On dynamic signaling congestion mitigation by overlapping tracking area lists

Sara Modarres Razavi, Di Yuan, Fredrik Gunnarsson and Johan Moe

Linköping University Post Print

N.B.: When citing this work, cite the original article.

On Dynamic Signaling Congestion Mitigation by Overlapping Tracking Area Lists

Sara Modarres Razavi[a,*], Di Yuan[a,b], Fredrik Gunnarsson[b], Johan Moe[b]

[a]*Department of Science and Technology, Linköping University, Sweden*
[b]*Ericsson Research, Ericsson AB, Sweden*

Abstract

Mitigating signaling congestion of tracing user equipments (UEs), adaptively to the changes in UE location and mobility patterns is a challenging issue in mobility management of Long Term Evolution (LTE) networks. Signaling congestion usually occurs due to many UEs behaving in a similar manner, e.g., massive and simultaneous UE mobility in a train movement scenario. LTE networks allow the use of tracking area lists (TALs), each being a list containing multiple tracking areas (TAs). The overlapping TAL scheme has been previously used for signaling congestion mitigation for snapshot scenarios. For maintaining the improved performance over non-list-oriented TA configuration over time, an automatic dynamic configuration framework, which is a key aspect in Self Organizing Network (SON), has been applied in this paper. We develop a linear programming model for optimal TAL configuration. Repetitively solving the model for different time intervals gives an evaluation framework on the performance of SON location management. Comprehensive numerical results are presented in this study, using a large-scale realistic network scenario. The experiments demonstrate the effectiveness of the SON dynamic framework in reducing the total signaling overhead of the network compared to the static TA. Moreover, the overlapping TAL scheme significantly improves the performance of the network in the tracking area update congested scenarios over the conventional TA configuration.

*Corresponding author. E-mail: sara.modarres.razavi@ericsson.com

1. Introduction

Location management, which is sometimes called reachability, aims to track and page user equipments (UEs) in cellular network. Tracking Area (TA) in Long Term Evolution (LTE) is a logical area-partition of the network, that each partition is a subset of cells [1]. In the idle state, a UE is in power-conservation mode and does not inform the network of each cell change. Hence, the location of idle UEs is known to the network to the granularity of TA. In the conventional TA scheme, when a UE passes a TA boundary, it sends an uplink signaling message to the Mobility Management Entity (MME). This procedure is called tracking area update (TAU). On the other hand, for placing a call to a UE, the MME sends downlink paging signaling messages to the cells inside the UE's current TA, in order to find the cell from which the UE can receive the call. UEs in active mode also perform TAUs as part of the handover (HO) procedure. However, while UEs are in active mode the corresponding signaling due to TAU is a small portion compared to the whole signaling including HO procedure. Therefore, the topic of this research becomes interesting while focusing on idle UEs, due to the fact that the number of idle UEs is significantly more than the active ones.

In designing the TA layouts of a network, the overheads resulted by TAU and paging signaling messages are the key parameters to be taken into account. Planning and optimization of TAs (or Location Areas (LAs) and Routing Areas (RA) for the previous generations of cellular networks) by considering signaling overhead have been dealt with in a large amount of literature [8, 10, 12, 20, 33, 34]. There are two new concepts in LTE networks, that need to be more explored and investigated with the objective of improving the network performance: One is the concept of self-organizing networks (SON) [4, 5], and two is the concept of TA list (TAL) [2]. Both of these concepts are standardized by 3rd Generation

Partnership Project (3GPP), Release 8, for LTE. In this paper, we aim to reduce the operating expenditure (OPEX) cost and improve the overall performance of LTE's location management by deploying these two concepts in the network.

Conventionally, TAs are manually configured and the configuration is static. As UE distribution and mobility patterns change over time, the TA configuration of the network does not perform optimally for all time intervals. In the conventional TA scheme, reconfiguring TAs, such as moving a cell from one TA to another, requires temporarily tearing down the cell causing service interruption [19]. Due to the excessive cost caused for the operator, it is not feasible to make small time interval reconfigurations in the conventional TA scheme. However in SON, an automated and efficient deployment of updated configurations is possible. For a stable optimization, a global view of the UEs' movements and call arrival rate statistics is a requirement. In a static TA, the average of this statistic is used for the TA design. However, in a dynamic framework, we incorporate time of day and day of week data patterns into our design, which can further strengthen the performance of the network over the static TA.

During signaling congestion, there is resource exhaustion for tracking UEs. Therefore, mitigating signaling congestion is to ensure no significant degradation in the quality of service (QoS) of a network. Normally, the TAU congestion occur in densely populated cities, where there is a huge number of UEs moving concurrently during specific times of day and night (rush-hours) [3]. Paging signaling congestion may also occur in scenarios where close-to-static UEs are simultaneously gathered at some hotspots. The focus in this paper is only on TAU congestion mitigation, with the assumption that it is a more critical issue in real-life deployments. The optimization framework used in this paper for mitigating TAU signaling congestion has been introduced in [24] and is based on the use of TA lists (TALs), which may contain overlapping TAs. Instead of updating the UE with one TA, the UE can be assigned a list of TAs, referred as a TAL. The UE performs a TAU, when it moves to a TA not being in its TAL. Hence, the location of a UE is known in the MME to the granularity of its allocated TAL. In the overlapping TAL scheme used in this paper, all cells

3

in one TA use the same collection of TALs, and they update their UEs with the same proportional use of each TAL. The standard TA design is considered as the underlying structure. The optimal proportional use of the TALs for congestion mitigation is obtained via a linear program (LP).

One difficulty in evaluating the performance of TAL is the existence of UEs in a same location with different TALs, and thus calculating the total number of TAUs depend on the specific mobility pattern of each individual UE. This information is not available, and therefore the performance of TAL configuration rely heavily on the validity of the assumptions made in the models [11, 21, 22]. The main advantage of applying overlapping TAL is that no mobility information is required for the calculation of the signaling overhead. Moreover, as all cells in the same area will assign TAL consistently based on their proportional usage, the performance evaluation can be effectively assessed.

The creation and reconfiguration of TALs are supported in SON architecture, therefore our dynamic optimization framework is cost efficient in implementation. One main advantage of using overlapping TAL scheme for SON is the possibility of entering the TAs with high number of TAUs at their boundaries and their neighboring TAs to the optimization formulation and hence applying it only to the congested area and not to the whole network. This will significantly reduce the problem size, which makes it practically useful to implement. However, we should consider that as TAs are geographically connected to each other, solving the congestion problems for subsets of TAs may omit the optimum solutions of TAL proportions for the overall network. Hence in this paper, we apply the model to the whole network and solve the problem of exponentially higher size. Numerical evaluations for mitigating massive number of TAUs demonstrates the effectiveness of applying overlapping TAL on a dynamic framework compatible to SON.

The remainder of the paper is constructed as follows. In Section 2, we review some works which are relevant to our study. Section 3 is devoted to description of the dynamic framework, basic notation and an illustrative example of the overlapping TAL scheme. The optimization formulation for using the over-

lapping TAL scheme for congestion avoidance of TAU is given in Section 4. In Section 5, we explain a mechanism to apply the proposed overlapping TAL scheme on a SON dynamic framework. In Section 6, we present results of performance evaluation of the proposed approach and compare the static TA with the SON framework. Conclusions are provided in Section 7.

2. Related works

As mentioned in Section 1, there is an impressive number of surveys on optimization models and algorithms of conventional TA configuration [8, 10, 12, 20, 33, 34]. On the design of unconventional location update and paging schemes, which form another line of research, see for example [6, 28, 35]. For thorough surveys on location management, we refer to [7, 36].

Recently, the concept of TAL in LTE has received a great attention due to its flexibility over the conventional TA concept [18, 27, 30, 31, 38]. Previous studies of using TAL for minimizing the total signaling overhead, with the assumption that a cell updates one common TAL to all UEs getting registered in that cell, have been previously presented in [11, 21, 22]. TALs, in forms of rings of cells, are used in [11]. In [21, 22], each cell belongs to a specific TAL. Any UE getting registered in the cell, is updated by that specific TAL. In [25], the authors prove that the optimum TAL configuration solution for minimizing the sum of TAU and paging signaling overhead is to assign only one TAL to each cell.

In general, to optimize the configuration of TALs, a major challenge is to estimate the signaling overhead. UEs in the same location may be registered to different TALs, thus the resulting overhead, in contrast to the conventional TA case, heavily depends on the forthcoming UE behaviors. Hence, assumptions on UE behavior, ranging from simple rule of thumb to statistical mobility models, are required. Nevertheless, previous works, in particular [21, 22], show that TAL is promising in extending the capability of the conventional TA scheme.

TAL is promising for solving a couple of issues imposed by the conventional TA scheme. For example, TAL can be used to prevent the ping-pong effect,

5

which is the frequent updates when a UE keeps hopping between two or more adjacent cells [27]. Another example is the train scenario associated with high uplink signaling traffic due to simultaneous updates of massive UEs crossing a boundary along a path [18, 23]. In [24], the authors generalize the TAL optimization formulation of a line-form train scenario in [23] to an arbitrary network topology. In the current paper, we use the TAL optimization model for congestion mitigation of TAU given in [24]. Hence, the following study is an extension to [24] by introducing a dynamic framework to the model and adapting the TAL optimization formulation to SON. Moreover, the large-scale realistic dynamic scenario used in performance evaluation of this paper not only present the potential of overlapping TAL on a SON paradigm, but also presents a close-to-reality overview of mobility behavior and call arrival rate of UEs in a network for one week time horizon.

The problem of load balancing and signaling congestion mitigation in location management are investigated in [17, 29, 32, 37]. In [17], the authors use randomized time distribution of cell reselection to avoid the mass mobility signaling congestion. However, their mechanism requires additional network configuration in order to provide the algorithm parameters. The main contributions of [29, 32] are models and algorithms for load balancing of the database servers in location update. However, most of the mechanisms of unconventional update and paging schemes require the use of non-standard system elements and parameters. The authors in [37] present a queuing model for location updates by group mobility, and examine the performance of using an information buffer at base stations to reduce update failures. The TAL scheme for signaling congestion mitigation used in the current paper allows a cell to belong to a collection of potentially overlapping TALs, which can be assigned to its UEs. As cells are fully consistent in the proportional use of the correlated TALs, the signaling overhead estimation becomes straightforward and there is no need for any UE mobility assumption.

Self-optimization have already been studied in 2G and 3G networks, typically with the objective of improving radio coverage and channel capacity, however in

6

recent years, there has been an increasing interest in the SON paradigm applied to mobility and location management of wireless networks [9]. The authors in [15] present a self-organized TAL mechanism to reduce the signaling overhead of a network of femtocells. The mechanism adapts the size of UE-specific TALs to the mobility speed and the paging arrival rate of each UE. Their performance evaluation is based on a simple hexagonal network topology, and a Markov-based mobility model. Our application focus is signaling congestion mitigation, and our proposed self-organized TAL framework is applied to a real-life large-scale network topology, and it is independent of UEs mobility model. Moreover, our study can also be applied to femtocell networks.

3. System model

In this section, first we propose a proper mechanism to use the data collected from SON for the design of conventional TA and TAL layout in a dynamic framework. Second, a detailed description of the basic elements representing the design model is given. Third, the overlapping TAL scheme is explained through an illustrative example. Table 1 summarizes the mathematical models notation and descriptions used in this paper.

3.1. SON dynamic framework

Until now, cellular operators have been using an off-line approach to configure TAs. Due to the complexity of TA reconfiguration, the operators mostly decide about the TA of each cell at the time of network deployment, and changes are made only in case of extreme performance degradations. In SON, the aim is to automate the optimization of TAs or TALs in varying network conditions, which results in reduction of OPEX costs. For example, when the user distribution and mobility patterns change, the network may try to move the TA boundaries to the areas where there are less UE distribution [31]. For the current off-line approach, this is either infeasible or highly expensive in terms of service interruption, meaning that changing the TA layout of an area requires temporarily tearing down the cells in that part.

Table 1: Mathematical models notation and descriptions

Parameters	
α	Call intensity factor (probability that a UE has to be paged)
β	Allowed percentage increase on the total paging
λ	Smoothing factor
ξ	Paging threshold value
$a_{v\ell}$	Binary indicator that identifies if TA v is in TAL ℓ
c^p	The overhead of a single paging
c^u	The overhead of a single TAU
h_{ij}	Number of UEs having HO from cell i to cell j
h_{vw}	Number of UEs having HO from TA v to TA w
n_v	Number of cells in TA v
u_i	Number of UEs in cell i scaled by the time proportion spent in i
u_v	Number of UEs in TA v scaled by the time proportion spent in v
z	Maximum TAU at TA boundaries in a congestion scenario
$C^P(d)$	Total paging signaling overhead of the data set d
$C^P_{additional}$	Total additional paging signaling overhead
$C^U(d)$	Total TAU signaling overhead in the data set d
D	Number of data sets
N	Number of cells
$\mathbf{S}$	Binary matrix presenting the TA configuration of all cells
T	Number of TAs
Sets	
$\mathcal{D}$	Set of all data in a dynamic framework
$\mathcal{L}$	Set of all possible TALs of $\mathcal{T}$
$\mathcal{N}$	Set of all cells
$\mathcal{T}$	Set of TAs currently in use
Variables	
s^d_{ij}	Binary variable that identifies if i and j are in the same TA in the data set d
x_ℓ	Continuous variable that represents the proportional use of TAL ℓ by all cells of the TAs in ℓ

8

In SON, there are software clients such as data collectors (DCs) which collect data to monitor network performance [9]. The optimization and new configuration decisions are based on the collected data, however no data is collected for the location and mobility behavior of idle UEs, which are the focus in the study of location management. Although the number of idle UEs is significantly more than the active UEs still we can estimate that the ratio of distribution of idle UEs in the network and the mobility pattern of these UEs would be similar to the HO and cell-load data which belong to active UEs. Other approaches for estimating the behavior of idle UEs include network simulation [34] and examining traffic density on roads across neighboring cells [10].

In this study, the collected load and HOs of each cell are the input data to our optimization formulation. We denote the set of cells in an LTE network by $\mathcal{N} = \{1, \ldots, N\}$ and u_i is the total number of UEs in cell i scaled by the time period that each UE spends in cell i. For the same time period, h_{ij} is the number of UEs moving from cell i to cell j. In this section, a mechanism to use the data collected from SON is proposed for the design of conventional TA and TAL layout in a dynamic framework. A realistic assumption considered in this study is that the similarity of data degrades by the increase of time interval.

The dynamic SON time framework is denoted by the set $\mathcal{D} = \{1, \ldots, D\}$, and D is the total number of data sets. The data entering the optimization model $\bar{u}_i(d)$ and $\bar{h}_{ij}(d)$ are defined to be the weighted sum of the previous value and the new data gathered in the last time slot $u_i(d - 1)$ and $h_{ij}(d - 1)$. The idea is to smoothly reduce the impact of old data sets, and not making a sharp change in the design of two consecutive time slots. With each collected data from SON, $u_i(d - 1)$ and $h_{ij}(d - 1)$, the new aggregated $\bar{u}_i(d)$ and $\bar{h}_{ij}(d)$ are computed by Equations (1) and (2). These two equations are referred to as "exponential filters" in the glossary of statistical terms, where λ is a constant between 0 and 1 that smooths the changes in $\bar{u}_i(d)$ and $\bar{h}_{ij}(d)$.

$$\bar{u}_i(d) = (1 - \lambda)\bar{u}_i(d - 1) + \lambda u_i(d - 1) \qquad (1)$$

$$\bar{h}_{ij}(d) = (1 - \lambda)\bar{h}_{ij}(d - 1) + \lambda h_{ij}(d - 1) \tag{2}$$

3.2. Basic elements

A TA design, which is a partitioning of cells in the network, can be represented by an $N \times N$ symmetric and binary matrix $\boldsymbol{S}$. As indicated in (3), s_{ij}, which is the entry of $\boldsymbol{S}$ at row i and column j, indicates whether or not cells i and j are in the same TA. Due to the fact that TAs do not overlap, the $\boldsymbol{S}$-matrix has the transitive property, meaning that for any three cells i, j, and k, $s_{ij} = s_{ik} = 1$ implies $s_{jk} = 1$.

$$s_{ij} = \begin{cases} 1 & \text{if cells } i \text{ and } j \text{ are in the same TA,} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The amount of signaling overhead of one TA update and one paging are denoted by c^u and c^p, respectively. The exact relationship between c^u and c^p depends on radio system specification [13]. In a dynamic framework, the probability that a UE has to be paged (i.e., the call intensity factor) is time-varying over different hours of a day. Since we apply the data of active UEs as an approximation of idle ones, the call intensity variability lies within the cell-load data of each time slot. Hence, for simplicity and without loss of generality, the parameter α, which was defined as the call intensity factor in [19, 20], is assumed to be a fixed load-scale parameter over d to only present the trade-off between TAU and paging.

The total TAU and paging signaling overhead for the given TA design $\boldsymbol{S}$ are denoted by C^U and C^P, respectively. Equations (4) and (5) compute the TAU and paging signaling overheads for each time slot d. If cells i and j are in different TAs, according to Equation (4) the movement between the two cells contributes to C^U. Based on Equation (5), the total paging overhead generated by cell i to other cells in the same TA is computed by the load of cell i and the size of that TA. Note that a paging overhead is generated in each of the cells in the TA, except the cell where the UE is located. This is considered by

the inner summation. By adding the paging overhead of all cells in the network (the outer summation), the total paging overhead is calculated for time slot d.

$$C^U(d) = \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}: j \neq i} c^u \bar{h}_{ij}(d)(1 - s_{ij}^d), \quad d \in \mathcal{D} \tag{4}$$

$$C^P(d) = \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}: j \neq i} \alpha c^p \bar{u}_i(d) s_{ij}^d, \quad d \in \mathcal{D} \tag{5}$$

In designing the TA layout, one has to deal with a tradeoff between C^U and C^P in (4) and (5). Very small TAs (e.g., one cell per TA corresponding to a diagonal $\boldsymbol{S}$ matrix) virtually eliminates paging, but causes excessive TAU, whereas TAs of too large size give the opposite effect.

3.3. Overlapping TAL scheme

In this section, the overlapping TAL scheme for TAU congestion mitigation is explained using a small example of a network with four TAs in Figure 1. The dark solid lines in the figure present the boundaries of each TA. Each of these TAs contain multiple cells and our assumption is that the TA layout in Figure 1 is optimum in terms of the total signaling overhead of the conventional TA scheme. Suppose TALs which may partially overlap in their composition are created and assigned to UEs. Overall for the example in Figure 1, there are a total of fifteen TALs, each being a non-empty subset of $\{1, 2, 3, 4\}$. Any UE in this network will be registered to one of these fifteen TALs and each TA belongs to its respective collections of TALs. For example, the cells in TA 1 potentially belong to eight TALs, which can be assigned to the UEs registered or updated in 1: $\{1\}$, $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{1, 3, 4\}$, $\{1, 2, 3, 4\}$. Each of these eight TALs has a proportional value of usage, and the sum of all these values is one.

One should consider that for example TAL $\{1, 3\}$ can be assigned to a UE from cells of both TAs 1 and 3. If 1 and 3 assign different proportional values of this TAL to their UEs, then the total number of UEs being registered to the TAL in the area of these two TAs will depend on the UEs mobility behaviors. One
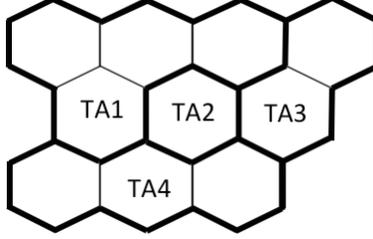
Figure 1: A small network with four TAs.

of the objectives in developing the overlapping TAL scheme is to have a scheme which the calculation of its signaling overhead and its performance evaluation are independent of individual UE mobility behavior. Hence, we impose an assumption on the assigned proportional use of each TAL. We assume that the two TAs apply the same proportional values for the TALs containing both TAs. We generalize the assumption to any TAL, meaning that all cells in each TA within the TAL are fully consistent in their proportional use of the TAL.

To formalize the optimization framework, we denote the proportional usage of TAL ℓ by a non-negative optimization variable x_ℓ. All cells in one TA will apply the same collection of TALs, and they update their UEs with the same proportional use of each TAL. For the network in Figure 1, the four TAs apply their respective collections of TALs by the Equations (6)-(9). Because the usage of any TAL ℓ is unified within the entire service area of ℓ, no matter if the UEs are currently stationary or making HO between cells in ℓ, x_ℓ accurately provides the proportion of users being registered to ℓ among all UEs in the area. Note that these UEs could have received ℓ from any of the TAs in the list.

$$x_1 + x_{12} + x_{13} + x_{14} + x_{123} + x_{124} + x_{134} + x_{1234} = 1 \tag{6}$$

$$x_2 + x_{12} + x_{23} + x_{24} + x_{123} + x_{124} + x_{234} + x_{1234} = 1 \tag{7}$$

$$x_3 + x_{13} + x_{23} + x_{34} + x_{123} + x_{134} + x_{234} + x_{1234} = 1 \tag{8}$$

$$x_4 + x_{14} + x_{24} + x_{34} + x_{124} + x_{134} + x_{234} + x_{1234} = 1 \tag{9}$$

12

The TAU and paging signaling overhead calculation of the overlapping TAL scheme do not require any additional data apart from cell load and HO statistics. For example to calculate the number of TAUs from TA 1 to TA 3 and the opposite direction, the aggregated HO statistics of cells in 1 and 3, denoting them by h_{13} and h_{31} are required. Among the UEs making HO in the direction $1 \rightarrow 3$, only the UEs which do not have 3 in their TAL make a TAU. Therefore, the TAU overhead is $c^u h_{13}(x_1 + x_{12} + x_{14} + x_{124})$, and with the same reasoning the TAU overhead is $c^u h_{31}(x_3 + x_{23} + x_{34} + x_{234})$ for the opposite direction. The total TAU between the two TAs with the overlapping TAL scheme is $c^u h_{13}(x_1 + x_{12} + x_{14} + x_{124}) + c^u h_{31}(x_3 + x_{23} + x_{34} + x_{234})$. As x-variables are all between zero and one, the TAU may attain values in the range $[0, c^u(h_{13} + h_{31})]$. Recall that, the total TAU between the two TAs is either zero or $c^u(h_{13} + h_{31})$ in the conventional TA scheme. As the overlapping TAL scheme is not constrained to satisfy the transitivity property, it has the potential to mitigate the congestion due to TAU signaling overhead [24].

4. TAU congestion mitigation by overlapping TAL optimization

We apply the min-max approach for congestion mitigation, that is, TALs are used to minimize the maximum TAU overhead between all pairs of TAs. We consider a conventional TA layout as the underlying structure. The current TAs of a network is denoted by $\mathcal{T} = \{1, \ldots, T\}$. Each TA $v \in \mathcal{T}$ is a set of cells. The TAs can be, for example, optimized for the total TAU and paging overhead, that is, the corresponding $\boldsymbol{S}$-matrix which leads to the minimum sum of (4) and (5) for each $d \in \mathcal{D}$. The set of all possible TALs of $\mathcal{T}$ are denoted by $\mathcal{L}$, which is the set of all non-empty subsets of $\mathcal{T}$. We use u_v to denote the total load of all cells inside TA $v \in \mathcal{T}$, and h_{vw} the total number of HOs between TAs v and w. These values can be easily aggregated from the load and HO values between cells. Parameter $a_{v\ell}$ is a binary indicator, denoting whether or not TA v is in TAL ℓ. Moreover, n_v denotes the number of cells in TA v. The aforementioned continuous variable x_ℓ represents the proportional use of TAL ℓ by all cells of the

TAs in ℓ. We use auxiliary variable z to denote the maximum number of TAU procedures executed between TAs. The min-max TAU optimization problem can then be modeled by the following linear program (LP).

$$\min \quad z \tag{10a}$$

$$\text{s. t.} \quad \sum_{\ell \in \mathcal{L}} a_{v\ell} x_\ell = 1, \quad v \in \mathcal{T} \tag{10b}$$

$$c^u \Big(h_{vw} \sum_{\ell \in \mathcal{L}} a_{v\ell}(1 - a_{w\ell})x_\ell + h_{wv} \sum_{\ell \in \mathcal{L}} a_{w\ell}(1 - a_{v\ell})x_\ell\Big) \leq z, \quad v, w \in \mathcal{T}, v \neq w \tag{10c}$$

$$\alpha c^p \sum_{\ell \in \mathcal{L}} \sum_{v \in \ell} \sum_{w \in \ell, v \neq w} u_v n_w x_\ell \leq \beta C^P, \tag{10d}$$

$$x_\ell \geq 0, \quad \ell \in \mathcal{L}. \tag{10e}$$

The first constraint set (10b) assures that, within any TA, the proportional usage values of TALs containing in the TA sums up to one. This is the generalization of what was explained in (6)-(9). The second set of constraints (10c) defines the maximum number of TAU procedures executed between all TA pairs. For each ordered pair v and w, if TAL ℓ contains v (i.e., $a_{v\ell} = 1$) but not w (i.e., $a_{w\ell} = 0$), then the TAU overhead between them is $c^u h_{vw} x_\ell$. Constraints (10d) bounds the total addition of paging overhead. For any list ℓ, the additional paging overhead imposed by paging UEs in TA $v \in \ell$ to TA $w \in \ell$ is $\alpha c^p u_v$, scaled by the number of cells in w and the variable x_ℓ that stands for the proportional use of ℓ. An upper bound on the increase of the total paging overhead is necessary, as otherwise the optimum would be to exclusively use the TAL containing all TAs (i.e., as if all cells form one TA). The bound is expressed as βC^P, where C^P is the total paging overhead of TA design $\mathcal{T}$, and β is a non-negative parameter denoting the allowed percentage of increase.

The number of variables in the LP formulation of (10) is exponential in the number of TAs in $\mathcal{T}$. In practice this is not a problematic issue, first due to the fact that (10) is an LP and it can be solved efficiently even for large number of variables. Second, because the region with high risk of excessive TAU is

typically confined to a part of the network.

The x-variables at the optimum of (10) are typically fractional. It means that the UEs moving from one TA to another will partially perform TAU in comparison to the conventional TA scheme. In other words, the sharp boundaries between the TAs are "blurred" by the proportional use of TALs.

5. TAL optimization mechanism in SON

The self-optimization mechanism can combine the static and dynamic framework based on the network behavior. The TAU congestion problem does not exist all the time, but only during certain time intervals. Hence, we can keep the conventional TA scheme until the time when TAU congestion reaches a certain threshold. According to the previous section, the objective of the proposed overlapping TAL is to minimize the maximum TAU overhead between the TA boundaries of the conventional TA scheme. Thus, whenever the auxiliary variable z exceeds the threshold value ξ $(z > \xi)$, we apply the overlapping TAL scheme to reduce this value and avoid the congestion in the network. Figure 2 provides a flowchart for the proposed TAL mechanism in SON.

In case of congested scenario, we assume that the new TAL configuration is received by each UE when it performs a TAU. In our dynamic framework, the load and HO of UEs in each time slot are used for TAL configuration calculation in the MME, therefore the probabilities of assigning each TAL in the MME are independent and more up-to-date based on the UEs mobility compared to the distribution of TAL-assigned UEs.

One challenging issue in dynamically applying the overlapping TAL scheme is making a proper assumption for the parameter β in (10). A relatively high value for β can completely eliminate the congestion problem, while it introduces a large amount of paging overhead in each time interval, whereas β of very small size might not solve the problem of congestion to the limit lower than ξ.

Based on our knowledge distributing a TA/TAL between two different MMEs is not a common practice, therefore the design of TA/TAL is usually defined
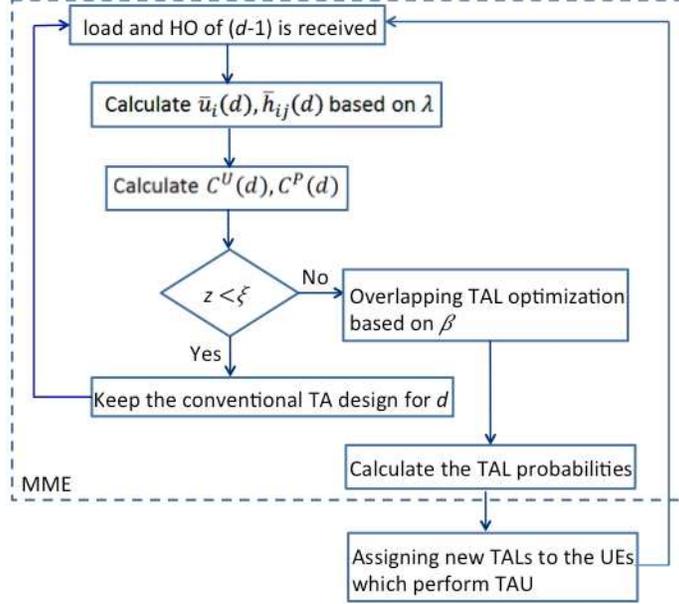
15

Figure 2: The flowchart of TAU congestion mitigation mechanism in SON.

in one MME. Our solution is to have localized TAL optimization procedure at each MME. As the assumption is that there are no TALs distributed in two different MMEs, the global optimization would not achieve a better solution. In this case, whenever a UE moves to another MME, then there will be a TAU procedure.

An alternative in using the overlapping TAL scheme is the fact that the set $\mathcal{T}$ can denote the TAs inside the congested region and not the total TAs of the entire network. Therefore, one mechanism can be that while MME monitors the TAU overheads and determines the congestion area, it can allow only those TAs with high TAU at their boundaries and their neighboring TAs to enter the optimization formulation (10). In this way the problem will significantly reduce size. However, in our study we considered the bigger picture and entered all the TAs of the network inside our optimization model and solved the problem of exponentially higher size. One main reason is that the congestion can happen in several parts of the network at the same time interval and as TAs are geograph-

ically connected to each other, solving the congestion problems individually for small subsets of TAs, may omit the optimum proportional use of TALs which could be found by solving the overall problem.

6. Performance evaluation

The realistic large-scale topology and data traffic of Lisbon down-town area network based on the one provided by the EU MOMENTUM project [26] is used in this study. The network consists of 60 sites and 164 cells. The traffic data corresponds to the load and HO of UEs at different times of day and week. Figures 3(a) and 3(b) illustrate the network topology. The sites are represented by disks. For every site, its cells are denoted by squares. The location of a square in relation to the site center shows the direction of cell antenna. The darkness of each cell is set in proportion to the accumulated cell load. A link is drawn between two cells if there is any HO between them, and the amount of HO is proportional to the thickness of the link. As putting the cells of one site (extended Node B, eNB) in different TAs is not typical, the TA designs are on site level.

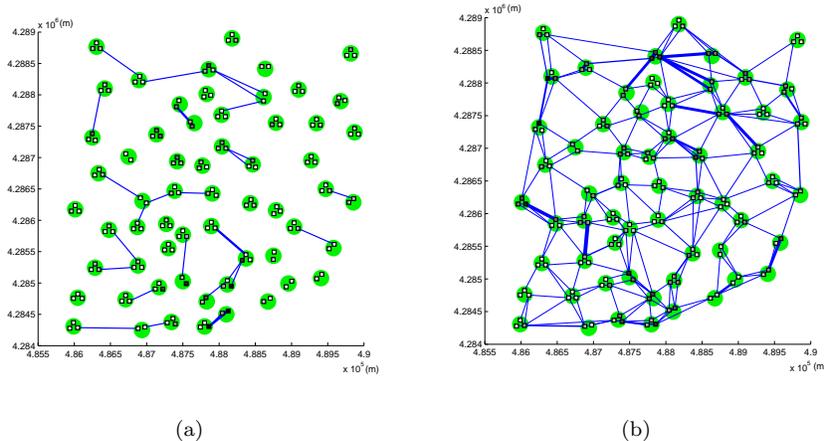(a)   (b)

Figure 3: Illustrations of Lisbon network topology and key statistics (a) traffic of Monday (3:00-4:00) (b) traffic of Friday (17:00-18:00).

Since we deal with a dynamic framework, for each time interval, there exists

17

a set of site load and HO statistics, which are used as an approximation for idle UEs existence and behavior. The time frame of one entire week is considered in the experiment and each time interval is one hour. Hence, the number of data sets D is 168. The traffic varies over different times of each day and different days of the week. This is especially observable when comparing the weekdays to weekends. The UEs' activities are very low during the night and higher during the day with two peaks at the beginning and the end of the office hours. For example, Figure 3(a) shows a low traffic between (3:00-4:00) of Monday morning, and Figure 3(b) shows a high traffic at (17:00-18:00) of Friday afternoon.

In the static TA, which is still the location management deployment in most cellular networks, a conventional TA configuration based on the average data of site load and HO of the entire time period (in our case: one week) is applied to all the data sets. In this study, the optimum conventional TA configuration is computed by CPLEX [14] using the model in [33]. The dashed lines in Figures 4(a) and 4(b) illustrate the total and TAU signaling overhead of the static conventional TA scheme during the examined week. In the entire experiment, the load scale parameter α is set to 0.05, and the overhead of a single update c^u is set ten times as much as c^p [16]. In the calculations, c^u is set to 1 cost unit, and hence c^p to 0.1 cost unit. The values presented on the y-axis of all the forthcoming diagrams are all based on this cost unit, and due to having the same unit, they are comparable. Table 2 summarizes all the numerical assumptions and values considered in the following performance evaluation study.

To apply SON on the standard TA scheme, the data is calculated according to the exponential filters explained in Section 3.1. This data is used to compute the conventional TA design of each time interval. However in SON, the TA layout is applied to the network's data of the next time interval. The reason for having this assumption is because in reality there is no possibility of going back in time and applying the optimum design to the same set of data. Note that the optimum design based on the history of UEs' behaviors might not perform well for the next time interval. The λ value in this study is set to 0.75, this means that the current time interval has three times more impact than the aggregated

18

Table 2: Numerical assumptions and values for performance evaluation.

Parameter	Description	Values
D	Number of data sets	168
N	Number of sites	60
T	Number of TAs	[11 - 30]
c^p	Cost of a single paging	0.1
c^u	Cost of a single update	1
α	Load scale parameter	0.05
β	Allowed percentage increase on the total paging	[0.05,0.1]
λ	Smoothing factor	0.75
ξ	Paging threshold value	100

data of the previous time intervals. Larger values of λ actually reduce the level of smoothing, and in case of $\lambda = 1$ the data applied is just the same as the original data $u_i(d-1)$ and $h_{ij}(d-1)$ for calculating the actual signaling overhead. While values of λ closer to zero have a greater smoothing effect and are less responsive to recent changes. The solid lines in Figures 4(a) and 4(b) illustrate the total and TAU signaling overhead of the SON conventional TA scheme during the examined week.

Figure 4(a) compares the total signaling overhead of static and SON conventional TA scheme during the examined week. The key statistics of the TA design are as follows. There are 26 TAs in the static design, whereas in the SON designs the number of TAs in different time intervals varies from a minimum value of 11 to a maximum of 30. The total signaling overhead of the static design during one week is $\sum_{d=1}^{D}(C_{static}^{U}(d) + C_{static}^{P}(d)) = 2.6304 \times 10^5$, while in SON it has a 35.9% lower value, $\sum_{d=1}^{D}(C_{SON}^{U}(d) + C_{SON}^{P}(d)) = 1.6864 \times 10^5$. According to Figure 4(b) which compares the TAU overhead of static and SON conventional TA scheme during the examined week, we can conclude that the improved total signaling overhead of dynamic SON framework over the

19
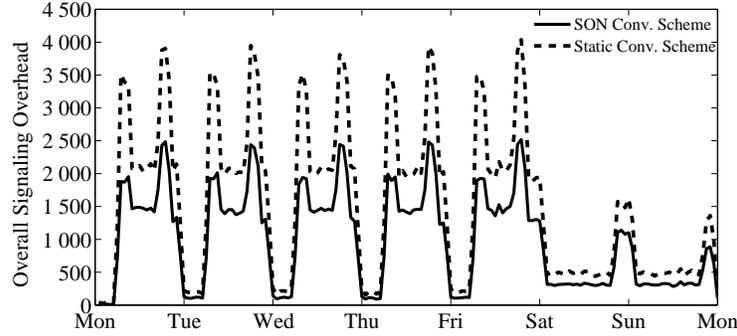
(a)
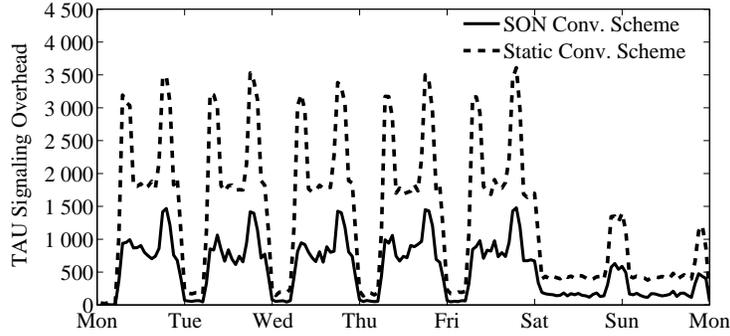
(b)

Figure 4: Performance comparison of static and SON conventional TA scheme (a) total signaling overhead (b) TAU signaling overhead of one-week data.

static one is a result of TAU overhead improvement. The total TAU overhead of the week is $\sum_{d=1}^{D} C_{static}^{U}(d) = 2.3283 \times 10^5$ for the static TA, and $\sum_{d=1}^{D} C_{SON}^{U}(d) = 9.0235 \times 10^4$ for SON, that is 61.24% reduction.

The next step is to evaluate the problem of TAU congestion while a conventional TA design based on minimization of total signaling overhead is applied in the network. For the same static and SON frameworks explained above, we have computed the maximum and average TAU overhead at TA boundaries of the conventional TA design for each time interval. Figure 5(a) shows the maximum value of TAU among all TA boundaries for both static (dashed line) and
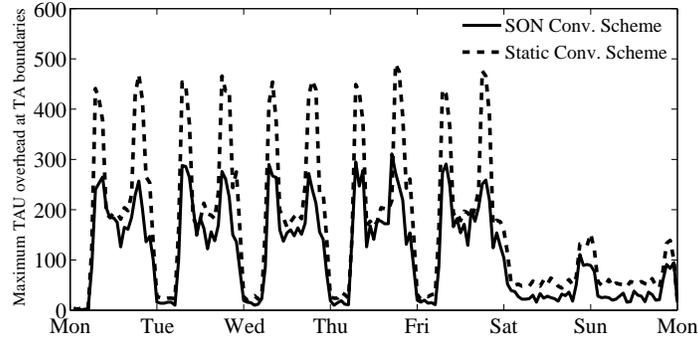
(a)
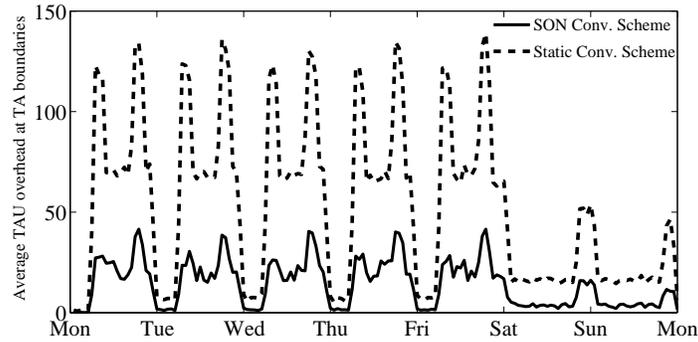
(b)

Figure 5: TAU Congestion condition of static and SON conventional TA scheme (a) Maximum TAU (b) Average TAU at TA boundaries during one-week data.

SON (solid line) frameworks. Although from the figure it is observable that the static framework has higher TAU congested TAs compared to SON in the whole time frame, the difference is more highlighted for two peaks in the morning and afternoon of the working days.

Figure 5(b) provides an illustration of the average TAU overhead at TA boundaries for the two frameworks. This explains that the SON conventional TA scheme outperforms the static TA mostly on reducing the average TAU overhead, however the problem of congested TAs still remains while we compare Figures 5(a) and 5(b). This motivates us to apply our overlapping TAL scheme

21

in order to mitigate the congestion situations.

6.1. Applying overlapping TAL for congestion mitigation

In this section our aim is to mitigate the TAU congestion in SON conventional TA scheme presented with solid line in Figure 5(a). Based on this figure, we observe that the problem of TAU congestion does not exist in all time intervals. We consider the threshold value $\xi = 100$, and hence we only apply the overlapping TAL scheme to data of time intervals in which the maximum TAU at the TA boundaries is higher than this threshold value.
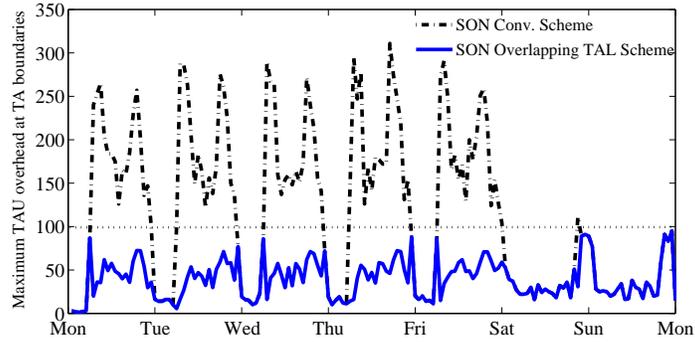
(a)
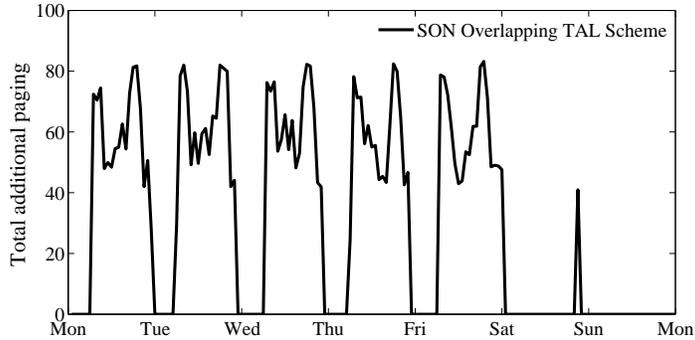
(b)

Figure 6: (a) TAU congestion mitigation by applying overlapping TAL scheme (b) Total additional paging ($\beta = 0.1$)

22

For each time interval, formulation (10) in Section 4 can be solved with different limit parameter β ranging from zero (i.e., the underlying TA design is kept) to the maximum possible in which $z = 0$ (i.e., all the TAs will belong to a list containing all of them). We assume $\beta = 0.1$, meaning that the model allows 10% increase in the total paging of each time interval in order to minimize the TAU congestion.

Figure 6(a) compares the TAU congested scenario of SON conventional TA scheme with the uncongested scenario while the overlapping TAL scheme is applied to the network. The horizontal line at 100 shows which time intervals where congested and hence we have applied the optimization model only to those time intervals. From the figure, we can observe that with $\beta = 0.1$ the maximum TAU remains well below the threshold value $\xi = 100$.

Our aim is to avoid any TAU congestion but not to reduce the TAU overhead as much as possible. Note that we are adding additional paging in each time interval by introducing the overlapping TAL concept. Figure 6(b) illustrates the amount of additional paging for each time interval for avoiding congestion. The total additional paging for the whole week is $\sum_{d=1}^{D} C_{additional}^{P}(d) = 5.1947 \times 10^3$ for $\beta = 0.1$.

In the final step of our performance evaluation, we decrease the value of β to 0.05 in order to see if the overlapping TAL scheme is able to mitigate any TAU congestion. Figure 7(a) compares the TAU congested scenario of SON conventional TA scheme with the SON overlapping TAL scheme while $\beta = 0.05$. The figure shows that the overlapping TAL scheme is able to mitigate the TAU congestion problem in most of the time intervals, and for the time intervals which still have congestion, the maximum TAU overhead at the boundaries has been significantly reduced. From Figure 7(b), we can observe that the total additional paging in the case where $\beta = 0.05$ is only $\sum_{d=1}^{D} C_{additional}^{P}(d) = 2.5973 \times 10^3$.

From the numerical results, we can conclude that our proposed overlapping TAL scheme is highly able to mitigate the TAU congestion in a SON dynamic framework. To fully mitigate the TAU congestion problem while not adding extra additional paging, SON can dynamically change the value of β based on the

23
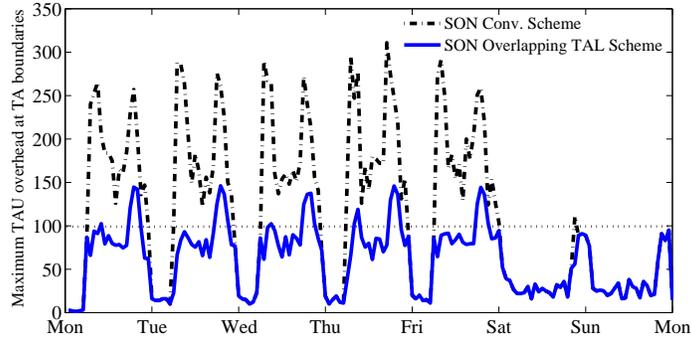
(a)
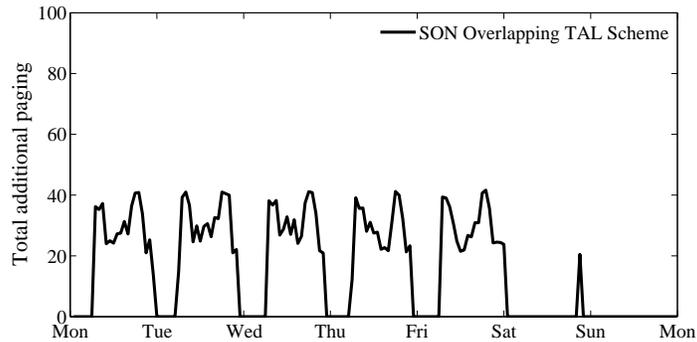
(b)

Figure 7: (a) TAU congestion mitigation by applying overlapping TAL scheme (b) Total additional paging ($\beta = 0.05$)

amount of congestion in different time intervals. Nevertheless, the performance of a fixed β is evidently satisfactory.

7. Conclusions

We proposed an automated process in location management of LTE networks within the paradigm of self-optimization networks. We tackled the problem of mitigating TAU signaling congestion in a dynamic framework by using overlapping tracking area lists (TALs) and a min-max optimization approach. The TAU and paging statistics are monitored and the proportional use of TALs are

adapted based on a linear program optimization formulation. Numerical evaluations for mitigating massive TAU demonstrates that the proposed dynamic optimization framework is very effective. Our study was limited in having the data of only one week, however by introducing more history of network's behavior, SON can significantly improve the network performance. By re-distributing the signaling overhead over the network area, TALs open up a new dimension of freedom in location management.

One extension line of this research is to build the overlapping TAL framework compatible for both minimization of the total signaling overhead, and mitigation of the signaling congestion depending on the priority of either problem in the network. Reasonable thresholds must be computed for triggering the overlapping TAL assignments according to the two problems.

8. Acknowledgments

References

[1] 3GPP TS 23.002, Release 9, version 9.1.0, Network architecture, 2009.
   http://www.3gpp.org/ftp/specs/archive/23_series/23.002/.

[2] 3GPP TS 23.401, Release 8, version 8.6.0, General Packet Radio Services (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, 2009.
   http://www.3gpp.org/ftp/specs/archive/23_series/23.401/.

[3] 3GPP TR 23.880, Release 8, version 0.1.0, Stage2 for registration in densely-populated area (RED), 2007.
   http://www.3gpp.org/ftp/specs/archive/23_series/23.880/.

[4] 3GPP TS 32.500, Release 8, version 8.0.0, Self organizing networks (SON), concepts and requirements, 2008.
http://www.3gpp.org/ftp/specs/archive/32_series/32.500/.

[5] 3GPP TS 36.902, Release 8, version 1.0.1, Self-configuring and self-optimizing network use cases and solutions, 2008.
http://www.3gpp.org/ftp/specs/archive/36_series/36.902/.

[6] I.F. Akyildiz, J.S.M. Ho, Y.B. Lin, Movement-based location update and selective paging for PCS networks, ACM/IEEE Transactions on Networking 4 (1996) 629–638.

[7] I.F. Akyildiz, J. McNair, J.S.M. Ho, H. Uzunalioglu, W. Wang, Mobility management in next generation wireless systems, in: Proceedings of IEEE 87 (1999) 1347–1384.

[8] Y. Bejerano, M.A. Smith, J. Naor, N. Immorlica, Efficient location area planning for personal communication systems, ACM/IEEE Transaction on Networking 14 (2006) 438-450.

[9] A. Buvaneswari, L.M. Drabeck, N. Nithi, M. Haner, P.A. Polakos, C. Sawkar, Self-optimization of LTE networks utilizing Celnet Xplorer, Bell Labs Technical Journal 15:3 (2010) 99-118.

[10] E. Cayirci, I.F. Akyildiz, Optimal location area design to minimize registration signaling traffic in wireless systems, IEEE Transactions on Mobile Computing 2 (2003) 76-85.

[11] Y.W. Chung, Adaptive design of tracking area list in LTE, in: Proceedings of 8th IEEE International Conference on Wireless and Optical Communications Networks (WOCN), Paris, 2011.

[12] I. Demirkol, C. Ersoy, M.U. Caglayan, H. Delic, Location area planning and cell-to-switch assignment in cellular networks, IEEE Transactions on Wireless Communications 3 (2004) 880-890.

[13] P. Garcia, V. Casares, J. Mataix, Reducing location update and paging costs in a PCS network, IEEE Transactions on Wireless Communications 1 (2002) 200-209.

[14] ILOG CPLEX 12.2, User's Manual, 2010.

[15] J. Ferragut, J. Mangues-Bafalluy, A self-organized tracking area list mechanism for large-scale networks of femtocells, in: Proceedings of International Conference on Communications (ICC), Ottawa, 2012.

[16] K. Kyamakya, K. Jobmann, Location management in cellular networks: classification of the most important paradigms, realistic simulation framework, and relative performance analysis, IEEE Transactions on Vehicular Technology 54 (2005) 687-708.

[17] T. Mach, R. Tafazolli, Mass mobility signaling congestion avoidance mechanism using randomized time distribution of cell reselections, in: Proceedings of International Conference on Telecommunications (ICT), Marrakesh, 2009, pp. 238-242.

[18] Mitsubishi Electric, Tracking areas sizes and tracking area list optimization, Technical Report in 3GPP TSG RAN WG3 Meeting, R3-071931, 2007.

[19] S. Modarres Razavi, D. Yuan, Performance improvement of LTE tracking area design: a re-optimization approach, in: Proceedings of the 6th ACM International Workshop on Mobility Management and Wireless Access (MobiWac '08), pp. 77-84, 2008.

[20] S. Modarres Razavi, D. Yuan, F. Gunnarsson, J. Moe, Performance and cost trade-off in tracking area reconfiguration: a Pareto-optimization approach, Journal of Computer Networks 56 (2012) 157-168.

[21] S. Modarres Razavi, D. Yuan, F. Gunnarsson, J. Moe, Exploiting tracking area list for improving signaling overhead in LTE, in: Proceedings of IEEE Vehicular Technology Conference (VTC Spring), Taipei, 2010.

[22] S. Modarres Razavi, D. Yuan, F. Gunnarsson, J. Moe, Dynamic tracking area list configuration and performance evaluation in LTE, in: Proceedings of IEEE Global Communications Conference Workshop (GLOBECOM Workshop), Miami, 2010.

[23] S. Modarres Razavi, D. Yuan, Mitigating mobility signaling congestion in LTE by overlapping tracking area lists, in: Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM), Miami, 2011, pp. 285-292.

[24] S. Modarres Razavi, D. Yuan, Mitigating signaling congestion in LTE location management by overlapping tracking area lists, Journal of Computer Communications 35 (2012) 2227-2235.

[25] S. Modarres Razavi, D. Yuan, Reducing signaling overhead by overlapping tracking area list in LTE, Proceedings of the 7th IFIP Wireless and Mobile Networking Conference (WMNC), Vilamoura, 2014.

[26] IST-2000-28088 Momentum project. http://momentum.zib.de.

[27] Motorola, Long Term Evolution (LTE): A technical overview, Technical White Paper, 2007.

[28] G.P. Pollini, I. Chih-Lin, A profile-based location strategy and its performance, IEEE Journal on Selected Areas in Communications 15 (1997) 1415-1424.

[29] R. Prakash, Z. Haas, M. Singhal, Load-balanced location management for cellular mobile systems using quorums and dynamic hashing, Wireless Networks 7 (2001) 497-512.

[30] NEC, Tracking area concepts and influence on paging load, Technical Report in 3GPP TSG RAN2 Meeting, R2-070655, 2007.

[31] QUALCOMM Europe and T-Mobile, New SON user case: tracking area optimization, Technical Report in 3GPP TSG RAN WG3 Meeting, R3-071594, 2007.

[32] S. Subramaniam, G. Krishnamurthi, Load balancing location management, in: Proceedings of IEEE International Conference on Communications (ICC), Helsinki, 2001, pp. 2835-2839.

[33] D.W. Tcha, T.J. Choi, Y.S. Myung, Location-area partition in a cellular radio network, Journal of the Operational Research Society 48 (1997) 1076-1081.

[34] T. Winter, Mobility management and network design for UMTS, in: Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2004.

[35] W. Wang, I.F. Akyildiz, G.L. Stüber, B. Chung, Effective paging schemes with delay bound as QoS constraints in wireless systems, Wireless Networks 7 (2001) 455-466.

[36] V.W.S. Wong, V.C.M. Leung, Location management for next generation personal communication networks, IEEE Network 14 (2000) 18-24.

[37] Y. Zhang, M. Fujise, Location management congestion problem in wireless networks, IEEE Transactions on Vehicular Technology 56 (2007) 942-954.

[38] ZTE, Self optimization for tracking areas, Technical Report in 3GPP TSG RAN WG3 Meeting, R3-071415, 2007.