by frequent service migrations [12]. However, these methods are mainly based on network load, user requests and other information to make decisions, without further consideration of user mobility [13]. As the user moves, the distance between the user and the MEC server where the service is located also changes relatively, and the original connection scheme may no longer be optimal.

In order to make accurate location prediction, the work [14] and [15] extracted features of multiple dimensions from users' historical information, such as network status features and social frequency features at different times, and effectively integrated them into a unified framework by using the factor graph (FG) model. All the above work on predicting mobility in service migration has a common assumption that we have perfect information about users over a period of time. However, in the actual environment, it is difficult to accurately predict the above users and network information. At the same time, for each decision moment, due to the lack of understanding of network environment parameters, users will consume additional communication costs in collecting system information. There are many research areas related to user mobility, such as context switching in cellular networks [16] and wireless ATM networks [17]. Nevertheless, these studies cannot be directly applied to service migration scenarios due to different decision spaces.

Many studies, e.g., [18–21], migrate the service to the vicinity of the current location of the user by means of the virtual machine dynamic migration technology, so as to ensure a low delay when the user use the service. However, this will entail significant service migration costs (such as additional network bandwidth usage and power resource consumption). To solve this problem, the change of network connection state is modeled by introducing user mobility have been investigated in [22–24]. Nevertheless, most of these schemes adopt the random walk model, rarely exploring the user's trajectory data and predicting the user's movement. In addition, these work pay less attention to the influence of QoS (such as network delay and migration cost) on edge server selection in service migration, so it is difficult to choose the optimal service migration strategy. In contrast, the method proposed in this paper can predict the location of the user at the next time, and considering the limitations of long-term prediction, it can make more intelligent decisions at the current time to avoid the cost of frequent service migration.

Different from the existing work, we proposed a mobile-aware dynamic service migration scheme, aiming at making more intelligent service migration decisions through location prediction, so as to reduce the service delay perceived by users and improve their service quality. We solved this problem by establishing a geometry-based user mobility prediction model and describing the service migration problem as a markov decision process. Finally, we developed a motion-aware dynamic migration algorithm.

3 User Mobility Model

In order to obtain the optimal solution of mobility-aware service migration, the foundation is to capture the dynamics of user mobility and try to precisely

Table 1. List of notations.

Notation	Description
$(x(t), y(t))$	The user's position coordinates
v_t	The speed of user at timeslot t
v_{max}	The maximum moving speed of user
D	The distance between the cellular network and TA boundary
$d(t)$	The distance between user and service at timeslot t
$P(cell_i/X_t)$	The probability that the user moves to cell i given the state X_t
$(v_{x(t)}, v_{y(t)})$	The velocity vector in $x(t)$ and $y(t)$ direction
Δv	The change of velocity between time $t+1$ and t
$u(t)$	The position of user at timeslot t
$w(t)$	The position of service at timeslot t
$cost_{com}(t)$	Communication cost at timeslot t
$cost_{mig}(t)$	Migration cost at timeslot t
$A(s)$	action space
$Size$	The size of the service to be migrated
$s(t)$	Initial state at timeslot t
$s^{'}(t)$	Intermediate state at timeslot t
S	The state space, including the location of the mobile user and the location of the base station where the service is located
π	Decision policy
$C_a(s_0)$	The sum of costs when taking action a in state s_0
$V^*(s_0)$	Discount sum cost when starting at state s_0
$P[a(s_0), s_1]$	Transition probability from state s_0 to the next initial state s_1
$\varphi_c, \varphi_l, \beta, \delta_c, \delta_l, \mu$	Parameters related to the service migration model
Dis	Maximum distance between user and service to maintain connection

predict the movement of users among the base stations in the near future. In this section, we propose user mobility model for prediction. To facilitate presenting the model in a formal way, we summarize all the notations used in the following discussions of this paper in Table 1.

In mobile edge computing, a user moves under the coverage of the base stations. An edge server is deployed to process the requests submitted by users

from one or multiple base stations. Without loss of generality, we assume that each base station is equipped with an edge server. For the cases when multiple base stations share an edge server, we focus on the coverage area of the edge server and regard the base stations as one cell.

To solve the migration problem effectively and efficiently, we consider a time-slotted model for formulating the user mobility, and optimal policies are obtained at the beginning of each time slot according to the *state* of the user. The length of the time-gap interval is denoted by t, and the state of a user at time t is expressed as Eq. (1) where $(x(t), y(t))$ are the coordinates of the user and $v(t)$ is the velocity.

$$X_t = \{x(t), y(t), v(t)\} \tag{1}$$

Service migration is more likely to occur when users move near or across the cell boundaries. However, how to define the boundaries depends on the velocity of the user mobility. Therefore, we dynamically define a *Target Area (TA)* according to the upper bound of the velocity which is denoted by v_{max}, and the width of a TA is given by Eq. (2).

$$D = v_{max} \cdot t \tag{2}$$

At each decision epoch, only the users in the TA are possible to move to another cell, which may trigger a service migration. All the users in the central area, otherwise, will remain in the cell during the time interval. Consequently, we need to calculate the possibility of a user to move to another cell only when the user is located in the TA, which can significantly reduce the computational overhead of calculating the optimal solution during the service processes.

Figure 2 shows the coverage areas of TAs in a cellular network. Conventionally, we use a hexagon to represent the coverage area of a cell, and multiple cells constitute a cellular network. Dashed lines illustrate the boundaries of the TA, and the coverage area of a TAs is the hexagon ring between the TA boundary and cell boundary.

Specifically, we illuminate our mobility model when a user is moving in the TA of a cell as Fig. 3. The center coordinates of the cell are denoted by (a_0, b_0), and the distance between the user and the cell center is expressed by $d(t) = \sqrt{(x(t) - a_0)^2 + (y(t) - b_0)^2}$. We define a stochastic variable θ to represent the moving direction of the user during the time slot at time t, and let θ_i denote the direction to the i-th vertex of the cell for $i = 0, 1, \ldots, 5$.

At time t, the probability of the user moving to cell i given the current state X_t has a general form expressed as Eq. (3), where $f(\cdot)$ is the probability density function of the moving direction.

$$P\left(cell_i | X_t\right) = \int_{\theta_i}^{\theta_{(i+1)\%6}} f\left(\theta | X_t\right) d\theta \tag{3}$$

Additionally, we define the stochastic variate of the velocity as $v_t = (v_{x(t)}, v_{y(t)})^T$ in x and y direction respectively, and thus θ can be simply calculated by the following equation.

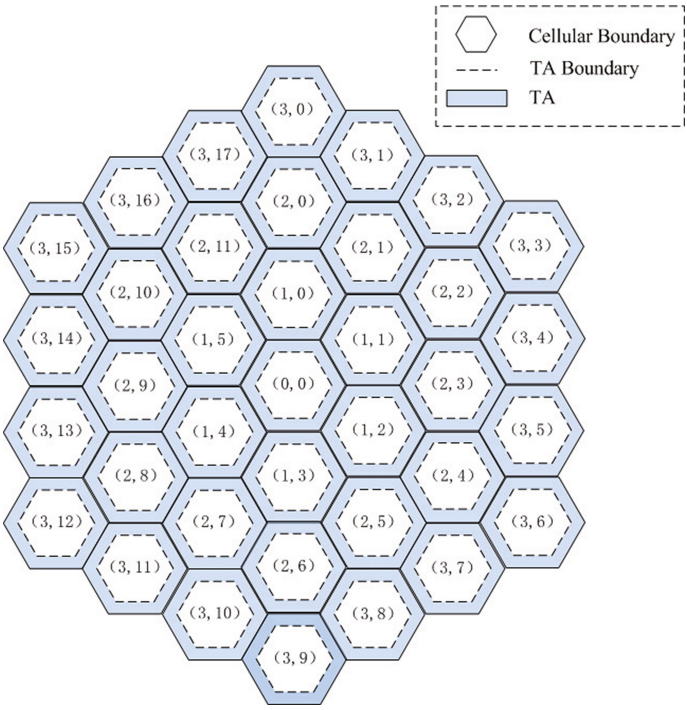


Fig. 2. Target areas in a cellular network.

$$\theta = g\left(v_t\right) = \arctan \frac{v_{y(t)}}{v_{x(t)}} \tag{4}$$

In most of the general cases, the probability density function of the velocity and its moving direction can be approximated by a Gaussian distribution [17]. Formally, we have

$$f(\theta|X_t) \sim N(\mu_\theta, \sigma_\theta^2) \tag{5}$$

where μ_θ is the mean and σ_θ is the standard deviation.

Afterward, the variate θ can be approximately expressed by Eq. (6).

$$\theta \approx g\left(v_t\right) + G\left(\Delta v\right) \tag{6}$$

where

$$G = \left.\frac{\partial g}{\partial v}\right|_{v=v_t} = \left[\frac{-v_{y(t)}}{v_{x(t)}^2 + v_{y(t)}^2}, \frac{v_{x(t)}}{v_{x(t)}^2 + v_{y(t)}^2}\right] \tag{7}$$

and

$$\Delta v = v_{t+1} - v_t \tag{8}$$

Δv is the change of velocity between time $t+1$ and t. Since v_t can be assumed to confrom to Gaussian distribution, and thus Δv also has a Gaussian

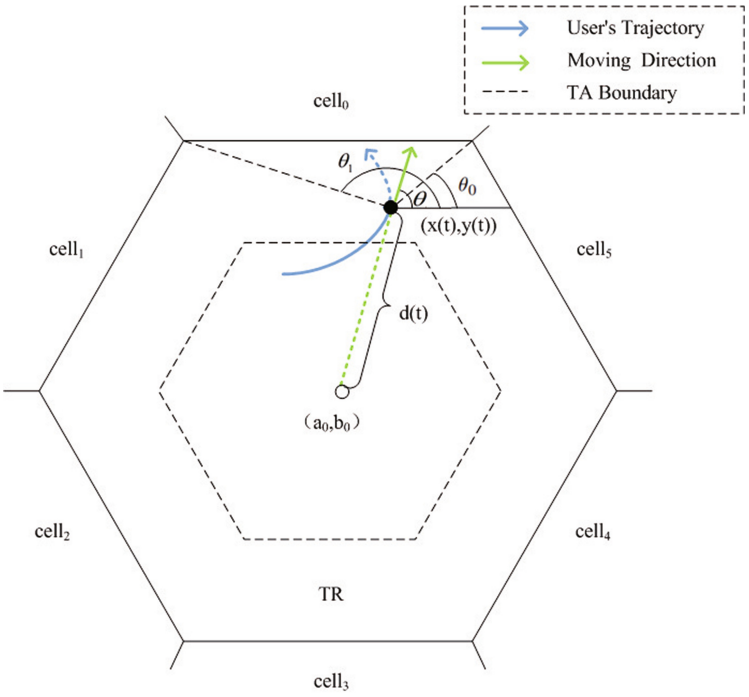


Fig. 3. User mobility model based on geometry.

distribution with the mean of $\mu_{\Delta v}$ and the variance of $\sigma^2_{\Delta v}$. With Eq. (6), we have

$$\mu_\theta = g(v_t) + \mu_{\Delta v} \tag{9}$$

$$\sigma_\theta = \sigma_{\Delta v} \tag{10}$$

In summary, one can conclude that the probability of a user moving to the i-th cell at the time slot $t+1$ can be calculated using the following expression.

$$P\left(cell_i | X_t\right) = \Phi\left(\frac{\theta_{i+1} - \mu_\theta}{\sigma_\theta}\right) - \Phi\left(\frac{\theta_i - \mu_\theta}{\sigma_\theta}\right) \tag{11}$$

In addition, in the cases when the speed of the mobile user is very slow, i.e., $\mu_{v_t} \approx 0$, $f(\theta|X_t)$ becomes a simple uniform distribution over $[0, 2\pi)$. Thus, the probability $P\left(cell_i|X_t\right)$ can be calculated by Eq. (12).

$$P\left(cell_i | X_t\right) = \frac{\theta_{i+1} - \theta_i}{2\pi} \tag{12}$$

4 Model and Algorithm of Service Migration

The user mobility model presented in the previous section can formulate the dynamic movements of users among different coverage areas of edge servers. With such model, in this section, we analyze the cost brought by migrating a service, and the formulate a dynamic optimization problem of service migration between two edge servers. An algorithm for solving the optimization problem is presented.

4.1 Cost Model

When predicting that a user will move to the coverage of another edge server with a high probability, a reasonable decision should be made whether to migrate the service being used by the user to the other edge server. The objective is to balance the trade-off between the QoS degradation and migration cost. For both of them, the distance should be fully taken into account for analysis, while other factors such as the size of the service should also be considered.

Let $u(t)$ and $w(t)$ denote the locations of the user and the service at time t, respectively. We assume that $w^{'}(t)$ is the location of the service that we try to migrate, according to the prediction result obtained form the mobility model. First, we analyze the communication cost which is closely related to the distance between the user and the service. The communication cost model has been studied by several existing literature, and hence we apply an exponential model [4], [18]. Specifically, we calculate the communication cost between the user at location $u(t)$ and the service located at $w^{'}(t)$ using the following expression.

$$cost_{com}(t) = \begin{cases} \varphi_c + \varphi_l \beta^{\|u(t)-w^{'}(t)\|} & u(t) \neq w^{'}(t) \\ 0 & u(t) = w^{'}(t) \end{cases} \tag{13}$$

where φ_c, φ_l and β are commonly non-negative parameters defined by the service provider, and β should be greater than 1.

The decision on triggering a service migration may bring in additional migration cost, and this type of cost depends on the distance between the locations of before and after the service migration, denoted by $w(t)$ and $w^{'}(t)$ respectively. Moreover, the size of the service $Size$ also has a certain impact on the cost of service transfer, since transferring a big service may consume considerable networking resources. Similar to the communication cost model, we define the migration cost using Eq. (14), where $\delta_c \geq 0$, $\delta_l \geq 0$, and $\mu > 1$.

$$cost_{mig}(t) = \begin{cases} \delta_c + \delta_l \mu^{Size \cdot \|w(t)-w^{'}(t)\|} & w(t) \neq w^{'}(t) \\ 0 & w(t) = w^{'}(t) \end{cases} \tag{14}$$

We define that the system state at the beginning of the time slot t is denoted by $s(t) = (u(t), w(t))$. Considering both the communication cost and migration cost, the total cost of our model is given by Eq. (15).

$$C(s(t)) = cost_{com}(t) + cost_{mig}(t) \tag{15}$$

4.2 Service Migration

Dynamic service migration is to make optimal decision at each time slot according to the system state. The user mobility model is applied to predict the location of a moving user in the near future, while the cost model is used to calculate the migration cost for supporting the optimal solution.

We let π denote a sequence of decision, which is a mapping between a state $s(t)$ and an action $a \in A(s)$. Let $a_\pi(s(t))$ denote the action taken by strategy π when the system is in state $s(t)$. This control action will trigger a state transition of the system from the current state $s(t)$ to an intermediate state $s^{'}(t) = ((u(t), w^{'}(t)) = a_\pi(s(t))$. Let $C_{a_\pi}(s(t))$ represent the sum of migration and communication costs incurred by taking the control action $a_\pi(s(t))$ in the time slot t, and we have $C_{a_\pi}(s(t)) = cost_{com}(t) + cost_{mig}(t)$. Starting from an initial state $s(0) = s_0$, the long term expected cost given a policy π is expressed as Eq. (16)

$$V_\pi(s_0) = \lim_{t\to\infty} E\left\{\sum_{\tau=0}^{t} \gamma^\tau C_{a_\pi}(s(\tau)) \,|\, s(0) = s_0\right\} \tag{16}$$

where $0 < \gamma < 1$ is the discount factor used to distinguish short-term costs from long-term costs. The long-term cost is multiplied by the discount factor, which means that in this model, the current short-term cost is more important than the uncertain long-term cost.

In this paper, the ultimate objective of service migration is to minimize the total cost given an initial state s_0, i.e.,

$$V^*(s_0) = \min_\pi V_\pi(s_0) \tag{17}$$

Equation (17) can be precisely formulated by a Markov Decision Process (MDP) with an infinite horizon discounted cost. The Bellman's equation of the MDP is shown by Eq. (18).

$$V^*(s_0) = \min_a \left\{C_a(s_0) + \gamma \sum_{s_1 \in S} P[a(s_0), s_1] \cdot V^*(s_1)\right\} \tag{18}$$

where $P[a(s_0), s_1]$ denotes the probability of the system to transfer from state $s^{'}(0) = s_0^{'} = a(s_0)$ to $s(1) = s_1$. The transition probability here is calculated by the mobility model presented in Sect. 3. With the relationship of state transitions, we have $s(t+1) = (u(t+1), w^{'}(t)) = (u(t+1), w(t+1))$, where $w(t+1) = w^{'}(t)$. In the following statement, the time symbol t will be omitted if not specified.

The solutions of the optimality equations include the minimum expected discounted total cost $V^*(s)$ and the optimal policy π. The optimal policy π indicates the migration target of the service given the state s. The procedures

of our mobility-aware dynamic migration (MODEM) algorithm in mobile edge computing are summarized as follows:

- The one-step cost at the k-th step is given by $C_a(s)$ when the system is in the state s and a control action a is selected. In this work, it is determined by the communication cost $cost_{com}(t)$ and the service migration cost $cost_{mig}(t)$.
- The state transfer mechanism is probabilistic which is controlled by the transition probabilities $P[a(s), s^{'}]$ of all states s, and the control action a is selected from all the feasible actions of state s. In this work, the transition probabilities are calculated by the user mobility model.
- The minimum cost function $V^*(s)$ includes the cost of one step and the minimum expected cost of all possible state transitions at the $(k+1)$-th step.

With all the calculations presented above, one can apply some well-known existing algorithms to solve the MDP problem. Considering the computational complexity of the algorithms, we select policy iteration algorithm, which is often able to find the optimal solution in the minimal number of iterations. The procedures of the policy iteration algorithm is shown in Algorithm 1. The output is the service migration strategy π in a certain state (that is, in the case of service migration, which server should be migrated from the current server to the surrounding server), which can minimize the total migration cost $V^*(s)$.

Algorithm 1. Policy-iteration algorithm based on mobility-aware

Input: $C_a(s)$, $P[a(s), s^{'}]$;
Output: $\pi*$, $V^*(s)$;

1: $\pi \in A(s)$ arbitrarily for all $s \in S$;
2: **for** $k = 0, 1, ...$ **do**
3: Compute the probability of transition $P\left[a\left(s\right), s^{'}\right]$ by solving Eq. (11);
4: Compute cost $C_a(s)$ by solving Eq. (15);
5: Solve $V^*(s) = C_a(s) + \gamma \sum_{s^{'} \in S} P[a(s), s^{'}] \cdot V^*(s^{'})$,
find $V^*(s)$ for π;
6: Update the policy according to $V^*(s)$,
$\pi_{k+1} \leftarrow \arg\min_{a} \left\{ C_a\left(s\right) + \gamma \sum_{s^{'} \in S} P\left[a\left(s\right), s^{'}\right] \cdot V^*\left(s^{'}\right) \right\}$;
7: **if** $\pi_{k+1} = \pi_k$ **then**
8: $\pi* \leftarrow \pi_k$;
9: **return** $\pi*$, $V^*(s)$;
10: **end if**
11: **end for**

It is worth noting that in practice, the maximum allowable distance between the mobile user and the service to maintain communication is usually bounded,

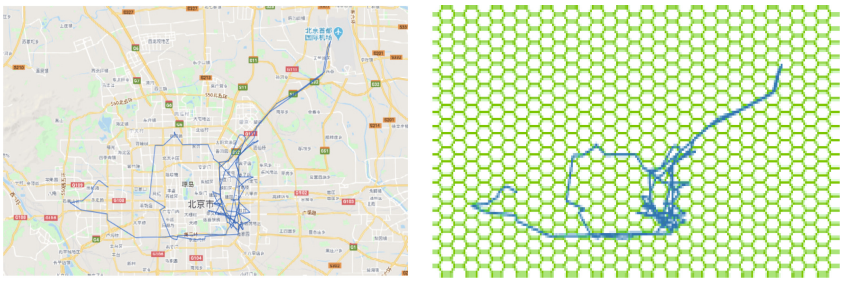
that is, although users can move in unbounded space, the state space controlled by the service is limited. Therefore, we assume that the maximum connection distance between the user and the service is Dis, then our policy only needs to focus on the state of $(u(t)-w(t)) \in [0, Dis]$. If the user moves to another service location where the distance exceeds the threshold Dis, the service migration is automatically triggered. Instead, we need to consider the mobility of users and the cost of service migration to make a reasonable migration decision.

5 Simulation Results

To validate the our approach, we conduct simulation experiments in MATLAB. A data set containing the GPS trajectories of 10,357 taxis in the city of Beijing is applied [25]. It has been released by Microsoft Research Asia, and covers the dates from Feb. 2 to Feb. 8 in the year of 2008. In our experiments, each active taxi is regarded as a mobile user in the MEC system, and its location varying with time is obtained from the longitude, latitude and time data from the data set.

The base stations are placed randomly, covering all the active area of the taxis. We simply assume that each taxi connects to its nearest base station measured by Euclidean distance. The base station connected to the taxi collects the taxi's service request, location and other parameter information. The length of a time slot is 60 s, and the base station calculates parameters such as the moving speed of the taxi based on the information obtained.

The relationship between the GPS track of the taxi and the cellular network is shown in Fig. 4. Subfigure (a) is the display of the GPS track of the taxi in the map, and subfigure (b) is the display of the GPS track of the taxi in the cellular network. Among them, the center of the hexagon is the base station, and the size of each hexagon is the coverage area of the base station. According to the survey, it is found that the 5G base station in densely populated areas is kept at about 200 m, and the suburban area is kept at about 500 to 1,000 m, so the radius of the hexagon in the experiment is set at about 400 m.



(a) (b)

Fig. 4. The relationship between the GPS trajectories of a taxi and cellular network

In order to validate the effectiveness of our MODEM approach, we select another three schemes for comparison as follows.

- **Non-Migration (NM)**: The service will not be migrated no matter how the user moves, unless the distance between them exceeds the maximum threshold Dis. Formally, in the area of $d(t) < Dis$, $a = 0$ always holds.
- **Always Migration (AM)**: We always migrate the service as soon as the user moves to another cell. Thus, the action variable a is always set to 1.
- **Random Walk Model (RWM)**: We still use the MDP algorithm to find the optimal solution, but replace our mobility model with a random walk model which has been widely applied in service migration [23]. We assume that the probability of the user leaving a cell is p, and thus the probability of staying inside the cell is $1 - 6p$.

5.1 Migration Cost Parameter Analysis

We analyze the algorithm by changing the migration cost parameters and the results are shown in Figs. 5 and 6.

As shown in Fig. 5, the total discounted cost of the NM approximates the cost of our MODEM algorithm when φ_l is small. The result can be explained by Eq. (13). the communication cost is relatively small when φ_l is small, and the policy is more in favor of NM. On the contrary, when φ_l is large, the total discounted cost of our MODEM algorithm approximates to the AM scheme. That is because, the communication cost is relatively large when φ_l is large, and the policy is more in favor of AM.

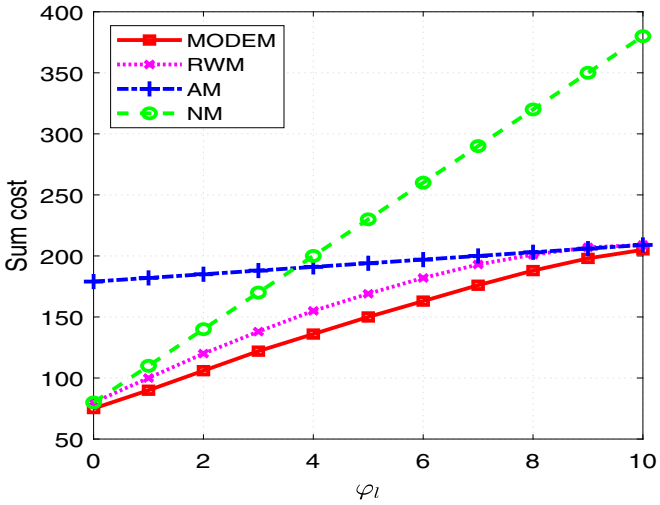


Fig. 5. Parameter analysis: the communication function parameter φ_l.

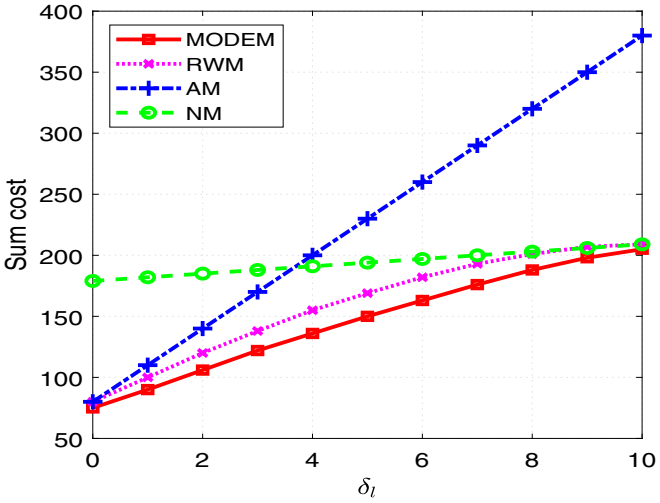


Fig. 6. Parameter analysis: the migration cost parameter δ_l.

In contrast, as shown in Fig. 6, the total discounted cost of the MODEM approximates the AM when δ_l is small. The result can be explained by Eq. (14), the migration cost is relatively small when δ_l is small, and the policy is more in favor of AM. On the contrary, when δ_l is large, the total discounted cost of our MODEM algorithm approximates to the NM scheme. That is because, the migration cost is relatively large when δ_l is large, and the policy is more in favor of NM. Moreover, since our MODEM algorithm is based on the mobility prediction model, its performance is always better than RWM.

5.2 The Impact of Maximum Communication Distance

This part analyzes the impact of the maximum distance Dis between the user and the service to maintain communication on the total cost.

The result is shown in the Fig. 7, for the AM scheme, although the communication distance between the user and the service does not change, as the maximum communication distance increases, the overhead caused by frequent migration is the largest. Compared with AM scheme, NM scheme does not require migration cost, however, with the increase of the maximum communication distance, the performance of this scheme will also be affected. In addition, compared with the first two schemes, the performance of RWM is better. That is because, the algorithm has the ability to find a better migration path, however, the algorithm is not sensitive to user mobility and the performance of the algorithm is mediocre. MODEM can find a better migration path while considering the user mobility, so it has better performance.

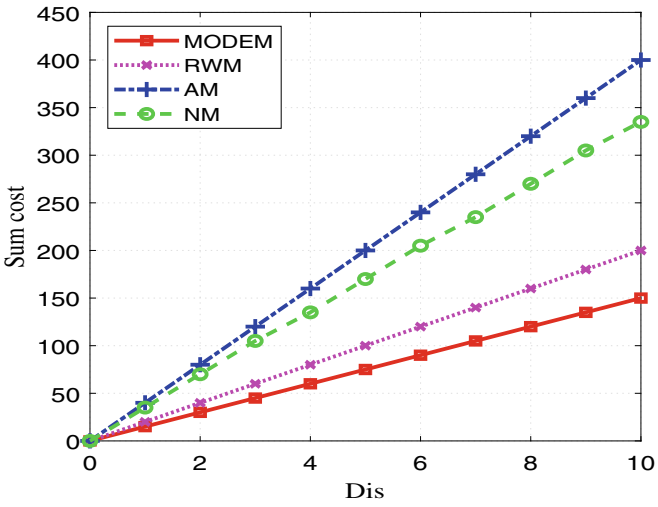


Fig. 7. The impact of the maximum distance *Dis* between the user and the service to maintain communication on the total cost.

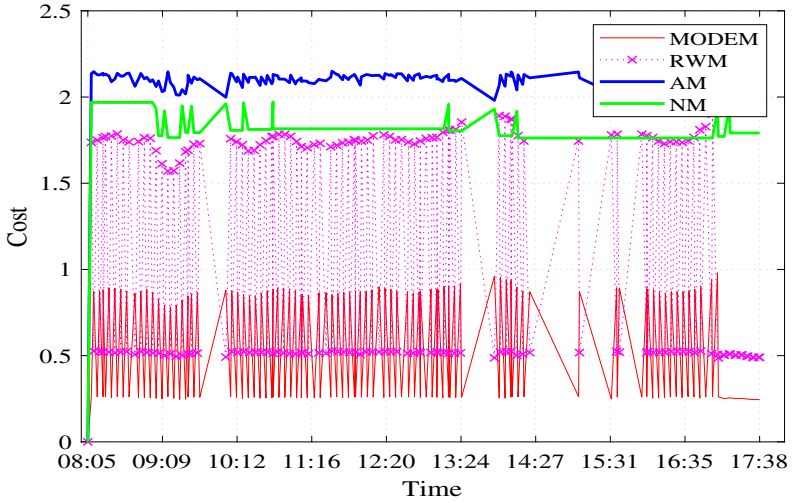


Fig. 8. Cost compared to alternative algorithms in trace-driven simulation.

5.3 Comparison of Simulation Results

The costs of the four migration strategies are compared by tracking the driver simulation and the results are shown in Fig. 8.

The track data of a certain taxi driver on a certain day was randomly selected (taxi data from 8:05 to 17:38 were selected in the experiment because taxi tracks were relatively active in the daytime) to compare the service migration cost. The

result is shown in Fig. 8, where the sparse part of the line is the absence of data (e.g. 14:27 to 15:40). The comparison results show that, in almost all cases, the proposed method has lower cost than other methods, which makes our algorithm better verified in the migration decision.

6 Conclusions

In this paper, we investigate the mobility-aware dynamic migration problem in mobile edge computing. Based on the analysis of the trajectory data from users, we propose a geometry-based user mobility model for predicting the probability of a user to move out from the coverage of the current edge server to another one. Considering the trade-off between QoS and cost, the service migration is formulated by a Markov decision problem, and an algorithm is designed for finding the long-term optimal solution. Finally, extensive simulations have been conducted to evaluate the effectiveness of the proposed algorithm, and the impacts of the model parameters are further analyzed. The experimental results show that our approach can dynamically find the optimal service migration decisions by reducing the cost while improving the QoS. This work is expected to provide a theoretical model and a practically solution of optimal service migration in MEC systems for mobile users.

Acknowledgment. This work was supported by National Natural Science Foundation of China (No. 61972414), Beijing Nova Program of Science and Technology (No. Z201100006820 082), Beijing Natural Science Foundation (No. 4202066), and Fundamental Research Funds for Central Universities (Nos. 2462018YJRC040 and 2462020YJRC001).

References

1. Villari, M., Fazio, M., Dustdar, S., Rana, O., Ranjan, R.: Osmotic computing: a new paradigm for edge/cloud integration. IEEE Cloud Comput. **3**(6), 76–83 (2016)
2. Abbas, N., Zhang, Y., Taherkordi, A., Skeie, T.: Mobile edge computing: a survey. IEEE Internet Things J. **5**(1), 450–465 (2018)
3. Ceselli, A., Premoli, M., Secci, S.: Mobile edge cloud network design optimization. IEEE/ACM Trans. Netw. **25**(3), 1818–1831 (2017)
4. Satyanarayanan, M.: The emergence of edge computing. Computer **50**(1), 30–39 (2017)
5. Peng, Q., et al.: Mobility-aware and migration-enabled online edge user allocation in mobile edge computing. In: 2019 IEEE International Conference on Web Services (ICWS), pp. 91–98. IEEE (2019)
6. Wang, S., Xu, J., Zhang, N., Liu, Y.: A survey on service migration in mobile edge computing. IEEE Access **6**, 23511–23528 (2018)
7. Taleb, T., Ksentini, A.: An analytical model for follow me cloud. In: Proceedings of IEEE GLOBECOM 2013, December 2013
8. Ksentini, A., Taleb, T., Chen, M.: A Markov decision process-based service migration procedure for follow me cloud. In: 2014 IEEE International Conference on Communications (ICC), pp. 1350–1354. IEEE (2014)

9. Wang, S., Urgaonkar, R., He, T., Zafer, M., Chan, K., Leung, K.K.: Mobility-induced service migration in mobile micro-clouds. In: Proceedings of IEEE MILCOM 2014, October 2014
10. Ouyang, T., Zhou, Z., Chen, X., et al.: Follow me at the edge: mobility-aware dynamic service placement for mobile edge computing. IEEE J. Sel. Areas Commun. **36**(10), 2333–2345 (2018)
11. Chowdhury, M., Rahman, M.R., Boutaba, R.: ViNEYard: virtual network embedding algorithms with coordinated node and link mapping. IEEE/ACM Trans. Netw. **20**(1), 206–219 (2011)
12. Minarolli, D., Mazrekaj, A., Freisleben, B.: Tackling uncertainty in long-term predictions for host overload and underload detection in cloud computing. J. Cloud Comput. **6**(1), 1–18 (2017). https://doi.org/10.1186/s13677-017-0074-3
13. Wang, S., Zhang, X., Zhang, Y., Wang, L., Yang, J., Wang, W.: A survey on mobile edge networks: convergence of computing, caching and communications. IEEE Access **5**, 6757–6779 (2017)
14. Kschischang, F.R., Frey, B.J., Loeliger, H.-A.: Factor graphs and the sum-product algorithm. IEEE Trans. Inf. Theory **47**(2), 498–519 (2001)
15. Wu, Q., Chen, X., Zhou, Z., Chen, L.: Mobile social data learning for user-centric location prediction with application in mobile edge service migration. IEEE Internet Things J. **6**(5), 7737–7747 (2019)
16. Xenakis, D., Passas, N., Merakos, L., Verikoukis, C.: Mobility management for femtocells in LTE-advanced: key aspects and survey of handover decision algorithms. IEEE Commun. Surv. Tutorials **16**(1), 64–91 (2013)
17. Liu, T., Bahl, P., Chlamtac, I.: Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks. IEEE J. Sel. Areas Commun. **16**(6), 922–936 (1998)
18. Ceselli, A., Premoli, M., Secci, S.: Mobile edge cloud net- work design optimization. IEEE/ACM Trans. Netw. 25(3), 1818–1831 (2017)
19. Nelson, M., Lim, B.-H., Hutchins, G., et al.: Fast transparent migration for virtual machines. In: USENIX Annual Technical Conference, General Track, pp. 391–394 (2005)
20. Zhou, A., Wang, S., Ma, X., Yau, S.S.: Towards service composition aware virtual machine migration approach in the cloud. IEEE Trans. Serv. Comput. **13**, 735–744 (2019)
21. Sung, J.-W., Han, S.-J., Kim, J.-W.: Virtual machine provisioning for computation offloading service in edge cloud. In: 2019 IEEE 12th International Conference on Cloud Computing (CLOUD), pp. 490–492. IEEE (2019)
22. Plachy, J., Becvar, Z., Strinati, E.C.: Dynamic resource allocation exploiting mobility prediction in mobile edge computing. In: IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–6. IEEE (2016)
23. Wang, S., Urgaonkar, R., Zafer, M., He, T., Chan, K., Leung, K.K.: Dynamic service migration in mobile edge computing based on Markov decision process. IEEE/ACM Trans. Netw. **27**(3), 1272–1288 (2019)
24. Machen, A., Wang, S., Leung, K.K., Ko, B.J., Salonidis, T.: Live service migration in mobile edge clouds. IEEE Wirel. Commun. **25**(1), 140–147 (2017)
25. Yuan, J., Zheng, Y., Xie, X., Sun, G.: Driving with knowledge from the physical world. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 316–324 (2011)