Modeling and Performance Analysis of a Tracking-Area-List-Based Location Management Scheme in LTE Networks

Tao Deng, *Student Member, IEEE*, Xian Wang, *Member, IEEE*, Pingzhi Fan, *Fellow, IEEE*, and Keqin Li, *Fellow, IEEE*

***Abstract*—In a cellular network, the network coverage is partitioned into nonoverlapping areas called registration areas to facilitate location management (LM) that keeps track of the location of a user equipment (UE) unit. In the second-generation Global System for Mobile Communications (GSM) and the third-generation Universal Mobile Telecommunications System (UMTS), a registration area is called a location area (LA) in the circuit-switched domain or a routing area (RA) in the packet-switched domain. In the fourth-generation Long-Term Evolution (LTE), a registration area is called a tracking area (TA). To tackle the drawbacks of the LA/RA-based LM scheme used in GSM and UMTS, which is intrinsically a static scheme, LTE adopts a dynamic LM scheme called a TA list (TAL)-based scheme. Under the TAL-based scheme, the network allocates to a UE unit a group of TAs referred to as a TAL instead of only a single TA. The UE can move freely within the allocated TAL without performing any location updates that are called TA updates (TAUs) in LTE. Only when moving into a TA that is not included in the allocated TAL does the UE need to perform a TAU, and afterward, the network will allocate, to the UE, a new TAL. The performance of the TAL-based scheme relies on the allocated TAL. If the allocated TAL is inappropriate with respect to the UE's mobility and traffic characteristics, the TAL-based scheme may produce adverse effects. To find an optimal TAL allocation method for global UE units that exhibit weak regularity in their mobility, this paper develops an embedded Markov chain approach to analyze the signaling cost of the TAL-based scheme. This paper distinguishes itself from existing studies in the following aspects. First, this paper follows LTE technical specifications, that is, the number of TAs in a TAL can vary and, furthermore, is upper bounded, whereas existing studies partially or completely ignore this stipulation. Second, this paper emphasizes the impact of a call handling model that dictates whether a TAU occurs after the completion of a call by considering two call handling models, i.e., a call plus location update and a call without location update. Third, as for the dependence among the cell residence time, the TA residence time, and the TAL residence time, this paper proposes the use of a fluid flow model to describe this dependence, which is simple but does not compromise accuracy. Analytical formulas for the total signaling cost of the TAL-based scheme due to TAU and paging operations, as well as formulas that are useful in designing an optimal paging scheme, are derived, and their accuracy is validated through Mont Carlo simulation. Then, numerical studies are carried out to investigate the impact of diverse parameters on the signaling cost, revealing that the network can allocate, to a UE unit, an optimal TAL that is consistent with the UE's mobility and traffic characteristics to minimize the signaling cost of the TAL-based scheme.**

***Index Terms*—Call-handling model, fluid flow model, location update (LU), Long-Term Evolution (LTE), tracking area list (TAL).**

I. Introduction

A. Motivation

IN a cellular network, the current location of a user equipment (UE) unit is constantly tracked by the network for the successful delivery of incoming calls to the UE, which is in charge of location management (LM). LM includes two operations, i.e., location update (LU) and terminal paging. LU is the process through which a UE unit dynamically reports its current location to the network so that the location database in the network can update the current location information of the UE. Paging is the process through which the network searches a called UE unit by broadcasting a paging message in a paging area when an incoming call arrives.

To facilitate LM, the coverage of the network is partitioned into nonoverlapping registration areas. Each registration area consists of a group of cells. In the second-generation (2G) Global System for Mobile Communications (GSM) and the third-generation (3G) Universal Mobile Telecommunications System (UMTS), a registration area is called a location area (LA) in the circuit-switched domain or a routing area (RA) in the packet-switched domain. In the fourth-generation Long-Term Evolution (LTE), a registration area is called a tracking area (TA). A 2G/3G network adopts an LA/RA-based LM scheme. Under this scheme, when entering a new LA/RA, the UE needs to perform an LU. When an incoming call arrives, all the cells in the LA/RA simultaneously page the UE. The LA- and RA-based schemes are essentially two static LM schemes, because all of the UE units have paging areas of the same size. These static schemes exhibit the following drawbacks. 1) These schemes are cost ineffective because they neglect the mobility

Manuscript received March 9, 2015; revised July 11, 2015; accepted August 19, 2015. Date of publication September 1, 2015; date of current version August 11, 2016. This work was supported in part by the National 973 Program under Grant 2012CB316100, by the National Natural Science Foundation of China under Grant 61171096, by the 111 Project under Grant 111-2-14, by the Sichuan Provincial Youth Science and Technology Fund under Grant 2014JQ0041, and by the Fundamental Research Funds for the Central Universities under Grant SWJTU12ZT02/SWJTU12CX093/11CX147/2682014ZT11. The review of this paper was coordinated by Prof. Y. Qian.

T. Deng, X. Wang, and P. Fan are with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China (e-mail: dengtaoswjtu@foxmail.com; drwangxian@gmail.com; p.fan@ieee.org).

K. Li is with the Department of Computer Science, State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/TVT.2015.2473704

and traffic characteristics of individual UE units. It is intuitive that UE units with high mobility should perform LUs less frequently; as in this case, the LU cost dominates the total signaling cost, whereas for UE units with low mobility, frequent LUs are beneficial in reducing the paging cost because in this case, the paging cost dominates the total signaling cost. Therefore, a cost-effective scheme needs to take into account the mobility and traffic characteristics of individual UE units. 2) These schemes will produce a ping-pong effect. Suppose that a UE unit moves back and forth between two neighboring LAs/RAs. In this scenario, the UE will perform repetitive LUs. This phenomenon is called the ping-pong effect. 3) These schemes will bring about the issue of uneven signaling distribution, because all the LUs will occur in the border cells of an LA/RA. When massive UE units (e.g., the UE units on a train), in a short time, pass through the border cells of an LA/RA, in the new cell in which these UE units are moving, a huge number of LU requests will be generated toward the network. This situation may, at times, lead to signaling congestion and affect the quality of service in the cell.

To overcome the drawbacks of static LM schemes, LTE adopts a novel LM scheme, which is called a TA list (TAL)-based LM scheme [1]. Under the TAL-based scheme, the network allocates to the UE a group of TAs instead of only one TA. The group of TAs is referred to as a TAL. The UE can move freely within the allocated TAL without performing any LU that is called a TA update (TAU) in LTE. Only when moving into a TA that is not included in the allocated TAL does the UE need to perform a TAU; after that, the network will allocate to the UE a new TAL. In [1], a central policy is applied for TAL allocation. Under this policy, the TA in which a UE unit performs a TAU is the central TA of the new TAL, indicating that the previous and the new TAL could overlap over some TAs. This policy can reduce the possibility that the UE quickly moves out of the new TAL. To adapt to the mobility and traffic characteristics of individual UE units, the network may allocate to different UE units different TALs. When an incoming call to the UE arrives, the UE is paged in all the cells within a TAL. Thus, the total signaling cost due to the TAU and paging operations can be abated by allocating a reasonable TAL to the UE. As for the number of TAs in a TAL, the works in [1] and [2] specify that it cannot exceed 16. Compared with the static LM scheme, the TAL-based scheme exhibits the following advantages. 1) This scheme is cost effective because the allocated TAL is per-UE based. 2) This scheme can avoid the ping-pong effect through adopting the central TAL allocation policy. 3) This scheme can solve the issue of uneven signaling distribution through providing different UE units with different TALs. From the given narrations, the performance of the TAL-based scheme is determined by the allocated TAL. If the allocated TAL is inappropriate to the UE's mobility and traffic characteristics, the TAL-based scheme may generate adverse effects. This paper aims to solve this issue.

B. Literature Review

The TAL-based scheme has been studied in the literature [3]–[16]. Based on the number of TAs in a TAL, the number of cells in a TA, and the study scenario, existing studies can be categorized into four groups as follows.

The first group consists of studies assuming that a TAL comprises only one TA, such as in [3]–[5]. In [3], Wu *et al.* proposed a distributed cell-based LA planning algorithm to reduce the cost of LM. In this algorithm, when deciding which cells belong to the same LA, the network has to collect information on the massive movement tracks of UE units. However, this information is unavailable to the network or too costly to collect. In [4], Toril *et al.* introduced a graph-partitioning algorithm to analyze the cost of LM. In [5], Razavi *et al.* developed a Pareto optimization approach to achieve a tradeoff between the total signaling cost of the network and TA reconfiguration cost. This tradeoff can be treated as a biobjective optimization problem, and a genetic algorithm (GA) was proposed to obtain the Pareto-optimal solutions. The studies in [3]–[5] considered the situation that a TAL includes one TA. Thus, this TAL allocation scheme is similar to a static LM scheme.

The second group is composed of studies assuming that each TA comprises only one cell. The works in [6]–[10] fall into this group. Using the concept of a movement-based LU (MBLU) scheme, in [6], an optimal movement threshold for TAL allocation is derived, based on the mobility and traffic characteristics of individual UE units. Adopting the notion of a distance-based LU (DBLU) scheme, in [7], a self-organized TAL allocation mechanism is developed to reduce the signaling cost and derived an optimal distance threshold for TAL allocation. However, this mechanism needs to introduce two extra timers to increase the system complexity. In addition, in [6] and [7], it is considered that a TAL is a concentric ring consisting of TAs. In this case, when the optimal movement/distance threshold is large, the number of TAs in a TAL may violate the 16-TA upper bound. In [8] and [9], Razavi *et al.* proposed a local search algorithm and a rule-of-thumb algorithm for TAL allocation to reduce the signaling cost in LTE. However, the two algorithms have the following defects. First, they assumed that a cell assigns one common TAL to all the UE units performing TAUs in this cell. Thus, this TAL allocation is not cost effective. Second, they cannot guarantee the reach of a global optimum, because each cell considers only its own signaling cost by neglecting the impact on neighbor cells. Third, the cells need to frequently transmit TAL reconfiguration information to all the UE units, resulting in expensive signaling cost.

The third group consists of studies considering 1-D study scenario, as in [11] and [12]. In static LA/RA-based schemes, when a large number of UE units cross an LA boundary, e.g., the case in train movement, there is a risk of LU signaling storm in a border cell due to a large amount of LU signaling cost generated in a short time. To solve this risk, in [11], an overlapping TAL method is introduced to mitigate the uplink signaling congestion. Through this method, the UE units on a train are allocated with different TALs so that the risk of LU signaling storm can be avoided. In [12], Liou *et al.* investigated the performance of the TAL-based scheme with three serial paging schemes using a 1-D random-walk model. Simulation and analytical results manifested that the performance of the TAL-based scheme always overmatches that of static schemes. When the cell residence times and call-to-mobility ratios are fixed, an optimal serial paging scheme can be chosen from the three schemes. Under the optimal serial paging schemes, when an incoming call arrives, the network may first page the last interacted cell in which a UE unit performs a TAU or makes a call. However, in commercial LTE scenarios, the network bearer context has no information on the last interacted cell [1]. Therefore, the optimal serial paging scheme is difficult to implement in LTE. Moreover, in the performance evaluation, when the number of cells in a TA is small, the number of TAs in a TAL is large. In this case, the number of TAs in a TAL may not meet the 16-TA upper bound.

The fourth group consists of studies assuming that a TAL comprises multiple TAs, each of which is composed of a group of cells, as in [13]–[16]. In [13], Paramo *et al.* proposed an X2-interface-based approach to reduce the paging cost through employing a location anchor to keep track of the location of UE units inside the TAL. Although this approach can reduce the paging cost, it will increase the system complexity and generate extra signaling cost. Moreover, this work introduced a discrete-time Markov chain with aggregated states to calculate the expected number of TAUs after a certain number of movements. In this introduced Markov chain method, each state represents a cell. However, this method is appropriate only for a situation where all the cells in a TAL are symmetric to each other, so that it can be applied only to some special cases. In more general situations, the mathematical analysis may be infeasible. In [14], Liou and Lin compared the performance of the TAL-based scheme with those of the DBLU and MBLU schemes through simulation. Simulation results proved that in many cases, the TAL-based scheme is better than the DBLU and MBLU schemes. Based on the overlapping TAL method introduced in [11], in [15], the research scenario is expanded from a 1-D scenario to arbitrary scenarios, and how to mitigate the downlink signaling congestion is investigated further. In [11] and [15], the total signaling cost is from the network perspective, which is different from other studies where the total signaling cost is from the UE perspective. According to the activity range of UE units, the UE units can be classified into two types, i.e., local and global. Local UE units represent those that are local dwellers whose mobility displays strong regularity [16]. For instance, a local UE unit may display the following living behavior. A user carrying a UE unit goes to work from his/her home in the morning and has lunch in another place at noon and then returns home in the evening. In contrast, global UE units represent those that are not local dwellers whose mobility exhibits weak regularity. For instance, a roaming user could visit any place of the current roaming area. In [16], Wang *et al.* developed an embedded Markov chain approach to analyze the total signaling cost of the TAL-based scheme for local UE units.

In summary, the existing studies mainly have the following problems. 1) Some studies assumed that a cell allocates a common TAL to individual UE units. Thus, this common TAL may not be cost effective for all the UE units. 2) To simplify the analyzing processes, some studies assumed that a TA includes only one cell, which is unreasonable, because a TA generally comprises multiple cells. 3) All the existing studies neglected the rule that the number of TAs in a TAL cannot exceed 16. 4) There lacks an accurate mathematical approach to calculate the signaling cost of the TAL-based scheme for global UE units. The purpose of this paper is to develop a comprehensive mathematical approach to address all these problems.

C. Our Contributions

Specifically, different from existing studies, the main contributions of this paper are as follows.

- This paper considers that each TA is a concentric ring consisting of a variable number of cells and observes the stipulation by LTE technical specifications, i.e., the number of TAs in a TAL can vary and, furthermore, is upper bounded.
- This paper emphasizes the impact of call handling models on the total signaling cost by taking into account two different call handling models, i.e., a call plus location update (CPLU) and a call without location update (CWLU) [17]. To simplify the mathematical analysis, most existing studies about LM consider only the CPLU model. However, the truth is that existing cellular networks employ only the CWLU model. In the performance evaluation, the impact of the two different call handling models on the total signaling cost is investigated.
- For mathematical analysis, this paper emphasizes the inherent dependence among the cell residence time, the TA residence time, and the TAL residence time through using a fluid flow model. The fluid flow model makes the mathematical modeling of complex problems such as the TAL-based LU scheme feasible, can greatly simplify the associated mathematical manipulation, and, at the same time, does not affect the accuracy of the results.
- To seek an optimal TAL allocation for global UE units, this paper develops an embedded Markov chain approach to analyze the total signaling cost of the TAL-based LM scheme under the two call handling models. To exemplify the modeling of the TAL-based scheme, we first consider a special case where the number of TAs in a TAL is 16. Certainly, we can also choose other cases in accordance with the 16-TA upper bound, because the number of TAs in a TAL is not an impacting factor in the modeling approach. Previously, the embedded Markov approach has been used to analyze LM in [18] and [19]. Analytical formulas for the total signaling cost due to TAU and paging operations, as well as formulas useful that are in designing an optimal paging scheme, are derived. The accuracy of these formulas is tested through simulation. With these formulas, the impact of diverse parameters on the total signaling cost is evaluated through numerical studies.
- Based on the developed approach, this paper further investigates the impact of the number of TAs in a TAL and the number of cells in a TA on the total signaling cost through numerical studies. With these numerical results, when the number of cells in a TA is fixed, based on the mobility and traffic characteristics of individual UE units, the network can allocate an optimal TAL to the UE that can minimize the total signaling cost.

The rest of this paper is organized as follows. Section II introduces the structures of a cell, a TA, and a TAL and talks

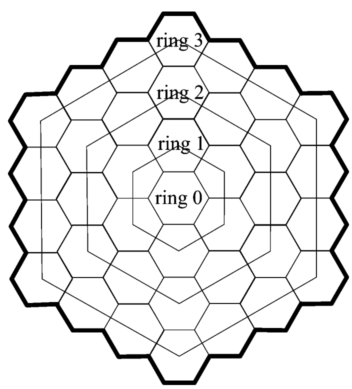


Fig. 1. TA structure with a radius of 4.

over the fluid flow model, traffic model, and the call handling models. Section III develops an embedded Markov chain approach to analyze the signaling cost of the TAL-based scheme under the CWLU call handling model when the number of TAs in a TAL is 16. Section IV does so under the CPLU model. Section V checks the accuracy of the derived formulas through simulations and conducts numerical studies to investigate the impact of various parameters on the total signaling cost. Section VI concludes this paper.

II. System Model

A. Cell, TA, and TAL Structures

To design a mathematical approach to analyze the signaling cost of the TAL-based LM scheme, the structures of a cell, a TA, and a TAL should be assumed.

Cell structure: Assume that all the cells in a network are regular hexagons of the same size so that each cell has six neighbor cells.

TA structure: Assume that all the TAs in the network have the same ring structure consisting of cells. Denote by d the radius of this structure. A TA consists of d rings of cells labeled as ring r's, $r = 0, 1, \ldots, d-1$. Ring 0 is the center cell. Ring r, $r = 0, 1, \ldots, d-1$, is composed of all cells that are r cells away from the center cell. Apparently, ring r, $r = 1, 2, \ldots, d-1$, has $6r$ cells so that the total number of cells in a TA is $3d^2 - 3d + 1$. Fig. 1 shows a TA with $d = 4$.

TAL structure: Denote by N_{TA} the number of TAs in a TAL. This paper considers that a TAL is constructed by TAs in concentric cycles around a center TA. In this TAL structure, all the TAs are identical. Ring 0 is the center TA in which a TAU occurs. When $1 < N_{\mathrm{TA}} \le 7$, ring 0 is surrounded by ring 1 that has $(N_{\mathrm{TA}} - 1)$ TAs. Since $N_{\mathrm{TA}} \le 16$, a TAL can have, at most, three rings of TAs. When $7 < N_{\mathrm{TA}} \le 16$, ring 1 is, in turn, surrounded by ring 2 that has $(N_{\mathrm{TA}} - 7)$ TAs. Fig. 2 shows a TAL with $N_{\mathrm{TA}} = 16$. In Fig. 2, label a TA in a TAL as "$\mathrm{TA}_{x,y}$," where x is the ring index, and y is the TA index, and subscripts x and y satisfy

$$y = \begin{cases} 0, & \text{if } x = 0 \\ 0, 1, \ldots, 5, & \text{if } x = 1 \\ 0, 1, \ldots, 8, & \text{if } x = 2. \end{cases} \qquad (1)$$

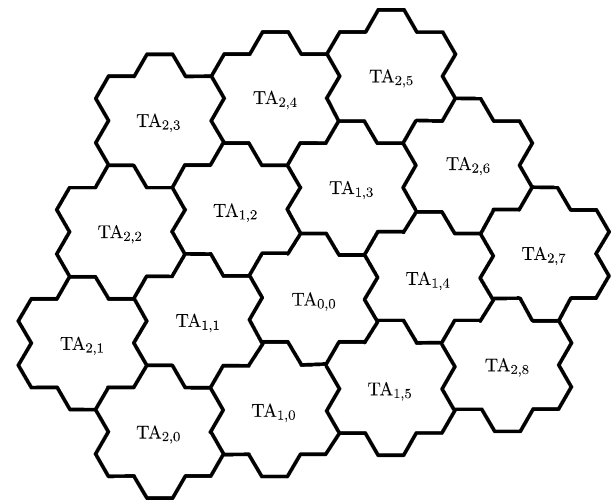


Fig. 2. TAL structure.

$\mathrm{TA}_{0,0}$ is the TA where the UE performs the last TAU. To facilitate mathematical analysis, the border TAs in a TAL can be divided into two types, vertex TAs and inner TAs [7]. A TA that has four/three neighbor TAs in the same TAL is called an inner/vertex TA. In Fig. 2, $\mathrm{TA}_{x,y}$'s, $x, y \in \{(2,0); (2,1); (2,3); (2,5); (2,7); (2,8)\}$ are vertex TAs, whereas $\mathrm{TA}_{x,y}$'s, $x, y \in \{(1,0); (1,5); (2,2); (2,4); (2,6)\}$ are inner TAs. Fig. 2 shows a TAL configuration, which is referred to as configuration 1, in which two vertex TAs and one inner TA are removed from the complete ring 2 that has six vertex TAs and six inner TAs. In fact, there is another possible configuration, which is referred to as configuration 2, in which two inner TAs and one vertex TA are removed from the complete ring 2. Section V-E will compare the signaling costs of the two configurations. This paper considers only a situation where all the removed TAs are adjacent. For nonadjacent situations, the corresponding mathematical analyses are similar.

B. Random-Walk Model

The design and analysis of any LM scheme depend on a selected mobility model. Random-walk models have been used to describe the UE movements in [12] and [20]–[24]. Zhou *et al.* in [20] used a random-walk model to describe node movements in mobile ad hoc networks. In this model, each node randomly and independently selects a movement direction and a velocity at the beginning of each time slot and walks along this direction at this velocity during the rest of the time slot. Afterward, the movement direction and velocity are reselected. The works in [12] and [22] adopted a 1-D random-walk model to describe UE movements in a 1-D cellular network, where one cell has only two neighbor cells. The works in [23] and [24] employed a 2-D random-walk model to describe UE movements in a 2-D cellular network. In this 2-D model, the cell residence time follows an independent and identical distribution. After residing in a cell for some time, a UE unit moves into the six neighbor cells with the same probability, i.e., 1/6 for a hexagonal cell structure. This paper adopts the 2-D random-walk model to describe UE movements among cells. The 2-D random-walk models used in [20] and this paper are different in terms of the timing when the movement direction and velocity can change. In [20], the two variables remain unchanged during a time slot, whereas in this paper, they can change at any time. In this case, as for UE movements among TAs, after leaving its current TA, the UE moves into the six neighbor TAs with the same probability of 1/6. As an inner TA has four neighbor TAs in the same TAL, after leaving an inner TA, the UE launches a TAU with probability 1/3. Similarly, when moving out of a vertex TA, this probability changes to 1/2.

The 2-D random-walk model can be accurately described by a 2-D Markov chain with a space size of $N_{\rm TA}(3d^2 - 3d + 1)$. When $N_{\rm TA}$ and d are large, the computational complexity of the Markov chain will be very high. Akyildiz *et al.* in [23] proposed an approach to reduce the computational complexity. The proposed approach can reduce the space size of the 2-D Markov chain from $N_{\rm TA}(3d^2 - 3d + 1)$ to about $(1/6)N_{\rm TA}(3d^2 - 3d + 1)$. To further decrease the space size, 1-D Markov chains were proposed to describe the random-walk model [17]. The space size of the 1-D Markov chain is $dN_{\rm TA}$. However, regardless of a 1-D or a 2-D Markov chain, it is difficult to tackle the mathematical analysis for the TAL-based scheme that has to consider the relation among a cell, a TA, and a TAL. To this end, this paper adopts a fluid flow model to characterize the dependence among the cell residence time, the TA residence time, and the TAL residence time for mathematical analysis. In simulation, this paper employs a 2-D random-walk model. Denote by $S_{d,d}$ the expected number of steps required to walk out a TA after entering the TA through ring $d - 1$. When the UE movements among cells are characterized by the 2-D Markov chain, $S_{d,d}$ can be given as [25]

$$S_{d,d} = \frac{3d^2 - 3d + 1}{2d - 1}, \quad d = 1, 2, \ldots. \tag{2}$$

C. Fluid Flow Model

A fluid flow model was proposed in [26]–[28] to compute the rate at which a UE unit residing in a closed region moves out of the border of the region. This rate is called the border-crossing rate. In [18] and [29]–[31], the fluid flow model has been used to characterize the residence time in closed regions, e.g., a cell and an LA. To apply the fluid flow model to describe the residence time in closed regions, the following two assumptions are necessary. First, the velocities of the UE in different places are independent and identically distributed, and the moving direction of the UE is uniformly distributed over $[0, 2\pi)$. Second, UE units are uniformly distributed throughout the closed region. Denoting by R_b the border-crossing rate, it follows from [18] that

$$R_b = \frac{L\overline{V}}{\pi A} \tag{3}$$

where $\overline{V}$, L, and A denote the average velocity of the UE and the perimeter and area of the closed region, respectively. Denote by $L_s/L_{\rm TA}$ the perimeter of a cell/TA and by $A_s/A_{\rm TA}$ the area of a cell/TA. Denote by R the side length of a hexagon cell. It follows that:

$$\begin{aligned} L_s &= 6R \\ L_{\rm TA} &= 6(2d-1)R \\ A_{\rm TA} &= (3d^2 - 3d + 1)A_s. \end{aligned}$$

Denote by ϵ the probability that after leaving an old cell, the UE moves into a new cell that belongs to the same TA as the old cell. Denote by $t_{\rm TA}$ and t_s the TA residence time and the cell residence time, respectively. Denote by $1/\eta$ the mean of t_s and by $1/\tau$ the mean of $t_{\rm TA}$. It follows that:

$$\tau = \eta(1 - \epsilon). \tag{4}$$

Thus, it follows from (3) and (4) that:

$$\begin{cases} \frac{\tau}{\eta} = \frac{A_s L_{\rm TA}}{L_s A_{\rm TA}} = \frac{2d-1}{3d^2-3d+1} \\ \epsilon = 1 - \frac{\tau}{\eta} = \frac{3d^2-5d+2}{3d^2-3d+1}. \end{cases} \tag{5}$$

That is, after crossing a cell, with probabilities ϵ and $1 - \epsilon$, the UE remains in and moves out of the current TA, respectively. This trick of quantitatively establishing the dependence between the events of cell and TA crossings was originally applied in [18] and [32]. The accuracy of this trick will be validated in Section V-A through simulation. Equation (5) manifests that $S_{d,d}$ defined in Section II-B can be given as [18]

$$S_{d,d} = \frac{1/\tau}{1/\eta} = \frac{3d^2 - 3d + 1}{2d - 1}, \quad d = 1, 2, \ldots$$

which is the same as (2). Therefore, in terms of $S_{d,d}$, the fluid flow model is also very accurate. Let $f_{\rm TA}(t)/f_s(t)$ be the probability density function of $t_{\rm TA}/t_s$. Denote by $f^*(s)$ the Laplace transform of a function $f(t)$, i.e.,

$$f^*(s) \triangleq \int_{t=-\infty}^{+\infty} f(t)e^{-st}dt.$$

This paper considers that t_s follows an exponential distribution with mean $1/\eta$. It follows from [18] that:

$$f^*_{\rm TA}(s) = \sum_{k=1}^{+\infty} \epsilon^{k-1}(1-\epsilon)\left[f^*_s(s)\right]^k = \frac{\eta(1-\epsilon)}{s+\eta(1-\epsilon)} = \frac{\tau}{\tau+s}. \tag{6}$$

That is, if t_s follows an exponential distribution, then $t_{\rm TA}$ must do so.

D. Traffic Model

This paper considers that the incoming calls follow a Poisson process with rate λ. Thus, the call interarrival time, which is denoted by t_c, follows an exponential distribution with mean $1/\lambda$. The Poisson process has been widely used to describe the incoming calls in [7], [12], and [16].

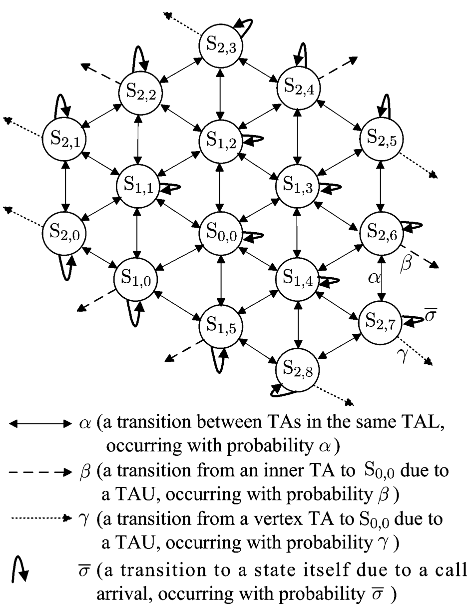


Fig. 3. Embedded Markov chain for the TAL-based scheme under the CWLU model when $N_{\mathrm{TA}} = 16$.

E. Call-Handling Models

A call handling model determines whether a TAU occurs after a call. This paper considers two call handling models, namely, a CPLU model and a CWLU model, which were first proposed by Li in [17]. Under the CPLU model, a TAU occurs after a call, whereas under the CWLU model, no TAU occurs after a call. The call handling models are used to investigate the impact of incoming calls on the total signaling cost of an LM scheme. This paper assumes that the call holding time is negligible. That is, it does not consider the impact of the call holding time. This assumption has been adopted in [12], [17], and [33]. There should be not much difficulty in extending the mathematical approach developed in this paper to take into account the influence of the call holding time.

III. Signaling Cost of the Tracking Area List-Based Location Management Scheme Under the Call Without Location Update Model

A. Embedded Markov Chain

Here, we exemplify the modeling of the TAL-based scheme through using a special case of $N_{\mathrm{TA}} = 16$. For other values of N_{TA}, the modeling is straightforward. The TAL-based scheme can be described by an embedded Markov chain. Fig. 3 shows such a chain for the TAL-based scheme under the CWLU model when $N_{\mathrm{TA}} = 16$. In the Markov chain, state $S_{x,y}$ represents the state in which the UE is currently residing in $\mathrm{TA}_{x,y}$. In particular, state $S_{0,0}$ is the state entered by the Markov chain after the occurrence of a TAU. Denote by $\mathbf{S}$ and S_{size} the state space of the Markov chain and its size, respectively. It follows that

$$\begin{cases} \mathbf{S} = \{S_{0,0}, S_{1,0}, S_{1,1}, S_{1,2}, S_{1,3}, S_{1,4}, S_{1,5}, S_{2,0}, \\ \qquad S_{2,1}, S_{2,2}, S_{2,3}, S_{2,4}, S_{2,5}, S_{2,6}, S_{2,7}, S_{2,8}\} \\ S_{\mathrm{size}} = 16. \end{cases}$$

Next, the transition probabilities of the Markov chain are derived. When the UE is in state $S_{x,y}$, $S_{x,y} \in \mathbf{S}$, there are two events that can cause the chain to leave state $S_{x,y}$, with one event being a call arrival and the other being the departure from $\mathrm{TA}_{x,y}$ into a new TA, e.g., $\mathrm{TA}_{x',y'}$. When the leave from state $S_{x,y}$ is due to a call arrival, after the completion of the call, the Markov chain will return to state $S_{x,y}$, because this paper has assumed that the duration time of a call is negligible, and hence, the UE remains in $\mathrm{TA}_{x,y}$ after the call. The corresponding transition probability from state $S_{x,y}$ due to a call arrival can be expressed as

$$\overline{\sigma} \triangleq \Pr(S_{x,y} \to S_{x,y}) = \Pr(t_c < t_{\mathrm{TA}}) = 1 - \sigma$$

where

$$\sigma \triangleq \Pr(t_c > t_{\mathrm{TA}}) = \frac{\tau}{\tau + \lambda}.$$

In the given equation, the variable t_c may also represent the residual call interarrival time, which is the time between the instant of entering a state and the instant the next call arrives. In view of the memoryless property of exponential distributions, the residual call interarrival time follows the same exponential distribution with the same mean as the call interarrival time. Thus, this paper no longer defines a new variable for the residual call interarrival time. Likewise, t_{TA} has the same characteristic as t_c. When the transition from state $S_{x,y}$ is due to the event of moving into a new TA $\mathrm{TA}_{x',y'}$, there are two different situations. The first situation is that $\mathrm{TA}_{x,y}$ and $\mathrm{TA}_{x',y'}$ are in the same TAL. It follows that the transition probability of this situation is:

$$\begin{aligned} \alpha &\triangleq \Pr(S_{x,y} \to S_{x',y'}, S_{x,y} \text{ and } S_{x',y'} \in \text{ the same TAL}) \\ &= \Pr(t_c > t_{\mathrm{TA}}, \mathrm{TA}_{x,y} \to \mathrm{TA}_{x',y'}) = \frac{\sigma}{6}. \end{aligned} \tag{7}$$

The second situation is that $\mathrm{TA}_{x',y'}$ does not belong to the TAL of $\mathrm{TA}_{x,y}$ so that a TAU will occur, and after the TAU, the Markov chain will transit to state $S_{0,0}$. In this case, this paper refers to $S_{x,y}$ as a border state. When $\mathrm{TA}_{x,y}$ is an inner TA, the transition probability of the second situation is

$$\begin{aligned} \beta &\triangleq \Pr(S_{x,y} \to S_{0,0}, \text{due to a TAU, } \mathrm{TA}_{x,y} \text{ is an inner TA}) \\ &= \Pr(t_c > t_{\mathrm{TA}}, \mathrm{TA}_{x,y} \to \mathrm{TA}_{x',y'}) = \frac{\sigma}{3}. \end{aligned} \tag{8}$$

Similarly, when $\mathrm{TA}_{x,y}$ is a vertex TA, the transition probability of the second situation is

$$\gamma \triangleq \Pr(S_{x,y} \to S_{0,0}, \text{due to a TAU, } \mathrm{TA}_{x,y} \text{ is a vertex TA}) = \frac{\sigma}{2}. \tag{9}$$

Denote by $\mathbf{P}$ the transition probability matrix of the embedded Markov chain, which is defined in the equation shown at the bottom of the next page.

Denote by Π_θ, $\theta \in \mathbf{S}$, the equilibrium probability of state θ. Denote by $\mathbf{\Pi}$ a row vector consisting of Π_θ, and the order of the entries in $\mathbf{\Pi}$ is the same as the order of the entries in $\mathbf{S}$.

The balance equations of the embedded Markov chain can be given as

$$\mathbf{\Pi P} = \mathbf{\Pi}. \tag{10}$$

Equation (10) combined with the following normalization condition:

$$\sum_{\theta \in \mathbf{S}} \Pi_\theta = 1$$

can yield a solution of $\mathbf{\Pi}$.

Denote by t_θ, $\theta \in \mathbf{S}$, the residence time in state θ. As aforementioned, a state transition is due to a call arrival or the crossing of a TA. When it is due to a call arrival, we have $t_c < t_{\mathrm{TA}}$ and $t_\theta = t_c$. When it is due to the crossing of a TA, we have $t_c > t_{\mathrm{TA}}$ and $t_\theta = t_{\mathrm{TA}}$. Therefore, it follows that:

$$t_\theta = \min(t_c, t_{\mathrm{TA}}), \theta \in \mathbf{S}.$$

Denote by T the expected time between two consecutive state transitions. It follows that:

$$T = \sum_{\theta \in \mathbf{S}} \Pi_\theta E(t_\theta) = \frac{1}{\tau + \lambda}. \tag{11}$$

Denote by $N_{\mathrm{TAU,ST}}$ the expected number of TAUs triggered by a state transition. It follows that:

$$\begin{aligned} N_{\mathrm{TAU,ST}} = \beta &\left(\Pi_{S_{1,0}} + \Pi_{S_{1,5}} + \Pi_{S_{2,2}} + \Pi_{S_{2,4}} + \Pi_{S_{2,6}}\right) \\ &+ \gamma \left(\Pi_{S_{2,0}} + \Pi_{S_{2,1}} + \Pi_{S_{2,3}} + \Pi_{S_{2,5}} + \Pi_{S_{2,7}} + \Pi_{S_{2,8}}\right). \end{aligned}$$

Denote by $N_{\mathrm{TAU,UT}}$ the expected number of TAUs occurring per unit time. Since the $N_{\mathrm{TAU,ST}}$ TAUs occur during time T, we have

$$N_{\mathrm{TAU,UT}} = \frac{N_{\mathrm{TAU,ST}}}{T} = (\lambda + \tau) N_{\mathrm{TAU,ST}}.$$

Denote by $N_{\mathrm{TAU,CWLU}}$ the expected number of TAUs of the TAL-based scheme under the CWLU model during the call interarrival time t_c. It follows that

$$N_{\mathrm{TAU,CWLU}} = N_{\mathrm{TAU,UT}} E(t_c) = \frac{\lambda + \tau}{\lambda} N_{\mathrm{TAU,ST}}.$$

B. Paging Scheme

Define a paging area as an area where the UE can be located when an incoming call arrives. There are two types of paging schemes, i.e., the parallel paging scheme and the sequential paging scheme. Under the parallel paging scheme, all the cells in the paging area are simultaneously paged. Under a sequential paging scheme, the paging area is partitioned into disjoint subareas, and these subareas are paged one by one until the UE is located. In general, the choice of paging schemes depends on the traffic characteristics and paging cost. For delay-sensitive traffic, the parallel paging scheme is the best choice, because it is easy to be realized and shortens the paging delay at the price of paging cost. If paging cost is an important factor and traffic can tolerate certain delay, we can choose a sequential paging scheme. Denote by $P(x, y)$ the probability that upon a call arrival, the UE resides in $\mathrm{TA}_{x,y}$. It follows that:

$$P(x, y) = \frac{\Pi_{S_{x,y}} E\left(t_{S_{x,y}}\right)}{\sum_{\theta \in \mathbf{S}} \Pi_\theta E(t_\theta)} = \frac{\Pi_{S_{x,y}}}{\sum_{\theta \in \mathbf{S}} \Pi_\theta}. \tag{12}$$

In [33], Zhu and Leung developed the necessary condition for a sequential paging scheme to be optimal and proposed a GA to search for an optimal sequential paging scheme that can guarantee the paging delay and, at the same time, minimize the paging cost. The distribution of $P(x, y)$ given in (12), combined with the algorithm proposed in [33], can lead to an optimal sequential paging scheme. For the purpose of demonstration, here, this paper considers only the parallel paging scheme. Denote by N_{paging} the number of cells paged to locate the UE. It follows that:

$$N_{\mathrm{paging}} = 16(3d^2 - 3d + 1).$$

$$\mathbf{P} = \begin{array}{c|cccccccccccccccc|} & S_{0,0} & S_{1,0} & S_{1,1} & S_{1,2} & S_{1,3} & S_{1,4} & S_{1,5} & S_{2,0} & S_{2,1} & S_{2,2} & S_{2,3} & S_{2,4} & S_{2,5} & S_{2,6} & S_{2,7} & S_{2,8} \\ S_{0,0} & \overline{\sigma} & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ S_{1,0} & \gamma & \overline{\sigma} & \alpha & 0 & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ S_{1,1} & \alpha & \alpha & \overline{\sigma} & \alpha & 0 & 0 & 0 & \alpha & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 \\ S_{1,2} & \alpha & 0 & \alpha & \overline{\sigma} & \alpha & 0 & 0 & 0 & 0 & \alpha & \alpha & \alpha & 0 & 0 & 0 & 0 \\ S_{1,3} & \alpha & 0 & 0 & \alpha & \overline{\sigma} & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & \alpha & \alpha & 0 & 0 \\ S_{1,4} & \alpha & 0 & 0 & 0 & \alpha & \overline{\sigma} & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \alpha & \alpha \\ S_{1,5} & \gamma & \alpha & 0 & 0 & 0 & \alpha & \overline{\sigma} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha \\ S_{2,0} & \gamma & \alpha & \alpha & 0 & 0 & 0 & 0 & \overline{\sigma} & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ S_{2,1} & \gamma & 0 & \alpha & 0 & 0 & 0 & 0 & \alpha & \overline{\sigma} & \alpha & 0 & 0 & 0 & 0 & 0 & 0 \\ S_{2,2} & \beta & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & \alpha & \overline{\sigma} & \alpha & 0 & 0 & 0 & 0 & 0 \\ S_{2,3} & \gamma & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & \overline{\sigma} & \alpha & 0 & 0 & 0 & 0 \\ S_{2,4} & \beta & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & \overline{\sigma} & \alpha & 0 & 0 & 0 \\ S_{2,5} & \gamma & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \overline{\sigma} & \alpha & 0 & 0 \\ S_{2,6} & \beta & 0 & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \overline{\sigma} & \alpha & 0 \\ S_{2,7} & \gamma & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \overline{\sigma} & \alpha \\ S_{2,8} & \gamma & 0 & 0 & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \overline{\sigma} \end{array}_{16 \times 16}$$

C. Total Signaling Cost

Denote by C_{CWLU} the total signaling cost of the TAL-based scheme under the CWLU model due to the TAU and paging operations during t_c. Denote by δ_{TAU} the signaling cost of performing a TAU and by δ_{paging} the signaling cost of performing a paging in a cell. It follows that:

$$C_{\mathrm{CWLU}} = \frac{\lambda+\tau}{\lambda} N_{\mathrm{TAU,ST}}\delta_{\mathrm{TAU}} + 16(3d^2-3d+1)\delta_{\mathrm{paging}}. \quad (13)$$

IV. Signaling Cost of the Tracking Area List-Based Location Management Scheme Under the Call Plus Location Update Model

A. Embedded Markov Chain

Here, we inherit the notations defined in Section III. Fig. 4 shows an embedded Markov chain for the TAL-based scheme under the CPLU model when $N_{\mathrm{TA}} = 16$. As in the case of the CWLU model, under the CPLU model, there are also two events that can trigger a transition from state $S_{x,y}$, i.e., the departure from $\mathrm{TA}_{x,y}$ and the arrival of an incoming call. When the transition from $S_{x,y}$ is due to the departure from $\mathrm{TA}_{x,y}$, the corresponding transition probabilities have been given in (7)–(9). When the transition from $S_{x,y}$ is due to a call arrival, as per the CPLU model, the UE needs to perform a TAU, so that in this case, after leaving $S_{x,y}$, the Markov chain will transit to state $S_{0,0}$. The corresponding transition probability from $S_{x,y}$ to $S_{0,0}$ due to a call arrival can be expressed as

$$\Pr(S_{x,y} \to S_{0,0}) = \Pr(t_c < t_{\mathrm{TA}}) = \overline{\sigma}.$$

Therefore, we reach the matrix of transition probabilities $\mathbf{P}$, which is shown at the bottom of the page. The set of balance equations can be given as

$$\mathbf{\Pi P} = \mathbf{\Pi}.$$

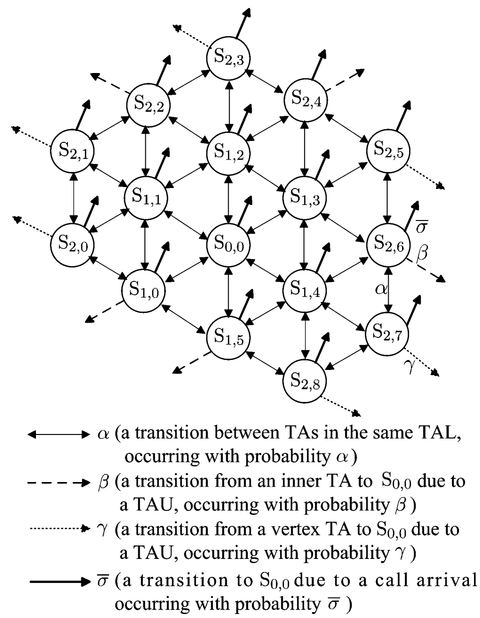


Fig. 4. Embedded Markov chain for the TAL-based scheme under the CPLU model when $N_{\mathrm{TA}} = 16$.

The given equation combined with the following normalization condition:

$$\sum_{\theta \in \mathbf{S}} \Pi_\theta = 1$$

can yield a solution of $\mathbf{\Pi}$.

B. Total Signaling Cost

Under the CPLU model, there are two types of TAUs, i.e., those occurring during the call interarrival time t_c and those occurring after a call. To avoid ambiguity, this paper refers to a TAU occurring after a call as TAU′, and hereafter, in

$$\mathbf{P} = \begin{array}{c|cccccccccccccccc|}
 & S_{0,0} & S_{1,0} & S_{1,1} & S_{1,2} & S_{1,3} & S_{1,4} & S_{1,5} & S_{2,0} & S_{2,1} & S_{2,2} & S_{2,3} & S_{2,4} & S_{2,5} & S_{2,6} & S_{2,7} & S_{2,8} \\
S_{0,0} & \overline{\sigma} & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
S_{1,0} & \overline{\sigma}+\gamma & 0 & \alpha & 0 & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
S_{1,1} & \overline{\sigma}+\alpha & \alpha & 0 & \alpha & 0 & 0 & 0 & \alpha & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 \\
S_{1,2} & \overline{\sigma}+\alpha & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 & \alpha & \alpha & \alpha & 0 & 0 & 0 & 0 \\
S_{1,3} & \overline{\sigma}+\alpha & 0 & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & \alpha & \alpha & 0 & 0 \\
S_{1,4} & \overline{\sigma}+\alpha & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \alpha & \alpha \\
S_{1,5} & \overline{\sigma}+\gamma & \alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha \\
S_{2,0} & \overline{\sigma}+\gamma & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
S_{2,1} & \overline{\sigma}+\gamma & 0 & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 \\
S_{2,2} & \overline{\sigma}+\beta & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 & 0 \\
S_{2,3} & \overline{\sigma}+\gamma & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 \\
S_{2,4} & \overline{\sigma}+\beta & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 \\
S_{2,5} & \overline{\sigma}+\gamma & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 & 0 \\
S_{2,6} & \overline{\sigma}+\beta & 0 & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 \\
S_{2,7} & \overline{\sigma}+\gamma & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha \\
S_{2,8} & \overline{\sigma}+\gamma & 0 & 0 & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0
\end{array}_{16\times 16}$$

TABLE I
COMPARISON BETWEEN THE SIMULATION AND ANALYTICAL RESULTS OF $P(x, y)$ WHEN $d = 3$

(x, y)		(0,0)	(1,0)	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(2,0)	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)	(2,8)
CWLU	analytical(10^{-2})	24.61	7.17	8.25	8.59	8.59	8.25	7.17	2.98	2.48	3.66	2.66	3.75	2.66	3.66	2.48	2.98
	simulation(10^{-2})	24.62	7.21	8.22	8.61	8.58	8.21	7.20	2.96	2.45	3.68	2.67	3.76	2.67	3.64	2.47	2.96
	rel.err.(%)	0.03	0.66	0.31	0.28	0.02	0.41	0.49	0.56	1.16	0.38	0.35	0.25	0.22	0.55	0.25	0.63
CPLU	analytical(10^{-2})	26.64	7.20	8.17	8.46	8.46	8.17	7.20	2.84	2.32	3.44	2.47	3.51	2.47	3.44	2.32	2.84
	simulation(10^{-2})	26.67	7.18	8.13	8.44	8.54	8.17	7.23	2.82	2.30	3.41	2.48	3.55	2.48	3.42	2.29	2.83
	rel.err.(%)	0.10	0.22	0.50	0.22	0.92	0.02	0.43	0.51	0.88	0.93	0.29	1.12	0.31	0.62	1.15	0.20

this paper, by TAU, we mean a TAU occurring during t_c. Denote by $\delta_{\mathrm{TAU}'}$ the signaling cost of performing a TAU$'$. Generally, $\delta_{\mathrm{TAU}} > \delta_{\mathrm{TAU}'}$, because after the completion of a call, a communication link between the UE and the network already exists so that the communication between the UE and the network required to process the TAU can be transmitted over this link. By contrast, during t_c, the UE stays in idle state so that there is no communication link between the UE and the network. The UE must first go through some procedure to establish a communication link to process a TAU, and this procedure will incur signaling cost. The expected number of TAUs triggered by a state transition can be expressed as

$$N_{\mathrm{TAU,ST}} = \beta\left(\Pi_{S_{1,0}} + \Pi_{S_{1,5}} + \Pi_{S_{2,2}} + \Pi_{S_{2,4}} + \Pi_{S_{2,6}}\right) + \gamma\left(\Pi_{S_{2,0}} + \Pi_{S_{2,1}} + \Pi_{S_{2,3}} + \Pi_{S_{2,5}} + \Pi_{S_{2,7}} + \Pi_{S_{2,8}}\right).$$

Thus, the expected number of TAUs occurring per unit time is

$$N_{\mathrm{TAU,UT}} = \frac{N_{\mathrm{TAU,ST}}}{T} = (\lambda + \tau)N_{\mathrm{TAU,ST}}.$$

It follows that the expected number of TAUs occurring during t_c is:

$$N_{\mathrm{TAU,CPLU}} = N_{\mathrm{TAU,UT}}E(t_c) = \frac{\lambda + \tau}{\lambda}N_{\mathrm{TAU,ST}}.$$

Denote by $N_{\mathrm{TAU}'}$ the number of TAUs occurring during t_c. It follows that:

$$N_{\mathrm{TAU}'} = 1.$$

The number of cells paged to locate the UE remains the same as in the case of the CWLU model. Denote by C_{CPLU} the total signaling cost of the TAL-based scheme under the CPLU model. It follows that:

$$\begin{aligned} C_{\mathrm{CPLU}} &= N_{\mathrm{TAU}'}\delta_{\mathrm{TAU}'} + N_{\mathrm{TAU,CPLU}}\delta_{\mathrm{TAU}} + N_{\mathrm{paging}}\delta_{\mathrm{paging}} \\ &= \delta_{\mathrm{TAU}'} + \frac{\lambda + \tau}{\lambda}N_{\mathrm{TAU,ST}}\delta_{\mathrm{TAU}} \\ &\quad + 16(3d^2 - 3d + 1)\delta_{\mathrm{paging}}. \end{aligned} \tag{14}$$

V. PERFORMANCE EVALUATION

Here, first, in Section V-A, the accuracy of the analytical results is checked through simulation. In Section V-B and C, the impact of various parameters on the total signaling cost of the two call handling models is assessed. In Section V-D, the impact of call handling models on the total signaling cost is investigated. In Section V-E, the effect of the TAL configuration on the expected number of TAUs under the CPLU model is demonstrated.

TABLE II
COMPARISON BETWEEN THE SIMULATION AND ANALYTICAL RESULTS OF $N_{\mathrm{TAU,CWLU}}$ AND $N_{\mathrm{TAU,CPLU}}$ AND $\lambda/\eta = 0.01$ AND $N_{\mathrm{CALL}} = 5 \times 10^5$

d		1	2	3	4	5	6
$N_{\mathrm{TAU,CWLU}}$	analytical	16.5700	7.1013	4.3604	3.1348	2.4447	2.0029
	simulation	16.5890	7.0899	4.3526	3.1293	2.4438	2.0094
	rel.err.(%)	0.1147	0.1650	0.1789	0.1754	0.0368	0.3245
$N_{\mathrm{TAU,CPLU}}$	analytical	16.3800	6.9140	4.1759	2.9530	2.2655	1.8263
	simulation	16.3580	6.9190	4.1849	2.9490	2.2723	1.8420
	rel.err.(%)	0.1343	0.0723	0.2151	0.1355	0.3002	0.8523

A. Accuracy Check of the Analytical Results Through Simulation

Here, the accuracy of the analytical results is checked through Monte Carlo simulation. In the simulations, as explained in Section II-B, this paper uses a 2-D random-walk model to describe UE movements among cells. This paper simulates a certain number of call arrivals, e.g., N_{CALL} call arrivals, to obtain a performance metric. In the simulation, we set $\delta_{\mathrm{paging}} = 1$, $\delta_{\mathrm{TAU}} = 50$, $\delta_{\mathrm{TAU}'} = 30$, $\lambda/\eta = 0.01$, and $N_{\mathrm{CALL}} = 5 \times 10^5$. Since this paper uses the parallel paging scheme, there is no need to check the accuracy of the number of cells paged to locate the UE (i.e., N_{paging}). Table I compares the analytical and simulation results of $P(x, y)$ when $d = 3$. Table II compares the analytical and simulation results of $N_{\mathrm{TAU,CPLU}}$ and $N_{\mathrm{TAU,CWLU}}$. In these tables, entry *rel. err.* represents the relative error of the analytical result to the simulation result. Tables I and II show that the simulation results closely match the analytical results, manifesting the accuracy of the mathematical approach proposed in this paper.

B. Impact of Diverse Parameters on C_{CWLU}

Here, the impact of parameters δ_{TAU}, δ_{paging}, N_{TA}, and λ/η on C_{CWLU} is investigated, and assume that $N_{\mathrm{TA}} = 16$ and $\delta_{\mathrm{paging}} = 1$.

Fig. 5 shows the total signaling cost of the CWLU model as a function of the radius of a TA (i.e., d) when $\lambda/\eta = 0.01$, 0.1, and 1. In Fig. 5, we observe the following. 1) The total signaling cost of the TAL-based scheme is a downward convex function of d. For the purpose of demonstration, here, we only explain for the case of $\lambda/\eta = 0.01$. Fig. 6 shows the total signaling cost, TAU cost (i.e., $N_{\mathrm{TAU,CWLU}}\delta_{\mathrm{TAU}}$), and paging cost (i.e., $N_{\mathrm{paging}}\delta_{\mathrm{paging}}$) of the CWLU model. In Fig. 6, it is shown that the TAU cost is a decreasing function of d, whereas the paging cost is an increasing function of d. There is a d, which is denoted by d_{opt}, before which, the decreasing rate of the TAU cost is larger than the increasing rate of the paging cost, and after which, the situation is just the reverse.

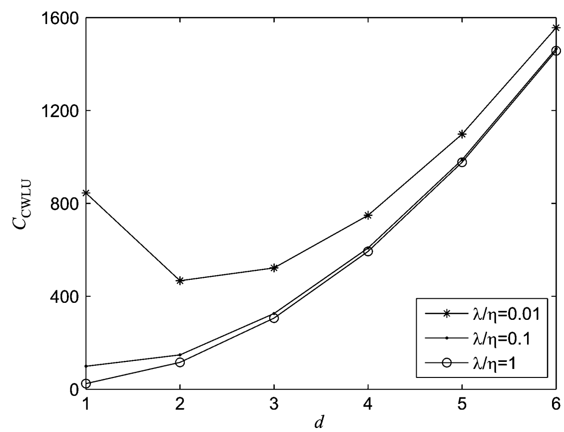


Fig. 5. $C_{\rm CWLU}$ versus d for various λ/η's. $\delta_{\rm TAU} = 50$.

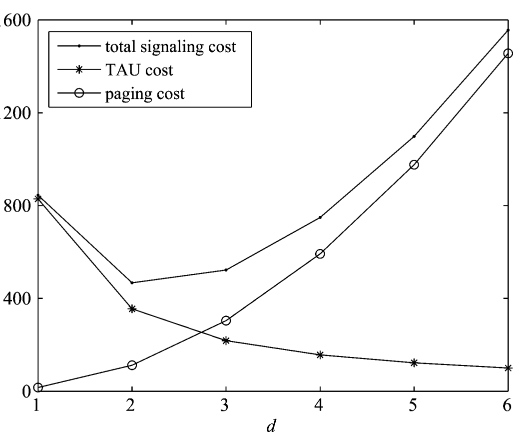


Fig. 6. Total signaling cost, TAU cost, and paging cost of the CWLU model versus d. $\lambda/\eta = 0.01$ and $\delta_{\rm TAU} = 50$.

This relation is clearer when expressed in mathematics. Denote by $C_{\rm TAU,CWLU}(d)$ the TAU cost of the CWLU model for a specific d. Similarly, we can define $C_{\rm paging,CWLU}(d)$. Suppose that d_1 and d_2 are two integers and that $d_1 < d_2$. It follows that when $d_1 < d_2 \leq d_{\rm opt}$, we have:

$$\frac{C_{\rm TAU,CWLU}(d_1) - C_{\rm TAU,CWLU}(d_2)}{d_2 - d_1} > \frac{C_{\rm paging,CWLU}(d_2) - C_{\rm paging,CWLU}(d_1)}{d_2 - d_1}.$$

It follows further that:

$$C_{\rm TAU,CWLU}(d_1) + C_{\rm paging,CWLU}(d_1) > C_{\rm TAU,CWLU}(d_2) + C_{\rm paging,CWLU}(d_2).$$

That is, before $d_{\rm opt}$, the total signaling cost is a decreasing function of d. Similarly, it can be proved that after $d_{\rm opt}$, the total signaling cost is an increasing function of d. Therefore, the total signaling cost is a downward convex function of d, and $d_{\rm opt}$ is the optimal d that can minimize the total signaling cost. 2) The larger the λ/η, the smaller the total signaling cost. The explanation is as follows. Since cell residence time t_s follows an exponential distribution, the cell-crossing events during t_c form an equilibrium renewal process [34]. Therefore, the expected number of cells crossed during call interarrival time t_c can be expressed as

$$\frac{E(t_c)}{E(t_s)} = \frac{1/\lambda}{1/\eta} = \frac{\eta}{\lambda}.$$

Thus, the increase in λ/η will lead to the decrease in the number of cells crossed during t_c and, accordingly, the decrease in the number of TAUs.

Fig. 7 shows the total signaling cost of the CWLU model as a function of the radius of a TA (i.e., d) when $\delta_{\rm TAU} = 20$, 40, and 80 under $\delta_{\rm paging} = 1$. In Fig. 7, we find that the larger the $\delta_{\rm TAU}$, the larger the total signaling cost. Furthermore, the impact of $\delta_{\rm TAU}$ is more pronounced when λ/η is small [see Fig. 7(a)].

Next, the impact of $N_{\rm TA}$ on the total signaling cost is investigated. With the previous narrations, $N_{\rm TA} = 16$ is not an effective TAL for all the UE units, because it generates a large paging cost. To assign effective TALs to UE units, the network can predict the mobility and traffic characteristics of individual UE units. Watanabe and Yabusaki in [35] proposed two schemes to obtain the UE's speed. Under one scheme, the UE measures its speed through a Global Positioning System receiver; under the other scheme, the UE estimates its speed based on the evolved Node B (eNB) coordinates' information advertised by the old and new eNBs. When the mobility characteristics can be obtained, the energy efficiency of networks can be improved [36], [37]. For instance, when a UE unit with high speed moves into a small cell from a macrocell, the network may prohibit the UE from doing a handover to the small cell to reduce the handover failure rate and improve the energy efficiency of backhaul networks. However, the detailed analysis process about the energy efficiency is out of the scope of this paper. In addition, the traffic arrival rate can be estimated at the network side by averaging the number of traffic arrivals during a certain period.

In Section III, we exemplified the modeling of the TAL-based scheme using a case of $N_{\rm TA} = 16$. It is straightforward to derive the analytical results for any other values of $N_{\rm TA}$. Fig. 8 shows the impact of d and $N_{\rm TA}$ on the total signaling cost of the CWLU model when $\delta_{\rm TAU} = 50$ and $\delta_{\rm paging} = 1$. In these figures, when d and λ/η are determined, there is a $N_{\rm TA}$, which is denoted by $N_{\rm TA,OPT}$, under which the total signaling cost is minimized. We use red circles to represent the minimum signaling cost. To facilitate observation, Table III summarizes the $N_{\rm TA,OPT}$'s in Fig. 8. Therefore, when the radius of a TA is planned by operators in advance, the network can allocate an optimal TAL of size $N_{\rm TA,OPT}$ for the UE, based on its mobility and traffic characteristics, to minimize the total signaling cost of the TAL-based scheme.

In Table III, we have the following observations. When $d \leq 3$, the smaller the λ/η, the larger the $N_{\rm TA,OPT}$. When $d \geq 7$, $N_{\rm TA,OPT} = 2$ for all the λ/η's. When d is small, e.g., $d = 2$, the smaller the λ/η, the larger the number of cells crossed during t_c. To reduce the TAU cost, the networks can assign a large TAL to the UE units. When $d \geq 7$, there are a

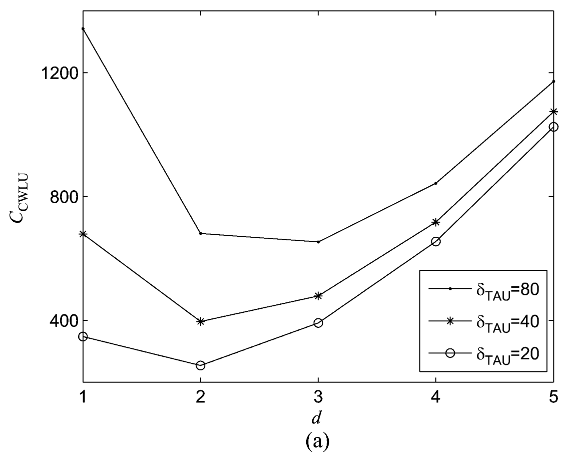


Fig. 7. C_{CWLU} versus d. (a) $\lambda/\eta = 0.01$. (b) $\lambda/\eta = 1$.

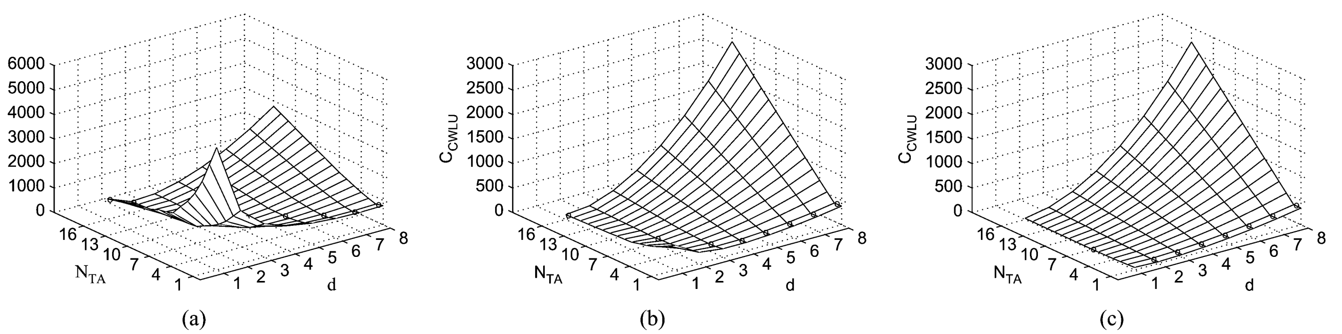


Fig. 8. C_{CWLU} versus N_{TA} and d. (a) $\lambda/\eta = 0.01$. (b) $\lambda/\eta = 0.1$. (c) $\lambda/\eta = 1$.

TABLE III
$N_{\mathrm{TA,OPT}}$'S UNDER DIFFERENT λ/η'S AND d'S

λ/η	0.01	0.1	1
d	1 2 3 4 5 6 ≥ 7	1 2 3 4 5 6 ≥ 7	1 2 3 4 5 6 ≥ 7
$N_{\mathrm{TA,OPT}}$	16 16 12 7 5 3 2	16 7 3 2 2 2 2	7 2 2 2 2 2 2

TABLE IV
COMPARISON OF THE SIGNALING COST OF THE COMMON TAL WITH THAT OF THE OPTIMAL TAL

λ/η	0.01	0.1	1
[9]	2247.4	229.2	27.4
our paper	844.4	98.8	22.0

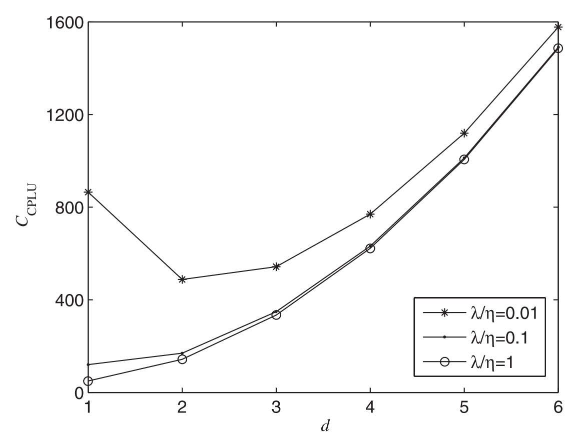


Fig. 9. C_{CPLU} versus d for various λ/η's. $\delta_{\mathrm{TAU}} = 50$ and $\delta_{\mathrm{TAU}'} = 30$.

large number of cells in a TAL; therefore, the number of TALs crossed during t_c is small, and the paging cost will dominate the total signaling cost. In this case, to reduce the paging cost, a small TAL should be assigned.

In [9] it is assumed that a cell assigns only one common TAL to all UE units registered in this cell. This approach of allocating TAL neglects the UE units' mobility and traffic traits. However, the optimal TAL derived in our paper is based on these traits. Table IV compares the signaling cost of the common TAL approach with the optimal TAL approach. In this table, it is assumed that a TA comprises only one cell (i.e., $d = 1$), and the common TAL includes five TAs. In this table, the signaling cost of the optimal TAL allocation is less than that of the common TAL allocation, manifesting the advantage of the optimal TAL derived in our paper.

C. Impact of Diverse Parameters on C_{CPLU}

In Section V-B, the impact of various parameters on the total signaling cost of the TAL-based scheme under the CWLU model is investigated. Here, we repeat this process for the CPLU model and consider that $\delta_{\mathrm{TAU}'} = 0.6\delta_{\mathrm{TAU}}$, $N_{\mathrm{TA}} = 16$, and $\delta_{\mathrm{paging}} = 1$. Figs. 9 and 10 show, respectively, the signaling cost of the CPLU model as a function of d when λ/η and δ_{TAU} vary. Generally, the observations that can be made from Figs. 9 and 10 are similar to those made from Figs. 6 and 7 for the CWLU model. The only exception is that when $\lambda/\eta = 1$, the impact of δ_{TAU} on C_{CPLU}, as shown in Fig. 10(b), is more

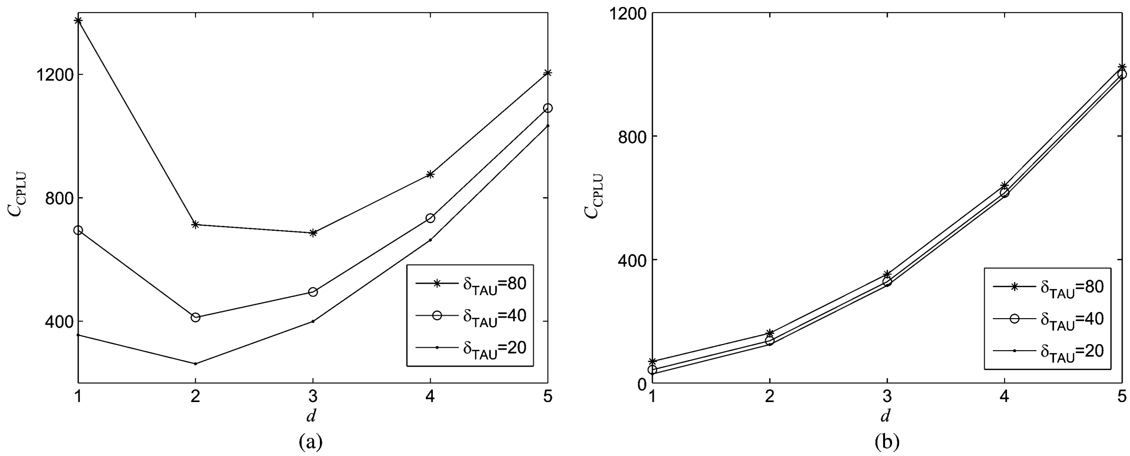


Fig. 10. C_{CPLU} versus d. (a) $\lambda/\eta = 0.01$. (b) $\lambda/\eta = 1$.

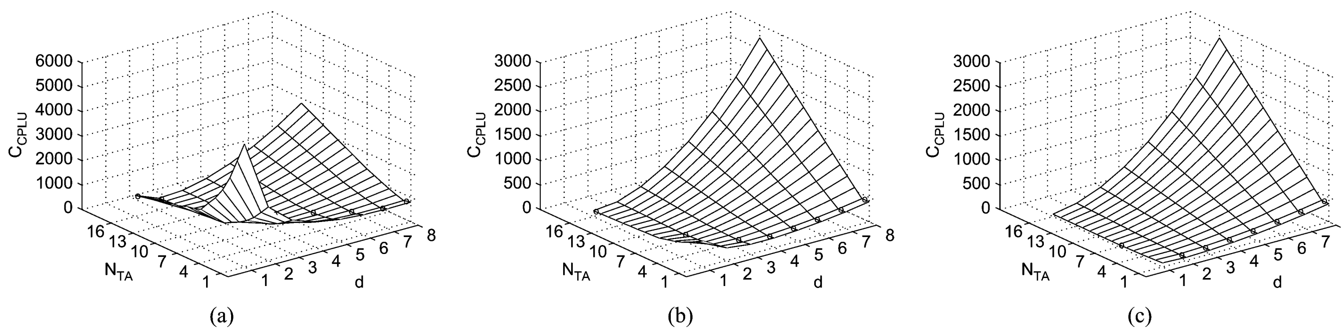


Fig. 11. C_{CPLU} versus N_{TA} and d. (a) $\lambda/\eta = 0.01$. (b) $\lambda/\eta = 0.1$. (c) $\lambda/\eta = 1$.

TABLE V
$N_{\mathrm{TA,OPT}}$'s UNDER DIFFERENT λ/η's AND d's

λ/η	0.01							0.1							1						
d	1	2	3	4	5	6	≥ 7	1	2	3	4	5	6	≥ 7	1	2	3	4	5	6	≥ 7
$N_{\mathrm{TA,OPT}}$	16	16	12	7	5	3	2	16	7	3	2	2	2	2	7	2	2	2	2	2	2

conspicuous that its impact on C_{CWLU}, as shown in Fig. 7(b). This exception is explained as follows. When $\lambda/\eta = 1$, be at the CWLU model or the CPLU model, there are quite a few TAUs. However, irrespective of the value of λ/η, there is always a TAU in the CPLU model. Since this paper assumed that $\delta_{\mathrm{TAU}'} = 0.6\delta_{\mathrm{TAU}}$, the conspicuous impact of δ_{TAU} on C_{CWLU} is, in fact, caused by $\delta_{\mathrm{TAU}'}$.

Next, we demonstrate the impact of N_{TA} on the total signaling cost of the CPLU model. Fig. 11 shows the signaling cost of the CPLU model for various λ/η's, d's, and N_{TA}'s. Table V summarizes the optimal size of a TAL (i.e., $N_{\mathrm{TA,OPT}}$) that can minimize the signaling cost. Contrasting Table III with Table V, we can find that the call handling model exerts no impact on $N_{\mathrm{TA,OPT}}$.

D. Impact of Call-Handling Models on the Total Signaling Cost

Here, we analytically compare the signaling cost of the two call handling models. Since the two models have the same paging cost, it follows that

$$C_{\mathrm{CPLU}} - C_{\mathrm{CWLU}} = (N_{\mathrm{TAU,CPLU}} - N_{\mathrm{TAU,CWLU}})\delta_{\mathrm{TAU}} + \delta_{\mathrm{TAU}'}.$$

Next, we quantify $N_{\mathrm{TAU,CPLU}} - N_{\mathrm{TAU,CWLU}}$ through considering two extreme cases. In the first extreme case, under the CWLU model, after the completion of a call, the UE remains in state $S_{0,0}$ so that in this case, the two models have the same number of TAUs during the call interarrival time t_c. In the second case, after the completion of a call, the UE resides in a border state and will move out of its current TAL right after the call. In this case, the CWLU model has one more TAU during t_c than the CPLU model. In any other more general cases, the difference between the signaling costs of the two models should be bounded by the differences in the two extreme cases. Therefore, we finally have

$$C_{\mathrm{CPLU}} - C_{\mathrm{CWLU}} \in \left[-\delta_{\mathrm{TAU}} + \delta_{\mathrm{TAU}'}, \delta_{\mathrm{TAU}'}\right].$$

For instance, under the conditions in Section V-A, when $d = 5$, $C_{\mathrm{CPLU}} - C_{\mathrm{CWLU}} = 21.1$. Therefore, compared with the total signaling cost (i.e., $C_{\mathrm{CPLU}} = 1119.3$ or $C_{\mathrm{CWLU}} = 1098.2$), the impact of the two handling models on the total signaling cost is insignificant.

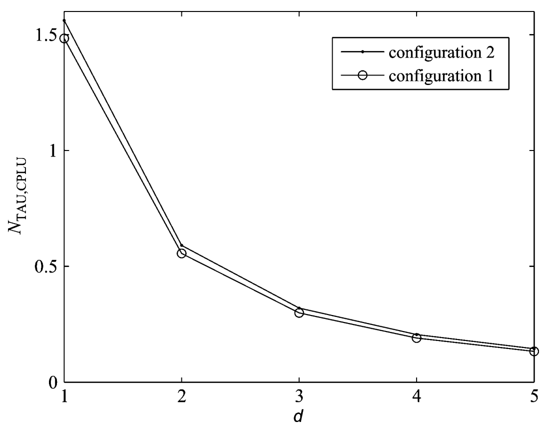


Fig. 12. $N_{\mathrm{TAU,CPLU}}$ versus d under two configurations. $\lambda/\eta = 0.1$.

E. Comparison Between the $N_{\mathrm{TAU,CPLU}}$'s Under Two TAL Configurations

As stated previously in Section II-A, when $N_{\mathrm{TA}} = 16$, there are two TAL configurations. As shown in Fig. 2, configuration 1 has two vertex TAs and one inner TA removed from the outermost ring, i.e., ring 2. The other configuration, i.e., configuration 2, has one vertex TA and two inner TAs removed from ring 2. When in a vertex TA, the probability of moving out of the current TAL is 1/2, whereas this probability becomes 1/3 when in an inner TA. Therefore, it can be anticipated that configuration 2 will have higher signaling cost than configuration 1, because the former has one more vertex TA than the latter. This anticipation is validated in Fig. 12, which compares the signaling cost of the two TAL configurations under the CPLU model. In Fig. 12, the increased signaling cost in configuration 2 is trivial compared with the signaling cost itself.

VI. Conclusion

To tackle the drawbacks of static LM schemes used in 2G and 3G cellular networks, LTE designs a new LM scheme, i.e., the TAL-based LM scheme. Instead of being allocated a single registration area in static LM schemes, under the new scheme, a UE unit is allocated a TAL that consists of multiple TAs. Within the allocated TAL, the UE can move freely without performing any TAUs, and only after moving out of the allocated TAL does the UE need to perform a TAU and then will be allocated a new TAL. The performance of the TAL-based scheme hinges on the allocated TAL. If the allocated TAL is inappropriate with respect to the mobility and traffic characteristics of the UE, the TAL-based scheme may produce an adverse effect on the LM overhead. The existing studies on the performance of the TAL-based scheme feature the following defeats. 1) Some studies assumed a TAL allocation approach that allocates a common TAL to all the UE units performing TAUs in the cell. First, this approach makes the TAL-based scheme vulnerable to the uneven distribution of the LU signaling and the ping-pong LU effect. Second, this approach failed to consider the diversity in the mobility and traffic characteristics of individual UE units so that the allocation common TAL cannot guarantee its effectiveness for all the UE units. Thus, this approach deprives the TAL-based scheme of its superiority over static LM schemes. 2) Some studies unreasonably assumed that a TA contains a single cell. 3) All the existing studies neglected the 16-TA upper bound set on the size of a TA by 3GPP. 4) In the existing studies, there is a lack of an accurate and, at the same time, feasible mathematical approach for calculating the signaling cost of the TAL-based scheme when it comes to global UE units that have weak regularity in their mobility.

To find an optimal TAL allocation approach for global UE units, this paper has developed a comprehensive mathematical approach to calculate the signaling cost of the TAL-based LM scheme. The following contributions distinguish the developed approach from those in existing studies. 1) The developed approach assumes that a TA contains a variable number of cells. 2) The developed approach obeys the 16-TA upper bound. 3) The developed approach emphasizes the impact of call handling models on the signaling cost by considering two different call handling models, i.e., the CWLU and the CPLU. For simplicity of mathematical analysis, most of the existing studies on LM considered only the CPLU model. However, the reality is that existing cellular networks use only the CWLU model. 4) The developed approach exploits a fluid flow model to characterize the dependence among the cell residence time, the TA residence time, and the TAL residence time. As proved in this paper, the fluid flow model is simple and accurate. Most of the existing studies did not consider this dependence. 5) The developed approach applies embedded Markov chains that are capable of taking into account a plethora of complex situations when describing the TAL-based scheme.

As attributed to the developed mathematical approach, this paper derives analytical formulas for the signaling cost of the TAL-based LM scheme caused by TAU and paging operations, as well as formulas that are useful in designing an optimal sequential paging scheme. The accuracy of these formulas was validated through simulation. Based on these formulas, numerical studies are conducted to investigate the impact of various parameters on the signaling cost. In these numerical studies, we mainly observed the following. 1) The total signaling cost is a downward convex function of the radius of a TA in terms of cells. 2) The call handling models exert insignificant influence on the total signaling cost. 3) Given the call-to-mobility ratio and the radius of a TA, there is always an optimal TAL size that can minimize the signaling cost.

Next, we discussed the operation complexity of the dynamic TAL-based LM scheme. As for the design complexity of the dynamic LM, the network must have the capability to determine the size of a TA. Moreover, as mentioned in Section II-A, the TAs constituting a TAL need to satisfy two conditions. 1) The number of TAs cannot exceed 16. 2) The TA where the UE performs the last TAU is the center TA. Therefore, to allocate a TAL to the UE, the network must have the topology information of networks. The information on the optimal TAL can be transferred from the network to the UE when the UE performs a TAU. In fact, the number of cells in a TA and the topology information are planned by operators in advance, so that the dynamic LM does not add extra operation complexity.

Finally, we briefly discussed the potential application of the proposed modeling approach in heterogeneous networks (HetNets). In HetNets, small cells are largely deployed within

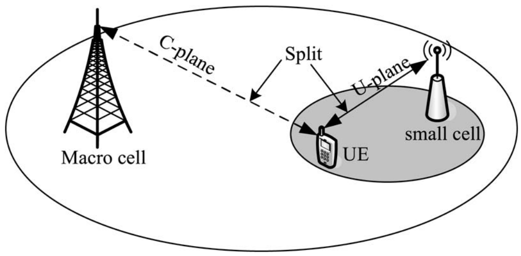


Fig. 13. Dual connectivity.

macrocells to enhance network capacity and provide better user experience. However, as the coverage of small cells is limited, HetNets pose challenge in terms of mobility management, e.g., frequent handovers. To overcome this challenge, Ishii *et al.* in [38] proposed a novel architecture, which is called the control-plane/user-plane split. In this architecture, if a UE unit resides in the overlapped region between a macrocell and small cells, the control plane must be supported by a more continuous and reliable coverage layer at lower-frequency macrocells, whereas the user plane can be provided by high-frequency small cells. This architecture is also called dual connectivity in 3GPP TS 36.300 [39], as shown in Fig. 13. In the dual connectivity, macrocells are responsible for sending paging messages and other control messages. When a macrocell and small cells within the coverage of the macrocell are allocated the same TA identity, the UE moving from the macrocell/small cell into the small cell/macrocell does not need to perform a TAU. Under such case, the impact of small cells on the LM can be ignored. Thus, the proposed mathematical model and the derived analysis results in this paper can also be applied to HetNets.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers, whose comments substantially enhanced the quality of this paper.

REFERENCES

[1] "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 13)," Third-Generation Partnership Project, Sophia Antipolis Cedex, France, Tech. Spec. Group Serv. Syst. Aspects 3GPP TS 23.401 V13.1.0, Dec. 2014.

[2] "Tracking areas sizes and tracking area list optimization," Mitsubishi Elect., Tokyo, Japan, Tech. Rep. 3GPP TSG RAN WG3 Meet. R3-071913, 2007.

[3] S.-L. Wu, J.-J. Chen, and W.-C. Chou, "Cell-related location area planning for 4G PCS networks with variable-order Markov model," *J. Syst. Softw.*, vol. 86, no. 10, pp. 2688–2699, Oct. 2013.

[4] M. Toril, S. Luna-Ramírez, and V. Wille, "Automatic replanning of tracking areas in cellular networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2005–2013, Jun. 2013.

[5] S. M. Razavi, D. Yuan, F. Gunnarsson, and J. Moe, "Performance and cost trade-off in tracking area reconfiguration: A Pareto-optimization approach," *Comput. Netw.*, vol. 56, no. 1, pp. 157–168, 2012.

[6] Y. W. Chung, "Adaptive design of tracking area list in LTE," in *Proc. IEEE WOCN*, 2011, pp. 1–5.

[7] J. Ferragut and J. Mangues-Bafalluy, "A self-organized tracking area list mechanism for large-scale networks of femtocells," in *Proc. IEEE ICC*, 2012, pp. 5129–5134.

[8] S. M. Razavi, D. Yuan, F. Gunnarsson, and J. Moe, "Exploiting tracking area list for improving signaling overhead in LTE," in *Proc. IEEE VTC*, 2010, pp. 1–5.

[9] S. M. Razavi, D. Yuan, F. Gunnarsson, and J. Moe, "Dynamic tracking area list configuration and performance evaluation in LTE," in *Proc. IEEE GLOBECOM Workshops*, 2010, pp. 49–53.

[10] S. M. Razavi, "Tracking area planning in cellular networks," Ph.D. dissertation, Dept. Sci. Technol., Linköping Univ., Linköping, Sweden, 2011.

[11] S. M. Razavi and D. Yuan, "Mitigating mobility signaling congestion in LTE by overlapping tracking area lists," in *Proc. 14th ACM Int. Conf. Model., Anal. Simul. Wireless Mobile Syst.*, 2011, pp. 285–292.

[12] R.-H. Liou, Y.-B. Lin, and S.-C. Tsai, "An investigation on LTE mobility management," *IEEE Trans. Mobile Comput.*, vol. 12, no. 1, pp. 166–176, Jan. 2013.

[13] D. Pacheco-Paramo, I. F. Akyildiz, and V. Casares-Giner, "X2-interface-based location management for small cell networks," in *Proc. IEEE GLOBECOM Workshops*, 2013, pp. 4976–4981.

[14] R.-H. Liou and Y.-B. Lin, "Mobility management with the central-based location area policy," *Comput. Netw.*, vol. 57, no. 4, pp. 847–857, Mar. 2013.

[15] S. M. Razavi and D. Yuan, "Mitigating signaling congestion in LTE location management by overlapping tracking area lists," *Comput. Commun.*, vol. 35, no. 18, pp. 2227–2235, Nov. 2012.

[16] X. Wang, X. Lei, R. Q. Hu, and Y. Qian, "Modeling of tracking area list-based location update scheme in long term evolution," in *Proc. IEEE ICC*, 2014, pp. 2325–2331.

[17] K. Li, "Analysis of distance-based location management in wireless communication networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 2, pp. 225–238, Feb. 2013.

[18] X. Wang, X. Lei, P. Fan, R. Q. Hu, and S.-J. Horng, "Cost analysis of movement-based location management in PCS networks: An embedded Markov chain approach," *IEEE Trans. Veh. Technol.*, vol. 63, no. 4, pp. 1886–1902, May 2014.

[19] X. Wang *et al.*, "Cost analysis of a hybrid movement-and time-based location update scheme in cellular networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5314–5326, Nov. 2015.

[20] L. Zhou, R. Q. Hu, Y. Qian, and H.-H. Chen, "Energy-spectrum efficiency tradeoff for video streaming over mobile ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 981–991, May 2013.

[21] L. Zhou, H. Wang, and M. Guizani, "How mobility impacts video streaming over multi-hop wireless networks?" *IEEE Trans. Commun.*, vol. 60, no. 7, pp. 2017–2028, Jul. 2012.

[22] Y.-B. Lin, L.-F. Chang, and A. Noerpel, "Modeling hierarchical microcell/macrocell PCS architecture," in *Proc. IEEE ICC*, 1995, vol. 1, pp. 405–409.

[23] I. F. Akyildiz, Y.-B. Lin, W.-R. Lai, and R.-J. Chen, "A new random walk model for PCS networks," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 7, pp. 1254–1260, Jul. 2000.

[24] Y.-C. Tseng and W.-N. Hung, "An improved cell type classification for random walk modeling in cellular networks," *IEEE Commun. Lett.*, vol. 5, no. 8, pp. 337–339, Aug. 2001.

[25] V. C. Giner, V. Pla, and P. Escalle-Garcia, "Mobility models for mobility management," in *Network Performance Engineering*, ser. Next Generation Internet, D. Kouvatsos, Ed. Berlin, Germany: Springer-Verlag, 2011, pp. 716–745.

[26] H. Xie, S. Tabbane, and D. Goodman, "Dynamic location area management and performance analysis," in *Proc. IEEE VTC*, 1993, pp. 536–539.

[27] G. Morales-Andres and M. Villen-Altamirano, "An approach to modelling subscriber mobility in cellular radio networks," in *Proc. 5th World Telecommun. Forum*, 1987, pp. 185–189.

[28] R. Thomas, H. Gilbert, and G. Mazziotto, "Influence of the movement of the mobile station on the performance of a radio cellular network," in *Proc. 3rd Nordic Semin.*, 1988, pp. 4–9.

[29] I. F. Akyildiz and W. Wang, "A dynamic location management scheme for next-generation multitier PCS systems," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 178–189, Jan. 2002.

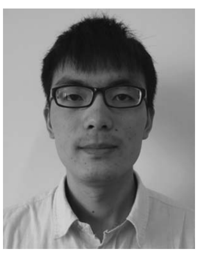
[30] C. Makaya and S. Pierre, "An analytical framework for performance evaluation of iPv6-based mobility management protocols," *IEEE Trans. Wireless Commun.*, vol. 7, no. 3, pp. 972–983, Mar. 2008.

[31] S. Mohan and R. Jain, "Two user location strategies for personal communications services," *IEEE Pers. Commun.*, vol. 1, no. 1, pp. 42–50, Jun. 1994.

[32] F. V. Baumann and I. G. Niemegeers, "An evaluation of location management procedures," in *Proc. IEEE 3rd Annu. Int. Conf. Univ. Pers. Commun. Rec.*, 1994, pp. 359–364.

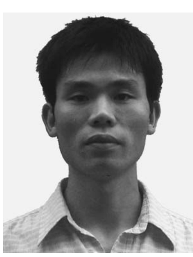
[33] Y. Zhu and V. C. Leung, "Derivation of moving distance distribution to enhance sequential paging in distance-based mobility management for PCS networks," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3029–3033, Nov. 2006.

[34] D. R. Cox, *Renewal Theory*, vol. 4. London, U.K.: Methuen, 1962.
[35] Y. Watanabe and M. Yabusaki, “Mobility/traffic adaptive location management,” in *Proc. IEEE VTC*, 2002, pp. 1011–1015.
[36] X. Ge, S. Tu, T. Han, Q. Li, and G. Mao, “Energy efficiency of small cell backhaul networks based on Gauss–Markov mobile models,” *IET Netw.*, vol. 4, no. 2, pp. 158–167, Mar. 2015.
[37] X. Ge *et al.*, “Spatial spectrum and energy efficiency of random cellular networks,” *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 1019–1030, Mar. 2015.
[38] H. Ishii, Y. Kishiyama, and H. Takahashi, “A novel architecture for LTE-B: C-plane/U-plane split and phantom cell concept,” in *Proc. IEEE GLOBECOM*, 2012, pp. 624–630.
[39] “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall, description; Stage 2 (Release 12),” Third-Generation Partnership Project, Sophia Antipolis Cedex, France, Tech. Spec. Group Radio Access Netw. 3GPP TS 36.300 V12.4.0, Dec. 2014.



Tao Deng (S’14) received the B.Eng. degree from Henan Normal University, Xinxiang, China, in 2012. He is currently working toward the Ph.D. degree with the Key Laboratory of Information Coding and Transmission, School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China.

His research interests include mobility management, software-defined networking, performance modeling for wireless networks, and machine-type communications.



Xian Wang (M’09) received the B.Eng. and Ph.D. degrees in communication and information systems from Southwest Jiaotong University (SWJTU), Chengdu, China, in 2002 and 2008, respectively.

He is currently an Associate Professor with the School of Information Science and Technology, SWJTU. His research interests include mobility management and performance modeling for wireless mobile networks.

Pingzhi Fan (M’93–SM’99–F’15) received the Ph.D. degree in electronic engineering from the University of Hull, Kingston upon Hull, U.K., in 1994.

He is currently a Professor and the Director of the Institute of Mobile Communications with Southwest Jiaotong University, Chengdu, China. He has over 200 research papers published in various academic English journals (IEEE/IEE/IEICE) and eight books (also edited) and is the inventor of 20 granted patents. His research interests include high-mobility wireless communications, fifth-generation technologies, wireless networks for big data, and signal design and coding.

Dr. Fan received the U.K. Overseas Research Students Award and the Outstanding Young Scientist Award from the National Natural Science Foundation of China. He is also a Chief Scientist of a national 973 research project. He served as a General Chair or a Technical Program Committee Chair for a number of international conferences and is the Guest Editor-in-Chief, Guest Editor, or editorial member of several international journals. He is the Founding Chair of the IEEE Vehicular Technology Society BJ Chapter and the IEEE Communications Society CD Chapter, as well as the Founding Chair of the IEEE Chengdu Section. He also served as a Board Member of the IEEE Region 10, IET(IEE) Council and of the IET Asia-Pacific Region.

Keqin Li (F’15) received the B.S. degree in computer science from Tsinghua University, Beijing, China, in 1985 and the Ph.D. degree in computer science from the University of Houston, Houston, TX, USA, in 1990.

He is a Distinguished Professor of computer science with the State University of New York, New Paltz, NY, USA. He has published over 360 journal articles, book chapters, and refereed conference papers. His current research interests include parallel computing and high-performance computing, distributed computing, energy-efficient computing and communication, heterogeneous computing systems, cloud computing, big data computing, central processing unit/graphics processing unit hybrid and cooperative computing, multicore computing, storage and file systems, wireless communication networks, sensor networks, peer-to-peer file-sharing systems, mobile computing, service computing, Internet of things, and cyberphysical systems.

Mr. Li has received several Best Paper Awards. He is currently with or has served on the Editorial Boards of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, and the *Journal of Parallel and Distributed Computing*.