eNB Selection for Machine Type Communications Using Reinforcement Learning Based Markov Decision Process

Yu-Jui Liu, Shin-Ming Cheng, *Member, IEEE*, and Yu-Lin Hsueh

***Abstract*—Machine type communication (MTC), as one of the most promising technologies in the future wireless communication, has brought mobile communication network into a new level. The breakthrough of cutting-edge technology and broad coverage of cellular networks in long term evolution advanced network constitute an ideal platform for ubiquitous MTC service provisioning on a large scale. However, under the traditional attach approach, the massive MTC devices always select the evolved NodeB (eNB) with the best signal quality for the attachment, thereby causing the network congestion and overload. As a result, it is necessary to design an efficient eNB selection scheme to avoid overload issue. In this paper, by modeling MTC arrivals using nonhomogeneous poisson process with the memoryless property, we formulate the eNB selection problem as a Markov decision process (MDP) and try to compute an optimal solution. Since the network parameters for MDP are not obtained easily, a learning-based algorithm with value-difference based exploration (VDBE) policy who could leverage both present conditions and expected future demands is further proposed. The performances of the proposed reinforcement learning (RL) VDBE, other RL-based, the legacy best-signal-quality, and MDP without RL eNB-selection schemes are analyzed in terms of blocking probability, transmission rate, and load balancing. The simulation results show that our scheme has the best performance on the blocking probability and loading balancing while an acceptable result on the transmission rate. It is suitable for the MTC environment with a highly changed number of MTC arrivals while without high requirement on the transmission rate.**

***Index Terms*—eNB selection, machine type communications, Markov decision process, reinforcement learning, value-difference based exploration.**

I. Introduction

MACHINE Type Communications (MTC), also known as Machine-to-Machine (M2M) communications in cellular networks, appear to be one of the most promising technologies in the future wireless communications since they enable the communications among machines without or with few human interactions. Various applications are facilitated by MTC devices, such as security, healthcare, education, transportation, and monitoring [1]. For example, MTC devices can provide health-care applications that monitor the status of the elders by sensors so as to offer instant feedbacks under certain circumstance [2]. 3GPP suggests that the deployment of MTC applications is standardized in Long Term Evolution Advanced (LTE-A) network for its higher capability and broader coverage area [3]. The ultimate goal of MTC is to enable comprehensive and ubiquitous connectivity for all the MTC devices in the environment.

In contrast with traditional Human-to-Human (H2H) communications using smartphones, such as voice and video streaming services, the data transmitted by MTC is smaller while the number of devices involved is much larger. It is expected that the number of MTC devices will proliferate to billions in 2020, and the number of MTC device connecting to a single base station (Evolved NodeB; eNB in LTE-A terminology) will be from 10000 to 100000 [4]. Moreover, the traffic patterns between MTC and H2H are different. In particular, most MTC applications generate concurrent, periodic, and short-period uplink traffic, and remain static compared to H2H devices. Taking temperature and humidity of environment as well as pulse and blood pressure of a monitored patient as examples, the amount of traffic transmitted or received by MTC devices is limited, thereby requiring smaller transmission rate comparing with traditional H2H services.

Considering the aforementioned characteristics of MTC, it is expected that simultaneous access attempts from a large number of devices to eNBs occur. For instance, a massive number of sensors or monitoring devices deployed in a railroad will be triggered when a train is about to pass. Scenarios like this happen frequently in the real world and might cause not only signaling load but also degradation of performance on the access network. As the example shown and suggested in [5]–[8], one of the MTC issues is the overload and congestion in radio access network.

The reason for causing congestion is that when performing random access procedure, MTC devices contend for the preamble by using random access procedure to access the LTE-A network [5]. Obviously, the random access procedure is an important issue, and plenty of the papers probing MTC are based on random access congestion problem [9], [10]. However, few are considering the impact from the perspective of eNBs, that is,

Manuscript received October 8, 2016; revised April 6, 2017; accepted July 5, 2017. Date of publication July 21, 2017; date of current version December 14, 2017. This work was supported by the Ministry of Science and Technology, Taiwan, under Grant 105-2628-E-011-001-MY3. The review of this paper was coordinated by Dr. Y. Song. *(Corresponding author: Shin-Ming Cheng.)*

The authors are with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: m10215078@mail.ntust.edu.tw; smcheng@mail.ntust.edu.tw; m10515096@mail.ntust.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2017.2730230

eNBs could suffer from the overload issue and fail to distribute the resource even when MTC devices finish random access procedure. Without a strong eNB selection approach, network load might fluctuate intensely, and overload could happen easily. As a result, it is necessary to design an efficient and robust eNB selection scheme MTC devices not only to fit the MTC characteristics but also to keep the network load as stable as possible.

Generally speaking, to exhibit temporal and spatial concurrence of MTC arrival, Beta distribution is a convincing way to demonstrate the burst arrival [5], [11]. In this paper, we utilize non-homogeneous Poisson process (NHPP) to modulate the Beta distribution [12] and formulate the eNB selection problem as Markov Decision Process (MDP) for the sake of memorylessness feature shown by both NHPP and MDP. By considering both information of network side (i.e., loading in eNBs) and of device side (i.e., the transmission rate), MDP could dynamically meet operator objectives while maximizing utility of MTC devices.

Moreover, when network parameters are not known, a Reinforcement Learning (RL) approach with Value-Difference Based Exploration (VDBE) policy [13], [14] is introduced to learn the requirements and preferences of MTC devices through interacting with them. In particular, RL could help MTC device to select the eNB dynamically either by random (known as exploration) or by previous experiences (known as exploitation). The advantage of the VDBE is that exploration actions are selected in situations when the knowledge about the environment is uncertain, which is indicated by fluctuating values during learning. In the learning process, the more information we have known, the higher probability we obtain to select the optimal eNB among the ones we know rather than choosing randomly.

We conduct extensive simulation experiments to evaluate the performance of the proposed RL-VDBE eNB-selection scheme, the legacy best-signal-quality scheme, and related RL-based schemes in terms of blocking probability, transmission rate, and load balancing. The simulation results show that our scheme has the best performance on the blocking probability and loading balancing while an acceptable result on the transmission rate. It is suitable for the MTC environment with highly changed number of MTC arrivals while without high requirement on the transmission rate.

The rest sections of this paper is organized as follows. Section II describes the preliminary and related work. In Section III, system model in this paper is presented. Section IV shows the details of problem formulation and proposed scheme corresponding to our system model. Follow by the simulations and numerical results in Section V. Finally, we conclude this paper in Section VI.

II. Related Work

In MTC network, it is expected to see an extreme scenario that a massive number of devices attempt to access the network. Generally speaking, when an MTC device finishes radio access procedure and obtains the exclusive preamble, it gets the right to connect eNB. So far most papers probing the related issue of MTC are around the level of random access procedure issue such as [9], [10], [15] in Access Class Barring scheme, [16]–[18] in Slotted Access scheme. The goal is to make the devices to access eNBs as many as possible.

On the other hand, from the point of the network terminal, massive devices pose a great challenge on eNBs. In [19], it first takes access intensity among eNB into consideration, several eNB selection algorithms are reviewed under different scenarios. In [6], different overload control approaches are proposed to avoid overload issue which is caused by random channel access of MTC devices. The authors propose a reinforcement learning based eNB selection algorithm that allows MTC devices to decide the target eNB. However, both papers only take network impacts into consideration, and the quality of experience is not concerned for MTC devices. Authors in [20] concern parameters in the user side (i.e., device throughput) and network side (i.e, transmission delay from eNB). Please note that, among all aforementioned papers [6], [19]–[21], MTC devices are regarded as decision makers and could select the target eNB on their own. However, it is difficult to obtain the precise network information at the user side. For example, instead of the real load condition, only transmission delay is considered, which might result in a sub-optimal solution. As a result, it is more appropriate to set the network side as decision maker such as [14], [22]–[25] due to the full understanding of networks.

Among selection issues, MDP is a powerful analytical tool for sequential decision making under uncertainty. In [23], traditional MDP requires explicit transition probability which is hard to set a precise value. Accordingly, [6], [14], [21], [22], [24], [26] integrate reinforcement learning to strengthen MDP by learning environments. This approach works particularly well in real environments for the reason that most environments are dynamic. Therefore, it is unreasonable and impractical to assume explicit transition probability in advance.

III. System Model

A. MTC Traffic Model

To consider the speical feature that a large number of IoT devices attempting to access the network in short time, we apply 3GPP model developed in [5] as our traffic model. In particular, it suggests that MTC packet arrivals over a given time period T with Beta distribution as follows

$$f_T(t) = \frac{t^{\alpha-1}(T-t)^{\beta-1}}{T^{\alpha+\beta-1}Beta(\alpha,\beta)}, \tag{1}$$

where $Beta(\alpha,\beta)$ indicates the Beta function, and parameters α and β are set as 3 and 4, respectively [11], [27], [28]. Please note that the original Beta distribution $f(t)$ over $[0,1]$ is rescale as $f_T(t)$ over a controllable time interval $[0,T]$ in this model [11]. By separating the arrival period into $\frac{T}{\Delta t}$ timeslots with index j, the access intensity in the j-th timeslots is $N\int_{t_j}^{t_j+\Delta t} f_T(t)$, where t_j is the time of the j-th timeslot and N is the expected number of MTC devices. According to the observation that the 3GPP model is equivalent to a modulated Poisson process (i.e., NHPP) with mean arrival rate $\lambda(t)$ in each timeslot [12], the expected number of arrival in timeslot t follows a modulated

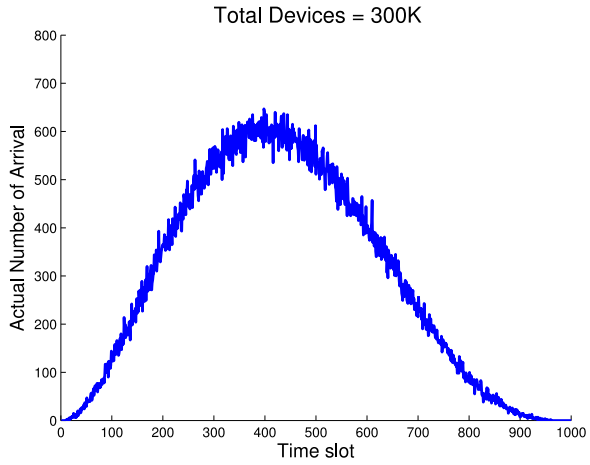


Fig. 1. 3GPP MTC traffic model: Interpretation as modulated Poisson process.

Poisson process as

$$\lambda(t) = f_T(t) \times \frac{\Delta t}{T} \times N. \tag{2}$$

Fig. 1 verifies the correctness of (2) with parameters $N =$ 300 K and $\frac{T}{\Delta t} = 10$ K. The value of N selected in this paper is much larger that suggested by [5] (i.e., 30 K) since the access from MTC devices must be more massive in the near future.

B. Network Model

We consider a LTE-A cellular network model consisting of irregular deployed eNBs. The first spatial distribution of eNBs is assumed to follow a homogeneous Poisson Point Process (PPP) [29] with density λ, where the number of eNBs is generated by using Poisson distribution while those eNBs are deployed in the networks by using Uniform distribution. PPP is considered as a good candidate model for the location of eNBs in cellular networks since tractable results on throughput are derived [30], [31]. To comprehensively evaluate the proposed scheme in a realistic environment, real data of eNBs deployments with the aid of OpenCellID project [32] is applied. The OpenCellID project maintains a complete and open database of eNB information worldwide, and we can easily retrieve latitude and the longitude information of eNBs according to the target urban area. Regarding the location of MTC devices, a uniform distribtuion is applied.

The successful reception of a transmission at an MTC device depends on if the Signal to Interfere and Noise Ratio (SINR) observed by the MTC is larger than an SINR threshold (denoted by θ). LTE-A standard suggests that the MTC to select the eNB which offers the highest recieved signal strength (know as the RSS-based solution). We assume that each eNB provides δ resource units for MTC devices, that is, δ MTCs can be served by one eNB simultaneously. At each arrival moment, MTC devices arrive in the form of NHPP model and Mobility Management Entity (MME) is served as the control center with full knowledge about the eNBs and MTC devices. Therefore, MME is able to selection the eNB according to different selection methods for each MTC devices.

C. Performance Metrics

To evaluate the load balance of LTE-A network due to numerious arrivals of MTC devices, blocking probability p_b and transmission rate μ are included as the performance metrics. Typcially, an MTC request will be blocked if the serving eNB has no more resource (i.e., overloading) or if the MTC device is located outside the coverage area of any eNB. As a result, the blocking probability p_b can be expressed as

$$p_b = \frac{N_f}{N}, \tag{3}$$

where N_f is denoted as the number of failed MTC requests. Moreover, we use transmission rate μ to understand the quality of transmission for the successful MTC connections, and it is defined as

$$\mu = \text{Bandwidth} \times \log_2(1 + \text{SINR}). \tag{4}$$

IV. Problem Formulation

As we mentioned in Section III-A, the arrival process of MTC devices follows modulated Poisson process, and thus the memoryless property is hold. Consequently, the eNB selection-making process can be formulated as a MDP, which typically acts as an analytical tool for sequential decision making in a dynamic environment.

A. Markov Decision Process (MDP) Formulation

In MDP, future states depend only on the current state rather than the former ones (i.e., it guarantees the Markov property). To apply MDP to formulate the eNB selection-making process, first we have to define network states, actions, transition probability, and reward. Next we use approaches such as linear programming, value iteration, and policy iteration to make an optimal decision. The details are descrbied as follows.

States: We simply define a state of an MTC device $n \in N$ as the eNB it selected. In particular, we say that an MTC device is in state s_k if it is connecting to the k-th eNB.

Actions: When an MTC device arrives, the center controller (i.e., MME) will take an action according to the current state s. In this case, the action $a_{(s,s')}$ indicates a transition from a certain state s to target state s'. If the action is available (i.e., the MTC device is within the coverage of eNBs), it is set as 1; otherwise, it will be 0, that is,

$$a(s, s') = \begin{cases} 1, & \text{if the MTC device is within the} \\ & \text{coverage area of eNB of state } s, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Transition Probability: The transition probability $p(s, s', a)$ means the probability of choosing a certain action, resulting in a transition from state s to s'. It depends on user arrival rate, preference, MME setting, and decision-making algorithm. However, these are not easy to be obtained in a real network. In this case, reinforcement learning is considered as a suitable alternative. The MME learns what action to take by trial-and-error instead of estimating user behaviors. Among

the existing RL algorithms, we select VDBE for its compatibility, and details of RL algorithms will be introduced in the next subsection.

Reward: The reward is obtained after the request of an MTC device is accepted by an eNB and no reward is given if the request is blocked. The reward consists of we consider received SINR at user side and network loading at network side. The load of k-th eNBs simply is defined as

$$L_k = \frac{\text{current number of users}}{\text{capacity}}. \tag{6}$$

Overall, let $r(s, s', a)$ denote the feedback received when an MTC device takes an action a and turns out to be state s' from s, it can be defined as:

$$r(s,a) = \begin{cases} r(s,s',a) = w_1 \times (1 - L_{s'}) \\ \qquad\qquad +w_2 \times \mu_{s'}, & \text{accepted}, \\ 0, & \text{blocked}, \end{cases} \tag{7}$$

where w_1 and w_2 are variables for normalization.

Obviously, knowing $r(s, a)$ and $p(s, s', a)$ is the basis to dervie an optimal policy. However, the transition probability $p(s, s', a)$ is really difficult to be estimated exactly. Alternatively, reinforcement learning comes as a solution because it can learn from previous experiences in a trial-and-error fashion, and choose the appropriate actions without explicit state transition probability.

B. Reinforcement Learning (RL)

In the proposed RL framework, MME plays a role of the agent, and interacts with MTC devices. At each discrete timeslot $t \in \{0, 1, 2, ...\}$, the agent is in a certain state $s \in S$. After the selection of an action $a \in A$, the agent receives a reward and passes into a successor state s'. The decision of choosing an action in a certain state is characterized by a policy $\pi(s) = a$. We denote $Q(s, a)$ as a state-action value function, that is, the expected discounted reward when starting in state s and selecting action a. Our goal is to find an optimal solution policy π_{opt} which maximizes the state-action function.

$$\pi_{\text{opt}}(s) = \arg\max_{a \in A} Q(s, a). \tag{8}$$

1) Q-Learning Function: In this case, value functions are learned by sampling observations of the interaction between the agent and its environment. At timeslot t, when state-action pair $p(s, a)$ is visited (i.e., the agent performs an action a in a state s), the agent earns reward $r(s, a)$, and ends up in state s' at $t + 1$. Then the Q-value function of state-action pair is updated as:

$$Q(s,a) = Q(s,a) + \rho \left\{ r(s,a) + \gamma \max_{a \in A} Q(s', a') - Q(s,a) \right\}. \tag{9}$$

2) Exploration and Exploitation: At every decision epoch, the agent makes the decision to maximize the effects of actions decided by the agent. Two modes can be applied as follows. In order to discover an effective action, the agent needs to try different actions in the beginning, thus the agent may select the possible action by random, this is called *exploration* mode. Otherwise, the agent will choose the action it has tried in the past and find the one which provides highest reward, which is known as *exploitation* mode. Exploitation is obviously suitable for a stable environment where the previous experience is useful while exploration is more appropriate to make a new discovery when environment changes rapidly. Since RL is an dynamic iterative learning algorithm, exploration and exploitation should be simultaneously performed.

By adjusting the probability between exploration and exploitation, many variants in RL-family are proposed, which have different properties. With massive MTC arrivals, our environment is highly dynamic, and we need a very flexible algorithm to take the advantages of exploration and exploitation efficiently.

3) Value-Difference Based Exploration (VDBE) Policy: In our work, we adopt VDBE, which introduces a state-dependent exploration probability, $\varepsilon(s)$, to tackle selection issue. Instead of a global parameter, $\varepsilon(s)$ in VDBE is state-dependent, which means every eNB has different exploration probability. It is more suitable in our network because the simultaneous condition of each eNB is not the same. The idea of VDBE is to be more explorative when the circumstance is uncertain, which is indicated by fluctuation (value-difference) function $f(s, a, \sigma)$ during learning, such as the beginning of learning process. The function $f(s, a, \sigma)$ is obtained by computing after each learning step, and is expressed as:

$$f(s,a,\sigma) = \frac{1 - e^{\frac{-|Q_{t+1}(s,a) - Q_t(s,a)|}{\sigma}}}{1 + e^{\frac{-|Q_{t+1}(s,a) - Q_t(s,a)|}{\sigma}}}. \tag{10}$$

On the other hand, exploration should be reduced as far as more information and knowledge is gained. Generally speaking, more information gained indicates a very small or no difference on reward value. The exploration probability $\varepsilon(s)$ can be expressed as

$$\varepsilon_{t+1}(s) = \delta \times f(s,a,\sigma) + (1 - \delta) \times \varepsilon_t(s), \tag{11}$$

where $\sigma \in [0.1]$ is a positive constant called inverse sensitivity, determining the influence of the selected action on the state-dependent exploration probability. δ is the number of actions in current state. At the beginning of the learning process, all exploration probabilities are initialized to arbitrary (e.g., $\varepsilon_t(s) = 1$ for all states).

The overall learning procedure is shown in Algorithm 1.

V. Numerical Results

This section presents the simulation results of the proposed VDBE scheme using the network model described in Section II. The performance of this scheme is compared to that of a pure MDP scheme without introducting RL, and a typical RSS-based scheme. The RSS-based scheme assumes that each MTC device always select eNB with best signaling quality. The MDP scheme without RL does not determine the impact the action has caused on the current state. The comparison could show the sole effects of two critical components of VDBE scheme, that is, MDP part and RL part. To further evaluate the importance of adaption of

Algorithm 1: Q-learning with VDBE policy..

Initialization:
- $\texttt{timeslot} \leftarrow 0$.
- $Q(s,a) \leftarrow 0, \forall s \in S$ and $\forall a \in A$.
- $\delta(s) \leftarrow 0, \forall s \in S$.

while do $\texttt{timeslot} \leq \texttt{ending time}$
  MME observes state s
  if exploration **then**
    chooses an action $a(s,a)$ randomly
  else
    chooses $a = \max_{a \in A} Q(s,a)$
  end if
  Perform a
  $\delta(s) \leftarrow \delta(s) + 1$
  Reward $r(s,a)$ gained
  Update Q-value according to eqn (9)
  Update $\varepsilon(s)$ according to eqn (11)
  $s \leftarrow s'$
  $\texttt{timeslot} \leftarrow \texttt{timeslot} + 1$
end while

VDBE scheme in high dynamice environment, where numerous MTC devices arrive massively, we also compare RL schemes with different polices as follows.

1) ε-greedy policy. At decision epoch the agent performs exploration with probability ε, and it exploits stored Q-values with probability $1-\varepsilon$. Here, ε is a tuning parameter between $0 < \varepsilon < 1$, which is sometimes changed, either according to a fixed schedule, or adaptively based on some heuristics. Typically, in order to improve long-term network reward, the exploration is never stopped, but rather reduce as time goes by.
2) softmax policy. Instead of a exploration probability, the softmax utilizes action-selection probabilities which are determined by ranking the value-function estimates using a Boltzmann distribution as follows.

$$P\{s_t = s, a_t = a\} = \frac{e^{\frac{Q_t(s,a)}{\tau}}}{\sum_b e^{\frac{Q_t(s,b)}{\tau}}}, \tag{12}$$

where τ is a positive parameter called temperature. High temperatures cause all actions to be nearly equitable, whereas low temperatures cause greedy action selections.

A. Network Environment

We consider a LTE-A cellular network model consisting of irregular deployed eNBs with two different distribution models. By applying PPP, we could easily model the distribution of eNBs with acceptable accuracy in numerical results. Moreover, by exploiting the real data of eNBs from OpenCellID project, a more realistic result will be retrieved. In particular, we consider 209 eNBs in around 20 km^2 from the central of Taipei city, Taiwan. Fig. 2 shows the filtered-out eNB locations using OpenCellID and PPP in Google Map of Taipei City. Under a specific eNB deployment, we generate the arrivals of MTC devices in each timeslot t by using our MTC traffic model and then deploy

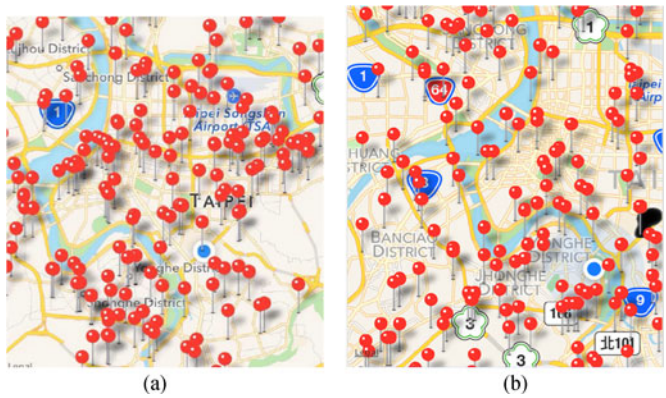


Fig. 2. eNB distribution scenarios. (a) Open Cell ID, (b) PPP.

TABLE I
PARAMETER SETUP

Parameter	Value
The size of target area	20 km * 20 km
Number of MTC devices N	from 100K to 1000K
Location of MTC devices	uniform distribution
Number of eNBs K	209
Location of eNBs	Open Cell ID and PPP
MTC Arrival rate/departure rate	3
Capacity of eNBs	from 2K to 3K
Learning epochs T	10K
SINR threshold θ	0 dB
σ in VDBE	0.3
τ in Softmax	1
ρ of Q-value	0.5
γ of Q-value	0.1

those MTC devices uniformly. The simulation experiments are built on Matlab platform, and the parameter setup follows the 3GPP model [5]. The details of parameter setup are shown in Table I. Please note that the parameters related RL schemes are determined according to the suggestions from [13]. As we mentioned in Section III-C, the performance metrics considered in the simulation experiment are blocking probabiilty p_b, transmission rate μ, and load balancing situation of the whole network.

B. Simulation Results

Effects of N on p_b under different eNB selection schemes: Figs. 3 and 4 respectively plot p_b against N in OpenCellID and PPP environments under RSS-based, MDP without RL, RL-softmax, RL-greedy, and RL-VDBE eNB selection schemes. We can observe a phenomenon that as N increases, p_b increases. It is simply due to that as N becomes larger, the eNB with fixed capacity could not accept more MTC requests, and p_b becomes larger.

The reason that RSS-based scheme gets worst performance is due to that the MTC devices always connect to the cloest eNB withoug considering if the eNB is overloading or not. This scheme simply only takes user side information into consideration. Regarding traditional MDP, it treats all possible eNBs as options, which might be more flexible than RSS-

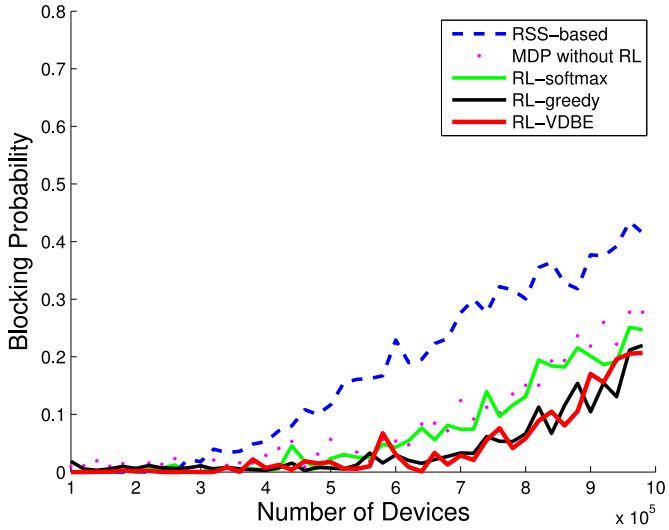


Fig. 3. Effects of the number of MTC devices on the blocking probability under different eNB selection schemes in PPP scenario.

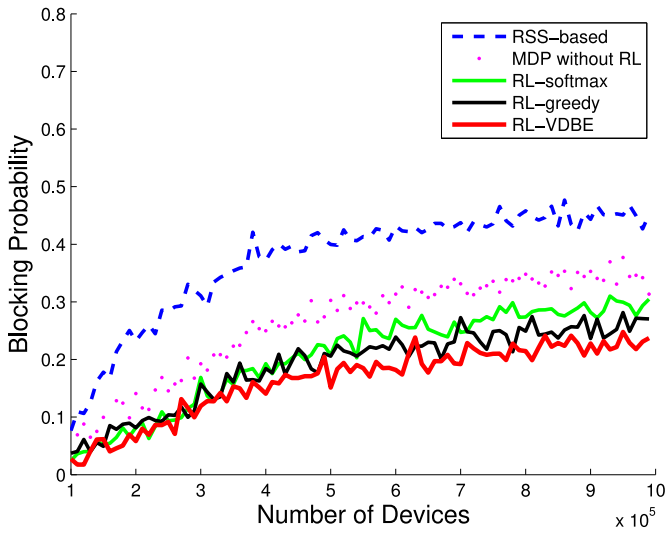


Fig. 4. Effects of the number of MTC devices on the blocking probability under different eNB selection schemes in OpenCellID scenario.

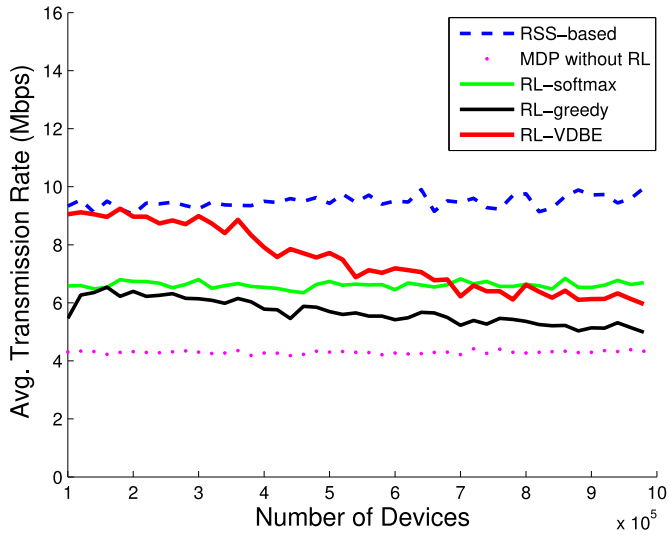


Fig. 5. Effects of the number of MTC devices on the average transmission rate for non-blocked MTC devices under different eNB selection schemes in PPP scenario.

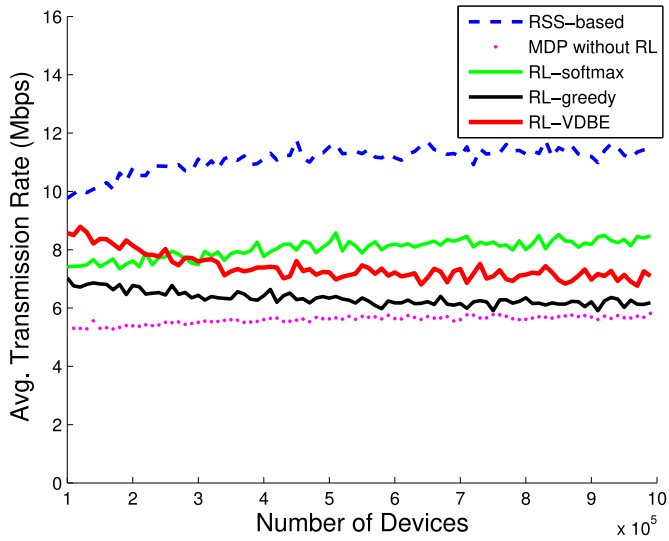


Fig. 6. Effects of the number of MTC devices on the average transmission rate for non-blocked MTC devices under different eNB selection schemes in OpenCellID scenario.

based scheme. However, without favoring the one with higher reward, MDP without RL could not get an optimal results. The rest RL schemes outperform the previous two schemes for not only considering network circumstances but learning to obtain the target eNB which provides higher reward. Among different policies in RL, VDBE is able to be pro-exploration when the situation is unsteady. In our simulation environment, the simultaneous access requests from numerous MTC devices and consequent drop-out could fluctuate the loading of eNBs in the target area. Hence, the VDBE policy performs particular suitable and better than softmax and greedy policies.

By comparing the results of Figs. 4 and 3, we observe that eNBs in OpenCellID scenario are more unevenly distributed, which makes it more difficult to reach the overall network balance than the PPP scenario with even distributed eNBs. We could get an exciting result that in the realistic environment where eNBs are distributed unevenly, the proposed RL-VDBE scheme outperforms all other schemes more significantly.

Effects of N on μ under different eNB selection schemes: Figs. 5 and 6 respectively plot μ averaged by non-blocked MTC devices against N in OpenCellID and PPP environments under RSS-based, MDP without RL, RL-softmax, RL-greedy, and RL-VDBE eNB selection schemes. These two figures provide a local view because only non-blocked devices are concerned. The results show that RSS-based gains the best performance. It is due to the reason that MTC devices will choose the eNB with highest RSS, thereby resulting in the highest transmission rate. The RL schemes have lower transmission rate since they also include network loading as the objective for the optimization. Regarding tradi-

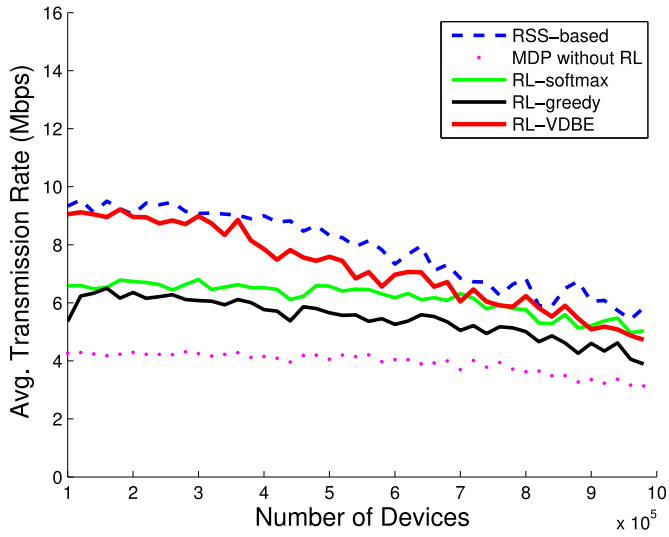


Fig. 7. Effects of the number of MTC devices on the average transmission rate for all MTC devices under different eNB selection schemes in PPP scenario.

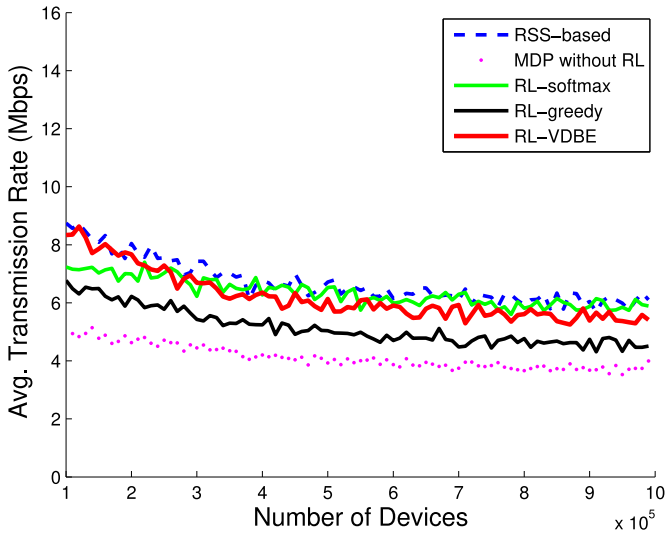


Fig. 8. Effects of the number of MTC devices on the average transmission rate for all MTC devices under different eNB selection schemes in OpenCellID scenario.

tional MDP, it does not treat the eNB with higher reward as preference, resulting in a complete random decisions and lowest transmission rate.

We can observe in these two figures that there is a decreasing trend for μ when N increases. It is due to the reason that as N increase, the number of non-blocked MTC devices increases while the total resource an eNB could offer is fixed. As a result, the average transmission rate for non-blocked MTC will slightly decrease. However, μ of RSS-based, RL-softmax, and MDP without RL schemes in OpenCellID scenoario (refer Fig. 6) slightly increase when N increases. As we can observe in Fig. 4, these three schemes have a higher p_b when N is larger. The higher average transmission rate for non-blocked MTC devices in these schemes are beneficial from smaller number of non-blocked MTC devices. These results can reflect a phenomenon that RSS-based, RL-softmax, and MDP without RL schemes facilitate the non-blocked MTC devices since a higher average transmission rate is supported, however, those schemes might not applied by the operator since smaller number of MTC devices can be served.

Figs. 7 and 8 respectively plot μ averaged by all MTC devices against N in OpenCellID and PPP environments under RSS-based, MDP without RL, RL-softmax, RL-greedy, and RL-VDBE eNB selection schemes. By taking both blocked and non-blocked MTC devices into consideration, the result could reflect the performance of the oveall network in a more fairness fashion. We can observe that although RSS-based scheme still has the highest transmission rate, the gap between RSS-based and RL-based schemes become smaller, especially in the case of OpenCellID scenario. The decending trend of μ when N increases becomes obvious and reason is same as we descrbied previously. The proposed RL-VDBE scheme performs better than RL-softmax when $N < 300000$ and performs slightly worse than RL-softmax when $N > 300000$. It implies that with $N < 300000$ (i.e., an appropriate number of MTC devices nowadays), RL-VDBE performs best in RL-family both in the perspecitves of p_b and μ. Please note that from the perspective of general MTC devices, relatively small transmission rate is required than H2H and thus the performance of RL-VDBE is acceptable even when $N > 300000$.

Loading of each eNB in OpenCellID scenario: Fig. 9 investigates the loading of each eNB in OpenCellID environment under RSS-based, MDP without RL, RL-softmax, RL-greedy, and RL-VDBE eNB at the peak arrival moment (i.e., $T = 5000$) when $N = 600000$. As we can see among them, RSS-based scheme leads to the most fluctuating outcome because it always chooses the eNB with strongest RSS without considering the condition of loading. An exciting result is that the proposed RL-VDBE scheme outperforms all other schemes from the perspective of load balancing since its value-difference feature could directly impact the exploration probability when the load fluctuates. That is, when the burst arrival occurs, the load of certain eNB will get much heavier, causing the exploration in such eNB state become higher in order to seek the better eNB.

C. Notes on Complexity

This subsection analyzes the computational complexity to demostrate the implementation cost of the proposed RL-VDBE scheme [33]. The computational complexity of the RL-based algorithms is limited by the number of update operations of the Q-value. Obviously, the agent can be in one state at any moment, and thus the computation complexity depends on the number of actions (denoted as n). We summarize the total number of operations required by each element of the RL-VDBE algorithm in Table II. The complexity of the proposed RL-VDBE scheme is acceptable comparing with other computational-consuming operations in MME.

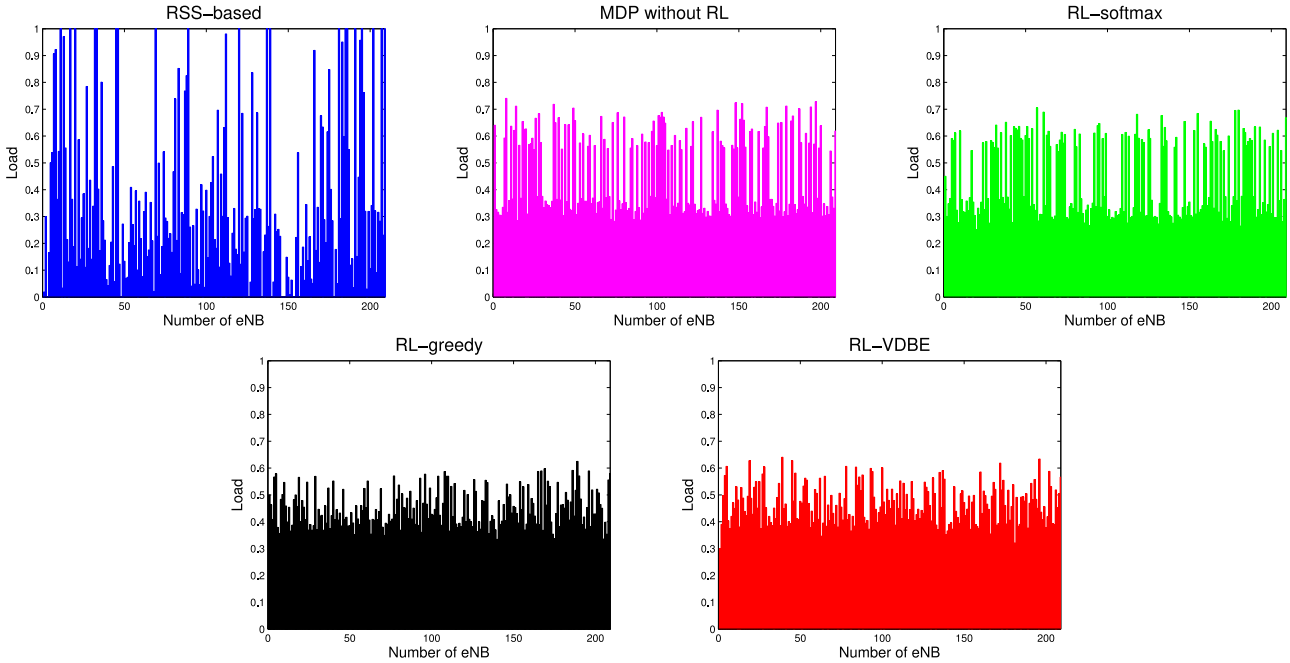


Fig. 9. Loading of each eNB under different eNB selection schemes when $T = 5000$ (i.e., peak arrival) and $N = 600000$ in OpenCellID scenario.

TABLE II
COMPUTATIONAL REQUIREMENT FOR RL-VDBE SCHEME

Step	Instruction
Identification of current and previous states	2 read
Choose action	n read $n-1$ comparisons
Update Q	$n+1$ read n comparisons
Update ε	4 read 3 division 2 exponentiation

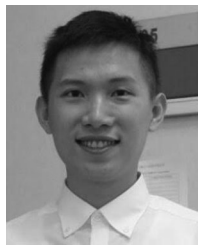
VI. CONCLUSION

In this paper, we attempt to tackle the MTC overload issue from the perspectives of both network side and device side. Due to the MTC traffic characteristic, it is estimated that a massive number of MTC devices will send requests to eNBs, resulting in the congestion and overload in RAN. By modeling massive MTC arrivals as an NHPP, we formulate the eNB selection problem as an MDP and integrate a learning-based scheme with policy, VDBE, to manage the uncertainty and fluctuation circumstances. The simulation shows that the proposed RL-VDBE scheme outperforms traditional RSS-based algorithm and standard MDP without the support of RL by lowering 10% to 20% blocking probability. More specifically, RL-VDBE scheme leads to better performance compared to other RL-based schemes in terms of load balancing according to our designated reward function. The reason behind the improved performance is that RL-VDBE scheme addressing exploration probability from each individual state rather than a global standard. Moreover, VBDE is capable of sensing the change of network environment and adapting the ratio between exploration and exploitation. In summary, the simulation results have proved that our proposed RL-VDBE scheme is particularly fitting in MTC to avoid overload circumstance.

REFERENCES

[1] 3GPP TR 22.368 v13.0.0, "Service requirements for Machine-Type Communications," Jun. 2014.
[2] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.
[3] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 525–549, Oct. 2015.
[4] S. Lucero and D. Bonte, "Cellular M2M connectivity services: The market opportunity for moblie operators, MVNOs, and other connectivity service providers," ABI Research Report, Dec. 2011.
[5] 3GPP TR 37.868 v11.0.0, "Study on RAN improvements for Machine-type Communications," Oct. 2011.
[6] M. Hasan, E. Hossain, and D. Niyato, "Towards understanding the fundamentals of mobility in cellular networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.
[7] F. Cao and Z. Fan, "Cellular M2M network access congestion: Performance analysis and solutions," in *Proc. IEEE 9th Int. Conf. Wireless Mobile Comput., Netw. Commun.*, Oct. 2013, pp. 39–44.
[8] A. G. Gotsis, A. S. Lioumpas, and A. Alexiou, "M2M scheduling over LTE: Challenges and new perspectives," *IEEE Veh. Technol. Mag.*, vol. 7, no. 3, pp. 34–39, Sep. 2012.
[9] J.-P. Cheng, C.-H. Lee, and T.-M. Lin, "Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," in *Proc. IEEE GLOBECOM Workshop*, Dec. 2011, pp. 368–372.
[10] T. P. de Andrade, C. A. Astudillo, and N. L. S. da Fonseca, "Random access mechanism for RAN overload control in LTE/LTE-A networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 7607–7612.
[11] M. Laner, P. Svoboda, N. Nikaein, and M. Rupp, "Traffic models for machine type communications," in *Proc. 10th Int. Symp. Wireless Commun. Syst.*, Aug. 2013, pp. 651–655.
[12] R. C. D. Paiva, R. D. Vieira, and M. Saily, "Random access capacity evaluation with synchronized MTC users over wireless networks," in *Proc. IEEE 73rd Veh. Technol. Conf.*, May 2011, 5 pages.
[13] M. Tokic, "Adaptive ε-greedy exploration in reinforcement learning based on value differences," in *Proc. 33rd Annu. German Conf. Adv. Artifi. Intell.*, Sep. 2010, pp. 203–210.
[14] M. E. Helou, M. Ibrahim, S. Lahoud, K. Khawam, D. Mezher, and B. Cousin, "A network-assisted approach for RAT selection in heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1055–1067, Jun. 2015.

[15] H.-L. He, Q.-H. Du, H.-B. Song, W.-Y. Li, Y.-C. Wang, and P.-Y. Ren, "Traffic-aware ACB scheme for massive access in machine-to-machine networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 2226–2231.
[16] M. K. Giluka, A. Prasannakumar, N. Rajoria, and B. R. Tamma, "Adaptive RACH congestion management to support M2M communication in 4G LTE networks," in *Proc. IEEE ANTS 2013*, Dec. 2013, 6 pages.
[17] S.-T. Sheu, C.-H. Chiu, Y.-C. Cheng, and K.-H. Kuo, "Self-adaptive persistent contention scheme for scheduling based machine type communications in LTE system," in *Proc. IEEE Int. Conf. Sel. Topics Mobile Wireless Netw.*, Jul. 2012, pp. 77–82.
[18] G. C. Madueno, S. Stefanovic, and P. Popovski, "Efficient LTE access with collision resolution for massive M2M communications," in *Proc. IEEE GLOBECOM Workshop*, Dec. 2014, pp. 1433–1438.
[19] C.-W. Chang, J.-C. Chen, C. Chen, and R.-H. Jan, "Scattering random-access intensity in LTE machine-to-machine (M2M) communications," in *Proc. IEEE GLOBECOM Workshop*, Dec. 2013, pp. 4729–4734.
[20] A. Mohammed, A. S. Khwaja, A. Anpalagan, and I. Woungang, "Base station selection in M2M communication using Q-learning algorithm in LTE-A networks," in *Proc. IEEE 29th Int. Conf. Adv. Inf. Netw. Appl.*, Mar. 2015, pp. 17–22.
[21] N. Abbas, S. Taleb, H. Hajj, and Z. Dawy, "A learning-based approach for network selection in WLAN/3G heterogeneous network," in *Proc. IEEE 3rd Int. Conf. Commun. Inf. Technol.*, Jun. 2013, pp. 309–313.
[22] J. Suga and R. Tafazolli, "Joint resource management with reinforcement learning in heterogeneous networks," in *Proc. IEEE 78th Veh. Technol. Conf.*, Sep. 2013, 5 pages.
[23] J. Buhler and G. Wunder, "Traffic-aware optimization of heterogeneous access management," *IEEE Trans. Commun.*, vol. 58, no. 6, pp. 1737–1747, Jun. 2010.
[24] K. Kittiwaytang, P. Chanloha, and C. Aswakul, "CTM-based reinforcement learning strategy for optimal heterogeneous wireless network selection," in *Proc. 2nd Int. Conf. Comput. Intell., Model. Simul.*, Sep. 2010, pp. 73–78.
[25] M. E. Helou, M. Ibrahim, S. Lahoud, and K. Khawam, "Radio access selection approaches in heterogeneous wireless networks," in *Proc. IEEE 9th Int. Conf. Wireless Mobile Comput., Netw. Commun.*, Oct. 2013, pp. 521–528.
[26] O.-C. Iacoboaiea, B. Sayrac, S. B. Jemaa, and P. Bianchi, "SON coordination in heterogeneous networks: A reinforcement learning framework," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5835–5847, Sep. 2016.
[27] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and C.-C. Hsu, "Performance evaluation of radio access network overloading from machine type communications in LTE-A networks," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops*, Apr. 2012, pp. 248–252.
[28] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A. C.-C. Hsu, "Overload control for machine-type-communications in LTE-advanced system," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 38–45, Jun. 2012.
[29] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*. Hoboken, NJ, USA: Wiley, 1995.
[30] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[31] S.-M. Cheng, W. C. Ao, F.-M. Tseng, and K.-C. Chen, "Design and analysis of downlink spectrum sharing in two-tier cognitive femto networks," *IEEE Trans. Veh. Technol.*, vol. 61, no. 5, pp. 2194–2207, Jun. 2012.
[32] "Open cell ID." [Online]. Available: http://opencellid.org/. Accessed on: Jun. 1, 2016.
[33] A. Chiumento, C. Desset, S. Pollin, L. V. der Perre, and R. Lauwereins, "Impact of CSI feedback strategies on LTE downlink and reinforcement learning solutions for optimal allocation," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 550–562, Jan. 2017.



Yu-Jui Liu received the B.S. degree in computer science from the National Chiao Tung University, Hsinchu, Taiwan, in 2013, and the M.S. degree in computer science and information engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2016. His research interests include machine type communication and machine learning.

Shin-Ming Cheng (S'05–M'07) received the B.S. and Ph.D. degrees in computer science and information engineering from the National Taiwan University, Taipei, Taiwan, in 2000 and 2007, respectively.

He was a Postdoctoral Research fellow at the Graduate Institute of Communication Engineering, National Taiwan University, from 2007 to 2012. Since 2012, he has been in the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, as an Assistant Professor. His current research interests include mobile networks, wireless communication, cyber security, and complex networks. He received the IEEE PIMRC 2013 Best Paper Award and the 2014 ACM Taipei/Taiwan Chapter K. T. Li Young Researcher Award.

Yu-Lin Hsueh received the B.S. degree in computer science and information engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 2016. He is currently working toward the M.S. degree in computer science and information engineering at the National Taiwan University of Science and Technology, Taipei, Taiwan. His research interest focuses on LTE networks.