Efficient Tracking Area Management Framework for 5G Networks

Miloud Bagaa, Tarik Taleb, *Senior Member, IEEE*, and Adlen Ksentini, *Senior Member, IEEE*

***Abstract*—One important objective of 5G mobile networks is to accommodate a diverse and ever-increasing number of user equipment (UEs). Coping with the massive signaling overhead expected from UEs is an important hurdle to tackle so as to achieve this objective. In this paper, we devise an efficient tracking area list management (ETAM) framework that aims to find optimal distributions of tracking areas (TAs) in the form of TA lists (TALs) and assigning them to UEs, with the objective of minimizing two conflicting metrics, namely paging overhead and tracking area update (TAU) overhead. ETAM incorporates two parts (online and offline) to achieve its design goal. In the online part, two strategies are proposed to assign in real time, TALs to different UEs, while in the offline part, three solutions are proposed to optimally organize TAs into TALs. The performance of ETAM is evaluated via analysis and simulations, and the obtained results demonstrate its feasibility and ability in achieving its design goals, improving the network performance by minimizing the cost associated with paging and TAU.**

***Index Terms*—5G, LTE, convex optimization, game theory.**

I. Introduction

ONE OF the main challenges of the upcoming 5G networks is to accommodate the high demand of data raised from the increasing number of devices. In this vein, deploying small cells should be considered with high interest to overcome this issue. $5G$ networks would deploy densely self-organizing low-cost and low power small base-stations. However, deploying high number of small cells would increase the signaling overhead caused by the tracking and paging of User Equipment (*UE*). Combined with the high number of *UEs* and Machine Type Communication (MTC) devices [1], [2], the use of small cells will introduce a major challenge in term of signaling overhead for $5G$ networks. In order to tackle the increased data rate expected from the usage of the envisioned $5G$ network, the signaling overhead should be minimized as much as possible.

Usually, the Radio Access Network (RAN) of a mobile operator is organized into a set of cells (including small cells) that covers several geographical areas. *UEs* in a specific area are attached to a base station (eNodeB), which manages their access to the mobile core network. *UEs* are usually in idle mode and have no call activity for some duration. When a connection request comes for a *UE* in idle mode, the Mobility Management Entity (*MME*) sends a signaling message, namely paging, to all eNodeBs to find the *UE*'s location (i.e., cell) in the network. Accordingly, in case a high number of *UEs* need to be paged, a massive number of downlink signaling messages have to be transmitted, resulting in high signaling overhead and wasting scarce resources of the mobile network. To overcome this issue, the Tracking Area (*TA*) concept has been introduced in Release 8 of the 3GPP mobile network specifications (i.e., replacing the Routing Area concept in previous releases). The key idea beneath the *TA* principle consists in grouping several cells or sites into one *TA*. *MME* keeps record of the location of *UEs* in idle mode at the *TA* granularity. Thus, when a connection setup request comes for a *UE* in idle mode, the *UE* in question is paged only within its current *TA*, which would mitigate the overhead of paging in the network.

Each time a *UE* moves to a new location and connects to a new cell not belonging to its current *TA*, the *UE* sends an uplink message, namely Tracking Area Update (*TAU*), to *MME*, which subsequently updates the *TA* of the *UE*. In this vein, it is worth noting that a *TA* is also defined as an area where the *UE* can move without transmitting *TAU* messages to *MME*. Despite the advantages of the *TA* concept in minimizing the paging overhead, it has the following limitations on the *TAU* signaling: (i) many *TAU* signaling messages might be generated due to ping-pong effect, i.e, a *UE* keeps hopping between two adjacent cells belonging to different *TAs*, which could be exacerbated in case of densely deployed small cells; (ii) the mobility signaling congestion due to a large number of *UEs* having a similar behavior, e.g. massive number of *UEs* simultaneously moving from one *TA* to another *TA* (train scenario); (iii) the use of *TA* strategy has the symmetry limitation: If two cells are in the same *TA*, then neither of them can be in any other *TA*. To overcome this limitation, Release 12 introduces the Tracking Area List (*TAL*) concept in order to simplify the *TA* configuration. The *TAL* concept aims for reducing the TAU signaling messages by grouping several *TAs* in one *TAL* and allowing the overlapping of *TAs*. Each time a *UE* visits a new *TA* that does not belong to its *TAL*, a *TAU* message is sent to the *MME*. Upon receiving the *TAU* message, *MME* assigns a new *TAL* to the *UE*. The new *TAL* should include the visited *TA*. Furthermore, Release 12 allows network operators to include up to 15 *TAs* in each *TAL* and the *MME* always adds the last visited *TA* to the list to overcome the problem of frequent updates due to ping-pong situations. Given

Manuscript received May 23, 2015; revised October 3 and December 28, 2015; accepted February 13, 2016. Date of publication February 26, 2016; date of current version June 7, 2016. This work was supported by the TAKE 5 project funded by the Finnish Funding Agency for Technology and Innovation (TEKES), a part of the Finnish Ministry of Employment and the Economy. The associate editor coordinating the review of this paper and approving it for publication was W. Saad.

M. Bagaa is with Aalto University, Espoo, Finland (e-mail: miloud.bagaa@aalto.no).

T. Taleb is with Sejong University and Aalto University, Espoo, Finland (e-mail: taleb-tarik@ieee.org).

A. Ksentini is with EURECOM, Sophia-Antipolis, France (e-mail: adlen.ksentini@eurecom.fr).

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

Digital Object Identifier 10.1109/TWC.2016.2535217

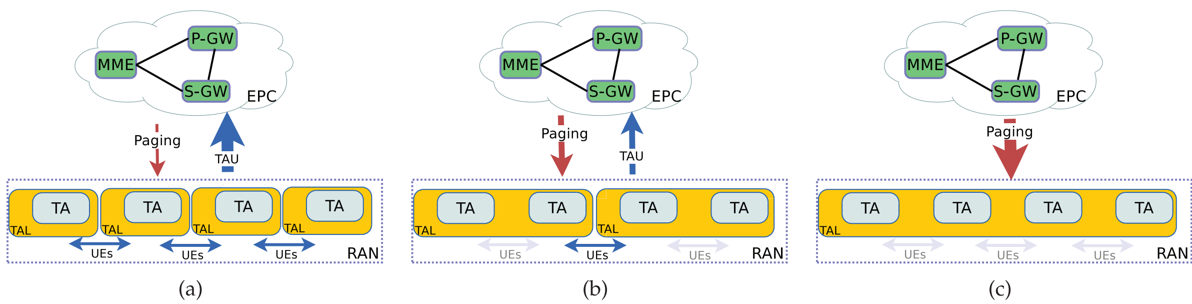


Fig. 1. The tradeoff between *TAU* and paging overhead in 4*G* and beyond mobile networks.

that *TALs* are overlapped, the above-mentioned limitations of conventional *TAs*, defined in Release 8, can be accordingly mitigated. However, the current LTE specifications do not provide any details on how to define *TALs* and allocate them to *UEs*.

Each time a *UE* moves to a new location and connects to a new *TA* not belonging to its current *TAL*, the *UE* sends a *TAU* message to *MME*. On the other had, when a connection request comes for a *UE*, the *MME* sends a paging message to all *TAs* (i.e., *TAL*) where the *UE* is registered. An increase in *TALs* size leads to a rise in paging signaling messages and a decrease in *TAU* signaling messages. Fig. 1 shows the tradeoff between *TAU* and paging overheads when forming *TALs*. In the figure, we assume that the network contains four *TAs* along a railway path, in which each *TA* has two other neighboring *TAs* on the left and the right sides. From Fig. 1(a), we observe that the organization of each *TA* in a separate *TAL* causes many *TAU* signaling messages in the network, which are generated and forwarded from the RAN to the evolved packet core (EPC). Whereas Fig. 1(b) and Fig. 1(c) show that increasing *TAL* size reduces *TAU* overhead and increases paging overhead. Fig. 1(c) shows that the *TAU* overhead can be ignored if all *TAs* are organized in the same *TAL*.

Several research works have been conducted to solve the *TAL* problem, whereby the aim is to capture the tradeoff that mitigates the overhead of *TAU* and paging messages when constructing and assigning *TALs* to *UEs*. Most of these solutions formulate the problem using a multi-objectives optimization technique to achieve a fair tradeoff between signaling messages overhead of *TAL* and paging, i.e. minimize both signaling messages due to *TAU* and paging. In this paper, we devise an efficient tracking area list management (*ETAM*) framework for 5G cloud-based mobile networks [3], [4]. The proposed framework consists of two independent parts. The first part is executed offline and is responsible of assigning *TAs* to *TALs*, whereas the second one is executed online and is responsible of the distribution of *TALs* on *UEs* during their movements across *TAs*. For the first part, we propose three solutions, which are: (a) *F-PAGING* favoring the paging overhead over *TAU*, (b) *F-TAU* favoring *TAU* over paging, and (c) *FOTA* (i.e., Fair and Optimal Assignment of *TALs* to *TAs*) for a solution that uses bargaining game to ensure a fair tradeoff between *TAU* and paging overhead. For the second part, two solutions are proposed to assign *TALs* to *UEs*. The computation load is kept lightweight in both solutions not to downgrade the network performance. Furthermore, both solutions do not require any additional new messages when assigning *TALs* to *UEs*. The first solution takes into account only the priority between *TALs*. As for the second one, in addition to the priority between *TALs*, it takes into account the *UEs* activities (i.e., in terms of incoming communication frequency and mobility patterns) to enhance further the network performance.

The remainder of this paper is organized as follows. Section II introduces some related research work. Section III presents the envisioned network model and formulates the target problem. It also presents an overview of the *ETAM* framework. Section IV presents the online part of the *ETAM* framework for assigning *TALs* to *UEs*. The three solutions proposed for the offline part of the *ETAM* framework are described in Section V. Section VI details a Markov-based analytical model for the three offline solutions. Besides the numerical results obtained by solving the Markov model, Section VII presents the simulation setup to evaluate the performance of *ETAM* and discusses the obtained results. Finally, the paper is concluded in Section VIII.

II. Related Work

Mitigating signaling overhead, due to *UE* mobility in cellular mobile networks, has attracted high attention during the last years. As stated earlier, in the Evolved Packet System (*EPS*), *MMEs* keep records of *UEs*' positions in order to adequately forward their relevant incoming connections. For this purpose, 3GPP introduced two types of signaling messages to support *UE* mobility: (i) paging messages from the network, namely *MME*, in order to find the locations of *UEs* in idle mode; (ii) *TAU* messages from *UEs* to *MME* to update their positions. A *TAU* message is sent each time a *UE* enters into a new location (cell) that does not belong to its current *TA*. Conventional *TA* assignment procedures whereby the network assigns only one *TA* for different *UEs* is not sufficient when *UEs* are highly mobile. Indeed, high number of *TAU* messages could be sent by *UEs* as they frequently cross their corresponding *TA* borders. An enhancement to the conventional procedure was envisioned to reduce *TAU* overhead by i) grouping several cells (i.e., eNodeBs) in one *TA* or ii) introducing delays between *TAU* messages sent by *UEs*. Another solution to reduce the impact of *TAU* messages on the network was proposed in [5] whereby queuing models and buffer information at eNodeBs are used to delay the *TAU* frequency.

To further alleviate the effect of *TAU* messages on the network performance, 3GPP has introduced the concept of *TAL* in Long Term Evolution (LTE), wherein each cell

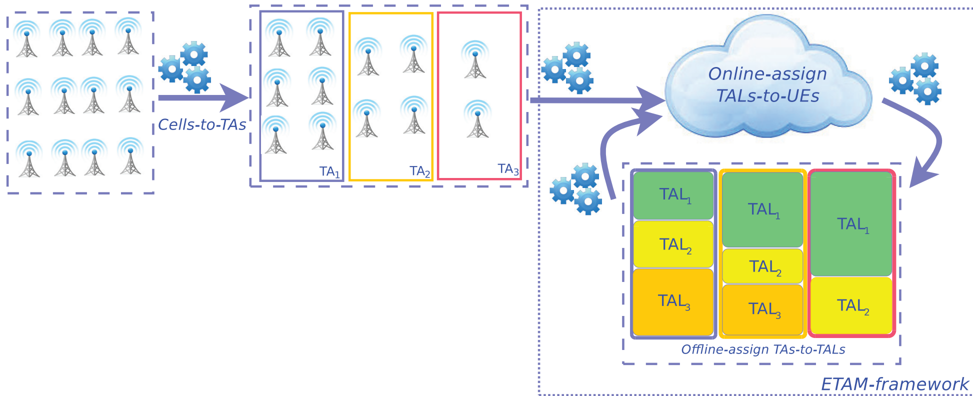


Fig. 2. The proposed framework for tackling *TAU* and paging overhead in 4*G* and beyond mobile networks.

(eNodeB) assigns different *TALs* to *UEs* [6], [7]. Since *TALs* are overlapped, the number of *UEs* performing *TAU* when crossing *TA* border drastically decreases. Besides reducing the number of *TAU* messages, *TAL* prevents the ping-pong effect, i.e., frequent *TAU* messages when a *UE* keeps hopping between adjacent *TAs*. Nevertheless, the current LTE specifications do not provide any details on how to define *TALs* and allocate them to *UEs*. To address this open issue, several solutions have been proposed. In [8], Chung et. al. proposed a solution that organizes cells into rings, where *UEs* in each ring use the same *TAL*. Solutions, proposed in [9] and [10], use the same concept as in [8] by assigning the same *TAL* to different *UEs* when visiting a cell in the network. However, all these solutions [8]–[10] have not fully explored the advantage of *TAL* against the conventional *TA* approach. In [7] and [11], Razavi et. al. overcome this limitation by allowing *UEs* residing in the same cell to register with different *TALs*. Indeed, in [7] they proposed a solution for congestion mitigation along a railway path. On the other hand, in [11] an extension of the former work is proposed with two new aspects: i) the solution is generalized for any arbitrary network instead of only train scenario; ii) a new solution that handles the extenuation of paging signaling messages via *TAL* management is proposed.

Generally speaking, assigning *TALs* to *UEs* shall depend on the mobility patterns of *UEs* as well as on their geographical distribution and density. *MME* may group, under the same *TAL*, a large number of *TAs* in an area that has low density to reduce the impact of *TAU* overhead on the network performance. Similarly, *MME* may group under the same *TAL* a small number of *TAs* serving a highly densed area. Indeed, to alleviate the impact of paging messages on the network performance, it is worth assigning more than one *TAL* to the same *TA*. To the best knowledge of the authors, most existing solutions focus only on the offline part for assigning the *TAs* to *TALs*. Moreover, they consider only the *TAU* overhead and ignore the paging overhead. The only research work that addressed both constraints is presented in [11], wherein Razavi et al. proposed two separate solutions, addressing the impact of *TAU* and paging overhead, respectively. Both solutions are based on multi-objectives optimization techniques for assigning the *TAs* to *TALs*. The first one tries to minimize the *TAU* overhead while setting paging as a constraint, and the second one minimizes the paging overhead while fixing the *TAU* overhead as a constraint.

In contrast to the existing works, in this paper, we propose a framework optimizing the management of *TALs* and consisting in: (i) an offline part that assigns *TAs* to *TALs*; (ii) an online part that assigns *TALs* to *UEs*. Two solutions are proposed to achieve the aim of the online part. The first one takes into account only the priority between *TALs*, whereas the second one, in addition to the priority between *TALs*, takes into account the *UE* behavior in terms of mobility and connection frequency. Regarding the offline part, we have devised three solutions, which differ from the existing ones on their way to cope with the problem. Indeed, most existing solutions assign the same *TAL*: i) to the same *TAs* in a static manner [8]–[10]; or ii) with the same probability [7], [11]. In contrast, the devised solutions dynamically assign the same *TAL* to different *TAs* with different probabilities. The first one, dubbed *F-PAGING*, is proposed for a network known with a high rate of paging (i.e., for voice call as well as for IP-based web applications) in comparing to the mobility rate. This solution maybe designated for small cities with high-density populations. The second one, dubbed *F-TAU*, is proposed for a network which is known with a high mobility rate compared to the paging rate. Such kind of solution maybe useful for a network known with low-density populations and/or high mobility. The last one, dubbed *FOTA*, is proposed to be generic for any kind of networks. It takes advantage of both previous solutions, jointly addressing the overhead due to both *TAU* and paging messages. *FOTA* uses Nash bargaining game to ensure a fair tradeoff between both conflicting overhead, i.e., *TAU* and paging signaling messages.

III. Envisioned Network Model and Framework Overview

A. ETAM Framework Overview

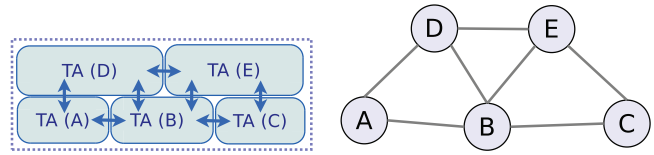
Fig. 2 depicts a general overview of the *ETAM* framework. We assume that the network is subdivided into N *TAs*, $\mathcal{N} = \{1, 2, \cdots N\}$. Each *TA* consists of a set of cells, whereby a cell is managed by an eNodeB (i.e., base station). As depicted in the figure, the geographically close eNodeBs can be grouped in the same *TA*, using any existing algorithm [12], [13], to optimize the network performance in terms of paging overhead. Initially, the *ETAM* framework starts by an inefficient solution and then converges, through iterations, to the optimal one. As depicted in Fig. 2, *ETAM* framework starts by considering each *TA* as a separated *TAL*. Then it executes, repetitively, two steps to converge to the optimal solution. The first step is the offline-assignment of *TAs-to-TALs*, whereas the second one is the online-assignment of *TALs-to-UEs*. To efficiently map between *TAs* and *TALs*, the information about *TAU* and paging signaling messages are transferred from the online step to the offline one. The latter enhances the mapping between *TALs* and *TAs* and then provides the former with the new mapping to optimize further the network performance. The online step is executed during a specified period D, where all the information about the *TAU* and paging overhead are gathered from the network to be transferred to the offline step. The duration D may be fixed by the network operator, but it can be changed when there is a noticeable update in the network.

Since there is no exact indication on the trajectory of *UEs*, during the online-assignment of *TALs-to-UEs*, we use a probability strategy to assign *TALs* to *UEs*. In each visited *TA*, *TALs* are assigned to visiting *UEs* with different probabilities. Indeed, the *TAL* that reduces more the *TAU* and paging signaling messages would have more priority to be assigned to a *UE*. There is a tradeoff between *TAU* and paging signaling messages. Clearly, the smaller the size of *TALs* is, the higher the *TAU* overhead is, but the smaller the paging overhead becomes. For the online-assignment of *TALs-to-UEs*, we consider two solutions. The first one takes into account only the priority between *TALs* that was learned from the offline step. Whereas, the second one, in addition to the priority between *TALs*, takes into account the *UEs* behavior, in terms of incoming communication frequency and mobility patterns. For the offline-assignment of *TAs-to-TALs*, we consider three different solutions, which define the core of our *ETAM* framework. It is worth recalling that (i) the first solution favors the paging overhead when forming *TALs*; (ii) the second one favors the *TAU* overhead; and (iii) the third solution uses the bargaining game theory to distribute *TALs* among *TAs* by capturing a fair tradeoff between *TAU* and paging overhead. The *TAL* that exhibits the highest fairness in the *TAU* and paging overhead has the highest probability to be assigned to a *UE*.

B. Network Model and Notations

Let Γ denote the set of all possible *TALs* in a mobile network, and let Γ_A denote the set of possible *TALs* that can be assigned to *UEs* in *TA* A. As mentioned earlier, each time a *UE* visits a new *TA* that does not belong to its *TAL*, a *TAU* message is sent to the *MME*. Upon receiving the *TAU* message, *MME* computes and sends a new *TAL* to the *UE*. The new *TAL* should include the visited *TA*. From Release 12 of the 3GPP specifications, the operator can specify for each *TAL* a list of up to 15 *TAs* and the *MME* always adds the last visited *TA* to the list to prevent the risk of ping-pong updates. For this reason, Γ is formed by considering the different possible combinations of *TAs*, such that the length of each element in Γ should be higher or equal to one and less than 16, i.e. each *TAL* $i \in \Gamma$ should contain at least 1 *TA* and at most 15 *TAs* to allow the *MME* to add the last visited *TA*.

Throughout the paper, we will refer to the example depicted in Fig. 3 in order to show how Γ should be constructed. In this example, we assume that the network consists of five *TAs*, named A, B, C, D and E. The blue arrows between *TAs* denote the movement of different *UEs* in the network. The movement of UEs can be deduced from the handover statistics of different eNodeBs or from the handover command messages sent by *MME*. To form Γ, we begin by forming the neighboring graphs G from the network as depicted in Fig. 3(b). An edge between two vertices (i.e., *TA*) A and B exists, if there is a *TAU* possibility between them. In Fig. 3(b), an edge is generated between the vertices A and B, if there is a blue arrow between *TAs* A and B in Fig. 3(a), which means the possibility of *UEs* movement between these *TAs*. In Fig. 3(b), we do not construct an edge between vertices A and E since a direct blue arrow does not exist between them; *UEs* cannot move from A to E without passing by another *TA* (i.e., B or D). Finally, Γ_A is formed from the neighboring graph G. Indeed, the different elements of Γ_A are those having all vertices of all sub-graphs of G that contain the vertex A and their length do not exceed 15. Thus, the vertices of a sub-graph of G that contain the vertex A are considered as one element in Γ_A. From Fig. 3, $\Gamma_A = \{\{A\}, \{A, B\}, \{A, D\}, \{A, B, C\}, \{A, B, D\}, \{A, B, E\}, \{A, D, E\}, \{A, B, C, D\}, \{A, B, C, E\}, \{A, B, D, E\}, \{A, B, C, D, E\}\}$. Finally, Γ is formed from different Γ_i as follows: $\Gamma = \bigcup_{i \in \mathcal{N}} \Gamma_i$. An element of Γ_i is a set, i.e. $\{A, B\}$ and $\{B, A\}$ are considered as the same element in Γ. From Fig. 3, $\Gamma = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{A, B\}, \{A, D\}, \{B, C\}, \{B, D\}, \{B, E\}, \{C, E\}, \{D, E\}, \{A, B, C\}, \{A, B, D\}, \{A, B, E\}, \{A, D, E\}, \{B, C, D\}, \{B, C, E\}, \{C, D, E\}, \{A, B, C, D\}, \{A, B, C, E\}, \{A, B, D, E\}, \{A, B, C, D, E\}\}$.



(a) (b)

Fig. 3. An example illustrating how to construct neighboring graphs G from an LTE network.

We assume that each *UE* has a specific probability to be called/paged (i.e., for voice call as well as for IP-based web applications). Further, each *UE* follows a different mobility pattern, hence the number of sites (cells) visited by each *UE* is different. In the online-assignment of *TALs-to-UEs* step, the network is monitored in order to track the number of signaling messages (i.e., *TAU* and paging) sent and received by different *UEs*. We denote by $\alpha = \{\alpha_1, \alpha_2 \cdots\}$ and $\beta = \{\beta_1, \beta_2 \cdots\}$ the probability of paging and *TAU* of *UEs* in the network, respectively. In other words, in the offline-assignment step, we have the information about different existing *UEs* in the network. We denote by Υ the different *UEs*. For each $UE_u \in \Upsilon$, we have its probability α_u to send a *TAU* message and its probability β_u to be called (i.e., cause a paging). We denote by $\gamma = \{\gamma_1, \gamma_2, \cdots\}$ the overhead of mobility and paging ratio of different *UEs*. γ_u denotes the overhead of mobility and paging ratio of UE_u, i.e. the ratio between the paging and the *TAU* of a UE_u. Formally, γ_u is computed as follows: $\gamma_u = \dfrac{\rho\alpha_u}{\rho\alpha_u + \tau\beta_u}$, where τ and ρ are the amount of overhead of one *TAU* operation and one paging message, respectively. Intuitively, the values of τ and

TABLE I
NOTATIONS USED IN THE PAPER.

Notation	Decription
Υ	The set of *UEs* in the network
$\mathcal{N}$	The set of *TAs* in the network
η_u	The number of cells (eNodeB) in *TA* u.
α_u	The probability that *UE* u gets paged during a period D.
β_u	The probability that *UE* u moves from *TA* to another i.e., mobility of *UE* u.
γ_u	The mobility and paging ratio of *UE* u.
Γ_i	The set of possible *TALs* that can be assigned to *UEs* in *TA* i.
$\digamma_i$	The sorted element of Γ_i.
$\mathcal{S}$	The matrix that ensures the mapping between *TAs* and *TALs* in the network.
$P_i(j)$	The probability of selecting a *TAL* j in *TA* i. Formally, $P_i(j) = \mathcal{S}_{ij}$.
Γ	The set of all possible *TALs* in the network.
h_{uv}	The number of handover between *TA* u and v.
τ	Overhead of one *TAU* operation.
ρ	Overhead of one paging message.
μ_i	The exponential distribution rate of the sojourn time of UEs in *TA* i
λ	The exponential distribution rate of the inter arrival time between two consecutive calls for a UE

ρ depend on the "radio system" [14]. Knowing that $\gamma_u \in [0, 1]$, the higher the value of γ_u is, the higher the number of paging of UE_u becomes in comparison to *TAU* messages. Accordingly, γ_u represents an important parameter to consider when designing *TALs* to assign to *UEs*. Indeed, when a *UE* has a high value of γ_u, meaning that it generates more paging messages than *TAU* messages, it is better to assign a *TAL* with a few number of *TAs* to reduce the paging overhead. However, if a *UE* has a low value of γ_u, meaning that it generates more *TAU* messages than paging, it is more appropriate to assign to it *TALs* with more *TAs* to reduce the *TAU* overhead.

Moreover, in the online-assignment of *TALs-to-UEs* step, we can deduce the number of *UEs* $h_{i,j}$ that moved from each *TA* i to another *TA* j. We define by $\mathcal{H}$ the matrix that represents the number of *UEs* that moved from different *TAs*. Each entry in the matrix $\mathcal{H}$ at row i and column j, denoted by $h_{i,j}$, indicates the number of *UEs* that moved from *TA* i to *TA* j. The value of $h_{i,j}$ can be deduced from the handover statistics of different eNodeBs or from the handover command messages sent by *MME*. Furthermore, each UE_i spends different times in different *TAs*. Let $\mathcal{M}$ denote the matrix that represents the duration spent by different *UEs* in different *TAs*. The rows in $\mathcal{M}$ represent the *UEs*, whereas the columns represent the different *TAs* in the network. The element $\mathcal{M}_{i,j}$ denotes the duration spent by UE_i in *TA* j. Note that, $\forall i \in \Upsilon$, $\sum_{j=1}^{N} \mathcal{M}_{i,j} = D$.

For the sake of readability, the notations used throughout the paper are summarized in Table I.

IV. ONLINE-ASSIGNMENT OF *TALs-to-UEs*

The mapping between *TAs* and *TALs* is represented through a matrix $\mathcal{S}$, where the rows are the different *TAs* and the columns are the different *TALs*. An element $\mathcal{S}_{i\ell}$, in the matrix $\mathcal{S}$, represents the probability to assign *TAL* ℓ in *TA* i to different *UEs*. Matrix $\mathcal{S}$ is first generated during the offline step and is used

i	1	2	3	4	5	6	7	8	9	10	11
$\digamma_A(\ell)$	A, B, C, D, E	A, B, C, D	A, B, C, E	A, B, D, E	A, D, E	A, B, E	A, B, D	A, B, C	A, D	A, B	A
$P_A(\ell)$	0.2	0.1	0.15	0.05	0.12	0.03	0.06	0.04	0.04	0.01	0.2
$\sum P_A(\ell)$	0.2	0.3	0.45	0.5	0.62	0.65	0.71	0.75	0.79	0.8	1

Fig. 4. *TALs* $\digamma_A$ and their probabilities P_A at *TA* A: an example.

then in the online step. Indeed, offline step generates Matrix $\mathcal{S}$ in a way that the *TAL* that optimizes more the network performance has a higher probability to be assigned to different *UEs*. From above, Γ_i, for $\forall i \in \mathcal{N}$, can be also defined as follows:

$$\Gamma_i = \{\ell, S_{i,\ell} \neq 0 \; for \; \forall \ell \in \Gamma \wedge i \in \ell\}$$

accordingly, when a *UE* visits a *TA* i, *MME* will assign to this *UE* a *TAL* from Γ_i. We denote by $\digamma_i$ the sorted element of Γ_i. *TALs* in $\digamma_i$ are sorted according to the number of *TAs* in each *TAL*, such that *TALs* having the smallest number of *TAs* are placed in the tail. $\digamma_i(\ell)$ represents the ℓ^{th} *TAL* of $\digamma_i$. We denote by $P_i(\ell)$ the probability to assign *TAL* $\digamma_i(\ell)$ by *TA* i to different *UEs*. $P_i(\ell)$ can be deduced from the matrix $\mathcal{S}$. Fig. 4 shows an example of $\digamma_A$ and P_A. In this example, $\digamma_A(1) = \{A, B, C, D, E\}$ and $\digamma_A(2) = \{A, B, C, D\}$.

The assignment of *TALs* to *UEs* should be lightweight in terms of computational cost and communication overhead. In this vein, the proposed solutions for this part are designed to be simple and easy to deploy. When a *UE* u visits a new *TA* A, the *MME* selects a new *TAL* $\digamma_A(\ell)$ from $\digamma_A$ according to the set of probability P_A. The *TAL* that has the highest probability would have more chance to be elected than the others. Then, the *MME* adds the last visited *TA* to $\digamma_A(\ell)$, to prevent the risk of ping-pong updates, before assigning it to *UE* u. It is worth noting that $\digamma_A(\ell)$ should be also assigned to each *UE* according to its mobility and paging features. Indeed, some *UEs* exhibit high mobility, while others are called more often. For this reason, unlike all existing works, in this paper we consider both the probability of each *TAL* $P_A(\ell)$ and the features of *UEs* when assigning *TALs* to different *UEs*. In this paper, two strategies are considered as explained below.

A. Assigning TALs to UEs Without Prioritization

In this strategy, we use only the probability of each *TAL* $P_A(\ell)$; i.e. no prioritization among *UEs* is considered. All *UEs* have the same priority to obtain any *TAL* from the visited *TAs*. This strategy could be used to reduce the involvement of *UEs* (and hence associated overhead and battery consumption) in the *TAL* assignment process. In this case, when a *UE* u visits a new *TA* A, the *MME* generates a random variable $\mathbf{V}_1 \in [0, 1]$ using a uniform distribution. Then, *TAL* ℓ is assigned to *UE* u as the one that satisfies the following condition:

$$\sum_{k=1}^{\ell-1} P_A(k) < \mathbf{V}_1 \leq \sum_{k=1}^{\ell} P_A(k)$$

Using the example depicted in Fig. 4, if $\mathbf{V}_1 = 0.38$, then *TAL* 3 would be assigned to *UE* u. By using this strategy, we ensure that *TALs* having higher probabilities will be more likely assigned to *UEs*. From above, we observe that the assignment of *TALs* to *UEs* without prioritization is light weighted. In fact, it is in the order of the generation of a random value $\mathbf{V}_1$ that follows a uniform distribution.

B. Assigning TALs to UEs With Prioritization

In this strategy, *UEs* exhibiting higher mobility rate than paging rate, should get *TALs* that have large number of *TAs* to mitigate the effect of *TAU* signaling. Employing the example depicted in Fig. 3, *TAL* $\{A, B, C, D, E\}$ is assigned to *UEs* that exhibit higher mobility features than paging, and that is to reduce the overhead of *TAUs*. Whereas, *TAL* $\{A\}$ is assigned to *UEs* having more paging than being highly mobile, and that is to reduce the impact of paging on the network performance. As discussed earlier, when a *UE* u visits a new *TA* TA_u, the *MME* in charge of TA_u, has the following information: (i) the matrix $\mathcal{S}$ and (ii) the overhead of mobility and paging ratio γ_u. We recall that the higher the value of γ_u is, the higher the number of paging is, i.e., in comparison to *TAU* (mobility).

To prioritize among *UEs* without impacting the probabilities of *TALs*, we define $F(\nu = x, k)$ as the cumulative distribution function of Poisson distribution until k, where ν is the mean value. Fig. 5 depicts $F(\nu = x, k)$ according to ν and k. When *UE* u visits *TA* A, *MME* computes for this *UE* its ν_u as $\nu_u = \lfloor \frac{1}{\gamma_u} \rfloor$. Since $\gamma_u \in [0, 1]$, then $\nu_u \geq 1$. Afterwards, a random variable $\mathbf{V}_2 \in [0, 100]$ is generated using a uniform distribution. Now, *TAL* ℓ is assigned to *UE* u as the one that satisfies the following condition:

$$\sum_{k=1}^{\ell-1} P_A(k) < F(\nu = \nu_u, \mathbf{V}_2) \leq \sum_{k=1}^{\ell} P_A(k)$$

From above, high values of γ_u mean that UE_u receives more paging messages than it issues *TAU* messages (due to mobility). For this *UE*, it is preferable to assign a *TAL* with small number of *TAs*. Note that large values of γ_u means small values of ν_u. From Fig. 5, *UE* u will have high probability to get a value in the vicinity of 1 and will be hence assigned *TALs* from the tail of $\mathcal{F}_A$ (i.e., *TAL* ℓ with small size). Whereas, when γ_u is small (i.e., *UE* u has high mobility features than paging), its ν_u will be large. Then, *UE* u has high probability to be assigned a *TAL* ℓ from the head of $\mathcal{F}_A$ (i.e., *TAL* ℓ with large size). The assignment of *TALs* to *UEs* with prioritization is also in the order of the generation of a random value $\mathbf{V}_2$ that follows a uniform distribution.

Theorem 1: *TAL* ℓ having the highest value of $P_A(\ell)$, has higher probability to be selected for different *UEs*.

Proof: Let TAL_ℓ denote the *TAL* that has the highest value of $P_A(\ell)$ at *TA* A. Formally, $P_A(\ell) = \sum_{k=1}^{\ell} P_A(k) - \sum_{k=1}^{\ell-1} P_A(k)$. We have two cases: (i) Assigning *TALs* from TAL_A to *UEs* without prioritization and (ii) Assigning *TALs* from TAL_A to *UEs* with prioritization. In the first case, a random probability $\mathbf{V}_1 \in [0, 1]$ is generated to select *TALs*. Whereas, in the second case, a random number $\mathbf{V}_2 \in [0, 100]$ is generated and then $F(\nu = \nu_u, \mathbf{V}_2)$ is computed. As *TAL* ℓ has the highest value of $P_A(\ell)$, for both cases it is more likely that $\mathbf{V}_1$ (resp., $F(\nu = \nu_u, \mathbf{V}_2)$) is in $[\sum_{k=1}^{\ell-1} P_A(k), \sum_{k=1}^{\ell} P_A(k)]$. Therefore, in both cases *TAL* ℓ that has the highest value of $P_A(\ell)$ is more likely to be selected by *UEs*. ∎

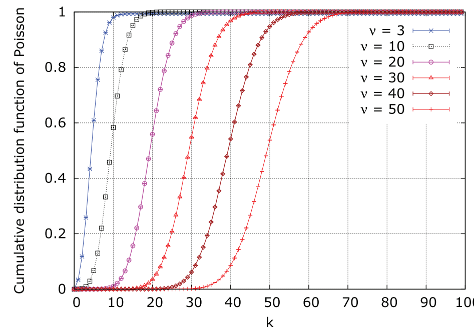


Fig. 5. The impact of ν values on the cumulative distribution function of Poisson.

Theorem 2: When assigning *TALs* to *UEs* via prioritization strategy, a *UE* u having higher speed (i.e., highly mobile) than paging ratio γ_u, is more likely to be assigned a *TAL* with large size to mitigate the effect of *TAU*.

Proof: Based on the above, the *UE* which has higher speed than paging ratio, has the smallest value of γ_u, and then, the highest value of ν_u. From Fig. 5, it is more likely to get $F(\nu = \nu_u, \mathbf{V}_2)$ in the vicinity of zero, and consequently select a *TAL* from the head of $\mathcal{F}_A$ that has a large size. ∎

V. Offline-Assignment of *TAs-to-TALs*

As discussed in Section III, this step is executed offline to allow the mapping between different *TAs* and *TALs*. At the end of this step, a matrix $\mathcal{S}$ is generated, whereby the rows represent the different *TAs* $\mathcal{N}$ and the columns represent the *TALs* Γ. An element $\mathcal{S}_{ij}$ in the matrix $\mathcal{S}$ refers to the probability that *TA* i assigns *TAL* j to different *UEs*. The sites (cells) belonging to the same *TA* i use the same row i in the matrix $\mathcal{S}$ to assign *TALs* to different *UEs*. As mentioned in Section III, the result of this step is used by the online step of our framework to assign different *TALs* to different *UEs*. In what follows, we present three problem formulations for optimizing *TALs* distribution in LTE and beyond networks. The two first optimizations are linear programs, whereas the last one is a convex optimization. As it is well known in the literature [15], the linear program and convex optimization have polynomial time complexity. It shall be noted that the result of the three solutions is the same matrix $\mathcal{S}$, however, with different elements $\mathcal{S}_{ij}$. The latter are considered as the variables for the problem optimizations. In the first optimization problem, we assume that the *TAU* overhead is dominator and we then propose a solution to optimize the network performance that favors *TAU* on paging. In the second solution, we propose an optimization problem whereby the paging overhead is dominator. Finally, we introduce *FOTA*, which aims at capturing the tradeoff between the *TAU* and paging overhead when assigning *TALs* to *TAs* (Fair and Optimal Assignment of *TALs* to *TAs* - *FOTA*), and ultimately to *UEs*.

In *FOTA*, a bargaining game is used to capture the tradeoff between *TAU* and paging.

A. Optimizing the Network Performance via the Reduction of TAU Overhead

In this subsection, we propose the solution, named *F-TAU*, that favors *TAU* when assigning *TAs* to *TALs*. In *F-TAU*, we seek the optimal distribution of *TALs* by applying the min-max approach. The aim is to minimize the maximum number of *TAU* messages. Formally, we aim to minimize the maximum aggregate number of *TAU* messages sent by *UEs* between any two *TAs* in the network. In this solution, we denote by $PAGING_{max}$ the maximum number of paging messages tolerated by the network. Its value could be fixed according to the capacity of *MMEs* in the network. Otherwise, $PAGING_{max}$ can be fixed to ∞. In this case, the optimal solution would converge to putting all *TAs* into the same *TAL* in order to reduce the *TAU* overhead. At this point, the optimization model which aims at reducing the *TAU* overhead can be formulated according to the following linear program ((1)...(6)):

$$\mathbf{min} \max_{\forall i,j\in\mathcal{N}\wedge i\neq j} \tau\left(\sum_{\ell\in\Gamma_i\wedge\ell\notin\Gamma_j} h_{ij}S_{i\ell} + \sum_{\ell\in\Gamma_j\wedge\ell\notin\Gamma_i} h_{ji}S_{j\ell}\right) \quad (1)$$

$S.t,$

$$\forall\ell\in\Gamma,\forall i\in\mathcal{N}\cap\ell, \mathcal{S}_{i\ell}\geq 0 \quad (2)$$

$$\forall\ell\in\Gamma,\forall i\in\mathcal{N}\cap\ell, \mathcal{S}_{i\ell}\leq 1 \quad (3)$$

$$\forall i\in\mathcal{N}, \sum_{\ell\in\Gamma}\mathcal{S}_{i\ell}=1 \quad (4)$$

$$\forall\ell\in\Gamma,\forall i\notin\mathcal{N}\cap\ell, \mathcal{S}_{i\ell}=0 \quad (5)$$

$$\rho\sum_{\ell\in\Gamma}\sum_{i\in\ell}\mathcal{S}_{i\ell}\left(\sum_{k\in\Upsilon}\alpha_k\mathcal{M}_{ki}\right)\left(\sum_{j\in\ell\wedge j\neq i}\eta_j\right)\leq PAGING_{max} \quad (6)$$

In the objective function (1), the number of *UEs* that transited from *TA* i (resp., j) is scaled by the variable $\mathcal{S}_{i\ell}$ (resp., $\mathcal{S}_{j\ell}$), which represents the proportional use of *TAL* ℓ by *TA* i (resp, j). It shall be also noted that the condition,"$\ell\in\Gamma_i\wedge\ell\notin\Gamma_j \Leftrightarrow \forall i,j\in\mathcal{N}, i\neq j, \forall\ell\in\Gamma : i\in\ell\wedge j\notin\ell$", aims at reducing the number of *UEs* moving between different *TAs* that do not belong to the same *TALs*. The first three constraints ((2)–(4)) are used to ensure that each *TA* $i\in\mathcal{N}$ can select its *TAL* from $\mathcal{S}_i$ with a fixed probability. The fourth constraint (5) ensures that a *TA* delivers *TALs* to *UEs* only if it belongs to this *TALs*. The last constraint (6) ensures that the sum of all paging overhead in the network should not exceed a predefined threshold $PAGING_{max}$. For any *TAL* ℓ, the overhead caused by paging *UEs* residing in *TA* $i\in\ell$ (by sending paging messages to all *TAs* $j\in\ell\wedge j\neq i$) is the number of sites η_j in these *TAs*, scaled by $\sum_{k\in\Upsilon}\alpha_k\mathcal{M}_{ki}$ and a variable $\mathcal{S}_{i\ell}$. Note that $\sum_{k\in\Upsilon}\alpha_k\mathcal{M}_{ki}$ is a constant that represents the paging overhead at *TA* i and $\mathcal{S}_{i\ell}$ represents the proportional use of i. Formally, $\sum_{k\in\Upsilon}\alpha_k\mathcal{M}_{ki}$ is defined as the sum of the probabilities of paging of each *UE* k scaled by its residence time in *TA* i.

B. Optimizing the Network Performance via the Reduction of Paging Overhead

In this subsection, we introduce *F-PAGING*, which favors the paging overhead when assigning *TAs* to *TALs*. As in *F-TAU*, we use the min-max approach as depicted in the linear program ((7)...(8)). In this linear program, the goal (7) is to optimize the network performance seeking the optimal distribution of *TALs* that minimizes the paging overhead. In this solution, we set the maximum amount of *TAU* overhead tolerated by the network to TAU_{max}. Its value could be defined according to the capacity of *MMEs* in the network. Otherwise, TAU_{max} can be fixed to ∞. In this case, the optimal solution would converge to putting each *TA* in a separate *TAL* in order to reduce the paging overhead. The linear program is formulated as follows:

$$\mathbf{min}\ \rho\sum_{\ell\in\Gamma}\sum_{i\in\ell}\left(\mathcal{S}_{i\ell}\left(\sum_{k\in\Upsilon}\alpha_k\mathcal{M}_{ki}\right)\sum_{j\in\ell\wedge j\neq i}\eta_j\right) \quad (7)$$

$S.t,$

(2)−(5) *and*

$\forall i,j\in\mathcal{N}\wedge i\neq j:$

$$\tau\left(\sum_{\ell\in\Gamma_i\wedge\ell\notin\Gamma_j} h_{ij}S_{i\ell} + \sum_{\ell\in\Gamma_j\wedge\ell\notin\Gamma_i} h_{ji}S_{j\ell}\right)\leq TAU_{max} \quad (8)$$

The first fourth constraints ((2)...(5)) are similar to the first linear program presented in the precedent section. The last constraint ensures that the total number of *TAU* messages sent by *UEs* when transiting between any two adjacent *TAs* $i\in\mathcal{N}$ and $j\in\mathcal{N}$ should not exceed the threshold TAU_{max}.

C. Trading off TAU Against Paging Using Nash Bargaining

In contrast to the conventional techniques (eg., weighted-sum method) used to solve the multi-objectives problems, which may not ensure a fair tradeoff between the conflicting objectives, *FOTA* uses a Nash bargaining game to achieve this tradeoff. As we have mentioned in Fig. 1, an increase in the size of TALs reduces the TAU signaling messages, however it has a negative impact on the paging signaling messages. Meanwhile, reducing TALs size has a negative impact on TAU signaling messages and positive impact on the paging signaling messages. The UE's mobility and call ratio have a great impact on the total number (i.e., TAU and paging) of signaling messages in the network. For a network characterized by a high mobility, we have to favor the reduction of TAU overheads in order to reduce the number of total signaling messages in the network. Whereas, for a network characterized by a high call ratio, the reduction of paging signaling messages significantly reduces the total signaling messages. In *FOTA*, *TAU* and paging overhead represent the conflicting objectives and are considered as two players in the bargaining game. The two players (i.e., *TAU* and paging signaling messages) would like to barter goods (i.e., total signaling messages). It was theoretically proven in [16] that the use of Nash bargaining game ensures a fair tradeoff between the players according to the network characteristics in terms of UE's mobility and call ratio. *FOTA* will favor the reduction of TAU overhead for a network characterized by a high mobility, whereas it will favor the reduction of paging overhead for a network characterized by a high call ratio. In what follows, some background on the Nash bargaining game is introduced and then *FOTA* solution is presented.

1) Nash Bargaining Model and Threat Value Game: Nash bargaining model can be viewed as a game between two players who would like to barter goods. This model is a cooperative game with non-transferable utility. This means that the utility scales of the players are measured in non-comparable units. This model is adopted in our proposed *FOTA* scheme to find a Pareto efficiency between the paging and *TAU* overhead. In our case, the players are the paging and *TAU* overhead which do not use the same unit. This model is based on two elements, assumed to be given and known to the players. First, the set of vector payoffs $\mathcal{P}$ achieved by the players if they agree to cooperate. $\mathcal{P}$ should be a convex and compact set. Formally, $\mathcal{P}$ can be defined as $\mathcal{P} = \{(u(x), v(x)), x = (x_1, x_2) \in X\}$, whereby X is the set of strategies of two players, and $u()$ and $v()$ are the utility functions of the first and second users, respectively. Second, the threat point, $d = (u^*, v^*) = (u((t_1, t_2)), v(t_1, t_2)) \in \mathcal{P}$, which represents the pair of utility whereby the two players fail to achieve an agreement. In Nash bargaining game, we aim to find a fair and reasonable point, $(\overline{u}, \overline{v}) = f(\mathcal{P}, u^*, v^*) \in \mathcal{P}$ for an arbitrary compact convex set $\mathcal{P}$ and point $(u^*, v^*) \in \mathcal{P}$. Based on Nash theory, a set of axioms are defined that lead to $f(\mathcal{P}, u^*, v^*)$ in order to achieve a unique optimal solution $(\overline{u}, \overline{v})$:

1) **Feasibility:** $(\overline{u}, \overline{v}) \in \mathcal{P}$.
2) **Pareto Optimality:** There is no point $(u(x), v(x)) \in \mathcal{P}$ such that $u(x) \geq \overline{u}$ and $v(x) \geq \overline{v}$ except $(\overline{u}, \overline{v})$. In other words, if $\mathcal{P}$ is symmetric about the line $u(x) = v(x)$, and $u^* = v^*$, then $\overline{u} = \overline{v}$.
3) **Independence of irrelevant alternatives:** If T is a closed convex subset of $\mathcal{P}$, and if $(u^*, v^*) \in T$ and $(\overline{u}, \overline{v}) \in T$, then $f(\mathcal{P}, u^*, v^*) = (\overline{u}, \overline{v})$.
4) **Invariance under change of location and scale:** If $T = \{(u'(x), v'(x)), u'(x) = \alpha_1 u(x) + \beta_1, v'(x) = \alpha_2 v(x) + \beta_2 for (u(x), v(x)) \in \mathcal{P}\}$, where $\alpha_1 > 0$, $\alpha_2 > 0$, and B_1 and B_2 are given numbers, then $f(T, \alpha_1 u^* + \beta_1, \alpha_2 v^* + \beta_2) = (\alpha_1 \overline{u} + \beta_1, \alpha_2 \overline{v} + \beta_2)$.

Moreover, the unique solution $(\overline{u}, \overline{v})$, satisfying the above axioms, is proven to be the solution of the following optimization problem:

$$\begin{cases} \mathbf{max}(u(x) - u^*)(v(x) - v^*) \\ \mathbf{s.t.} \\ \quad (u(x), v(x)) \in S \\ \quad (u(x), v(x)) \geq (u^*, v^*) \end{cases}$$

A general geometric interpretation of the Nash bargaining game is shown in Fig. 6.

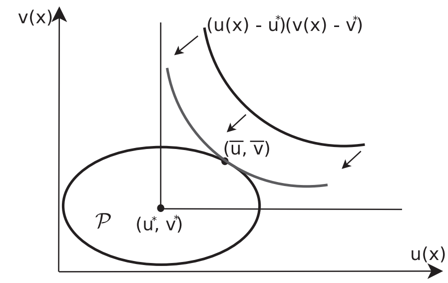


Fig. 6. The geometric interpretation of the Nash bargaining game.

2) Fair and Optimal TALs Assignment: We denote by $d = (TAU_{worst}, PAGING_{worst})$ the threat point of our bargaining game that solves *FOTA*. In contrast to conventional bargaining game, the utility function of each player, (i.e., *TAU* and paging overhead) in our model, is the opposite of its cost. In other words, $(TAU_{worst}, PAGING_{worst}) \geq (f(\mathcal{S}), g(\mathcal{S})), \forall \mathcal{S} \in X$, where $f()$ and $g()$ are the utility functions of *TAU* and paging overhead players, respectively. The tradeoff problem between *TAU* and paging overhead can be modeled as a convex optimization problem ((9)...(13)).

$$\mathbf{max}\ (TAU_{worst} - f(\mathcal{S}))(PAGING_{worst} - g(\mathcal{S})) \tag{9}$$

$S.t,$

$(2)-(5)$ *and*

$\forall i, j \in \mathcal{N} \wedge i \neq j :$

$$\tau \left(\sum_{\ell \in \Gamma_i \wedge \ell \notin \Gamma_j} h_{ij} S_{i\ell} + \sum_{\ell \in \Gamma_j \wedge \ell \notin \Gamma_i} h_{ji} S_{j\ell} \right) \leq f(\mathcal{S}) \tag{10}$$

$$\rho \sum_{\ell \in \Gamma} \sum_{i \in \ell} \mathcal{S}_{i\ell} \left(\sum_{k \in \Upsilon} \alpha_k \mathcal{M}_{ki} \right) \left(\sum_{j \in \ell \wedge j \neq i} \eta_j \right) \leq g(\mathcal{S}) \tag{11}$$

$$f(\mathcal{S}) \leq TAU_{worst} \tag{12}$$

$$g(\mathcal{S}) \leq PAGING_{worst} \tag{13}$$

In the optimization problem, in addition to matrix $\mathcal{S}$, we added two variables $f(\mathcal{S})$ and $g(\mathcal{S})$ that represent the maximum values of *TAU* and paging overheads in the network, respectively. The use of Nash bargaining game in *FOTA* ensures fairness among the players (*TAU* and paging overheads) and produces a Pareto optimal solution. From the second and the third axioms of the bargaining game, we can deduce that *FOTA* yields a fair Pareto optimal solution according to the threat point $(TAU_{worst}, PAGING_{worst})$, which represents the performance thresholds of *TAU* and *paging* overheads, respectively. Let $\mathcal{S}^{TAU}$ and $\mathcal{S}^{PAGING}$ be the optimal solutions of the linear programs ((1)...(6)) and ((7)...(8)), respectively. Then, we can define $PAGING_{worst}$, $PAGING_{best}$, TAU_{worst} and TAU_{best} as follows:

1) $PAGING_{worst} = \rho \sum_{\ell \in \Gamma} \sum_{i \in \ell} \left(\left(\sum_{k \in \Upsilon} \alpha_k \mathcal{M}_{ki} \right) \sum_{j \in \ell \wedge j \neq i} \eta_j \mathcal{S}^{TAU}_{i\ell} \right)$
2) $PAGING_{best} = \rho \sum_{\ell \in \Gamma} \sum_{i \in \ell} \left(\left(\sum_{k \in \Upsilon} \alpha_k \mathcal{M}_{ki} \right) \sum_{j \in \ell \wedge j \neq i} \eta_j \mathcal{S}^{PAGING}_{i\ell} \right)$
3) $TAU_{worst} = \max_{\forall i, j \in \mathcal{N}, i \neq j} \left(\tau \left(\sum_{\ell \in \Gamma_i \wedge \ell \notin \Gamma_j} h_{ij} S_{i\ell} + \sum_{\ell \in \Gamma_j \wedge \ell \notin \Gamma_i} h_{ji} S^{PAGING}_{j\ell} \right) \right)$

4) $TAU_{best} = \max_{\forall i,j \in \mathcal{N}, i \neq j} \left(\tau \left(\sum_{\ell \in \Gamma_i \wedge \ell \notin \Gamma_j} h_{ij} S_{i\ell} + \sum_{\ell \in \Gamma_j \wedge \ell \notin \Gamma_i} h_{ji} S^{TAU}_{j\ell} \right) \right)$

It is easily noticeable that $PAGING_{best} \leq PAGING_{worst}$ and $TAU_{best} \leq TAU_{worst}$. Fig. 7 illustrates the physical interpretation of the trade-off between TAU and paging overheads. From this figure, we can observe that a reduction in TAU signaling messages increases the number of paging signaling messages, and vise versa. *FOTA* aims at finding the Pareto optimal point $(f(\overline{\mathcal{S}}), g(\overline{\mathcal{S}}))$ between TAU and paging overhead. The slope of $\mathcal{P}$ would vary according to the network characteristics, in terms of UE's mobility and paging ratio, which have an impact on the Pareto optimal point $(f(\overline{\mathcal{S}}), g(\overline{\mathcal{S}}))$.

The values of $PAGING_{best}$, $PAGING_{worst}$, TAU_{best} and TAU_{worst} are obtained by updating the linear programs ((1)...(6)) and ((8)...(8)) as follows:

$$\textbf{min } f(\mathcal{S}) \tag{14}$$

$S.t$,

(2)–(5) *and*

$\forall i, j \in \mathcal{N} \wedge i \neq j$:

$$\tau \left(\sum_{\ell \in \Gamma_i \wedge \ell \notin \Gamma_j} h_{ij} S_{i\ell} + \sum_{\ell \in \Gamma_j \wedge \ell \notin \Gamma_i} h_{ji} S_{j\ell} \right) \leq TAU_{best} \tag{15}$$

$$\rho \sum_{\ell \in \Gamma} \sum_{i \in \ell} \mathcal{S}_{i\ell} \left(\sum_{k \in \Upsilon} \alpha_k \mathcal{M}_{ki} \right) \left(\sum_{j \in \ell \wedge j \neq i} \eta_j \right) \leq PAGING_{worst} \tag{16}$$

$$PAGING_{worst} \leq PAGING_{max} \tag{17}$$

$$TAU_{best} \leq f(\mathcal{S}) \tag{18}$$

$$\textbf{min } g(\mathcal{S}) \tag{19}$$

$S.t$,

(2)–(5) *and*

$\forall i, j \in \mathcal{N} \wedge i \neq j$:

$$\tau \left(\sum_{\ell \in \Gamma_i \wedge \ell \notin \Gamma_j} h_{ij} S_{i\ell} + \sum_{\ell \in \Gamma_j \wedge \ell \notin \Gamma_i} h_{ji} S_{j\ell} \right) \leq TAU_{worst} \tag{20}$$

$$\rho \sum_{\ell \in \Gamma} \sum_{i \in \ell} \mathcal{S}_{i\ell} \left(\sum_{k \in \Upsilon} \alpha_k \mathcal{M}_{ki} \right) \left(\sum_{j \in \ell \wedge j \neq i} \eta_j \right) \leq PAGING_{best} \tag{21}$$

$$PAGING_{best} \leq g(\mathcal{S}) \tag{22}$$

$$TAU_{worst} \leq TAU_{max} \tag{23}$$

The optimization problem shown in the linear program ((9)...(13)) is non-convex. Using the approach proposed in [17], the problem can be transformed to convex-optimization problem without changing the solution. The key idea is to introduce the log function which is an increasing function. Therefore, the optimization problem is reformulated as follows:

$$\textbf{max} \log((TAU_{worst} - f(\mathcal{S}))) + \log((PAGING_{worst} - g(\mathcal{S}))) \tag{24}$$

S.t,

(2)–(5) *and*

$\forall i, j \in \mathcal{N} \wedge i \neq j$:

$$\tau \left(\sum_{\ell \in \Gamma_i \wedge \ell \notin \Gamma_j} h_{ij} S_{i\ell} + \sum_{\ell \in \Gamma_j \wedge \ell \notin \Gamma_i} h_{ji} S_{j\ell} \right) \leq f(\mathcal{S}) \tag{25}$$

$$\rho \sum_{\ell \in \Gamma} \sum_{i \in \ell} \mathcal{S}_{i\ell} \left(\sum_{k \in \Upsilon} \alpha_k \mathcal{M}_{ki} \right) \left(\sum_{j \in \ell \wedge j \neq i} \eta_j \right) \leq g(\mathcal{S}) \tag{26}$$

$$f(\mathcal{S})) \leq TAU_{worst} \tag{27}$$

$$g(\mathcal{S})) \leq PAGING_{worst} \tag{28}$$

Theorem 3: The optimization problem ((24)...(28)) is convex and admits a unique solution.

Proof: To prove the unicity of the solution, we have to show that the optimization problem in ((24)...(28)) is convex. It shall be stated that for an optimization problem to be convex, the objective function should be convex, the equality constraints should be linear, and the inequality constraints should be convex [15]. For our optimization problem ((24)...(28)), the equality and the inequality constraints are linear. This also means that the inequality constraints are convex. Thus, to show that the optimization problem in ((24)...(28)) is convex, it is sufficient to prove that the objective function is convex. In the optimization problem ((24)...(28)), we have TAU_{worst} and $PAGING_{worst}$ as constant values, whereas $f(\mathcal{S})$ and $g(\mathcal{S})$ are variables. For the sake of simplicity, we denote TAU_{worst}, $PAGING_{worst}$, $f(\mathcal{S})$ and $g(\mathcal{S})$ by A, B, x and y, respectively. Thus, the objective function becomes $\textbf{max} \log(A - x) + \log(B - y)$. Based on [15], the convex optimization problem should be minimized. For this reason, the objective function is transformed, without changing the solution as follows: $\textbf{min } P = -(\log(A - x) + \log(B - y))$. To prove that the optimization problem ((24)...(28)) is convex, it is sufficient to show that the Hessian matrix $\mathbf{H}$ of P is positive definite.

$$\begin{pmatrix} \frac{\partial^2 P}{\partial^2 x} & \frac{\partial^2 P}{\partial x \partial y} \\ \frac{\partial^2 P}{\partial y \partial x} & \frac{\partial^2 P}{\partial^2 y} \end{pmatrix}$$

Computing the different components of the Hessian matrix, we obtain

$$\frac{\partial^2 P}{\partial x \partial y} = \frac{\partial^2 P}{\partial y \partial x} = 0$$

$$\frac{\partial^2 P}{\partial^2 x} = \frac{1}{(A - x)^2} > 0$$

$$\frac{\partial^2 P}{\partial^2 y} = \frac{1}{(B - y)^2} > 0$$

It follows that the Hessian matrix is diagonal with positive eigenvalues. Therefore, the Hessian matrix is positive definite, the optimization problem is thus convex and admits a unique solution. ■

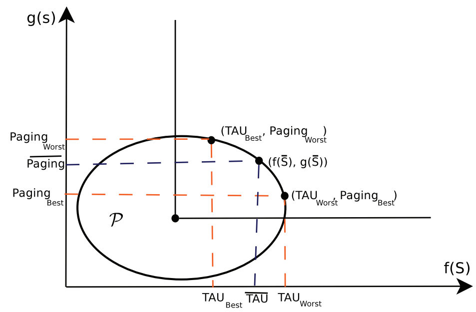


Fig. 7. The geometric interpretation of the tradeoff between *TAU* and paging overhead using Nash bargaining game.

VI. Analytical Model

In this section, we introduce a Markov-based model for analyzing the three offline solutions, *F-TAU*, *F-PAGING* and *FOTA*. We use the same intuition to model the three solutions, since the main difference between these solutions is the output matrix $\mathcal{S}$. To ease the explanation of the proposed analytical model, let us consider the network topology depicted in Fig. 8. The possible *TALs* for Fig. 8 is $\Gamma = \{\{\eta_1\}, \{\eta_2\}, \{\eta_3\}, \{\eta_1, \eta_2\}, \{\eta_1, \eta_3\}, \{\eta_2, \eta_3\}, \{\eta_1, \eta_2, \eta_3\}\}$. We numerate the elements in Γ from 1 to 7, respectively. Now, we consider the following matrix $\mathcal{S}$, which can be produced via *F-TAU*, *F-PAGING* or *FOTA*:

$$\mathcal{S} = \begin{bmatrix} 0.3 & 0 & 0 & 0.2 & 0.5 & 0 & 0 \\ 0 & 0.3 & 0 & 0.3 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0.1 & 0.4 & 0.5 \end{bmatrix}$$

We denote by H the expected probability of movement of a *UE* in the network. H can be deduced from $\mathcal{H}$. Each element $\hbar_{i,j}$ in H can be computed as follows:

$$\forall i \in \mathcal{N}, \hbar_{i,j} = \frac{h_{i,j}}{\sum\limits_{\forall j \in \mathcal{N}} h_{i,j}}$$

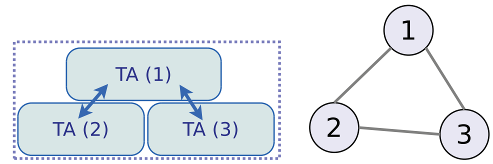
Considering the example of Fig. 8, H is:

$$H = \begin{bmatrix} 0 & 0.1 & 0.9 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{bmatrix}$$

Let M denote the expected duration of a *UE* in each *TA*. Formally, M is a vector with a size L. Each element M_i in M represents the time that the *UE* can spend in *TA* i. M_i can be computed as follow:

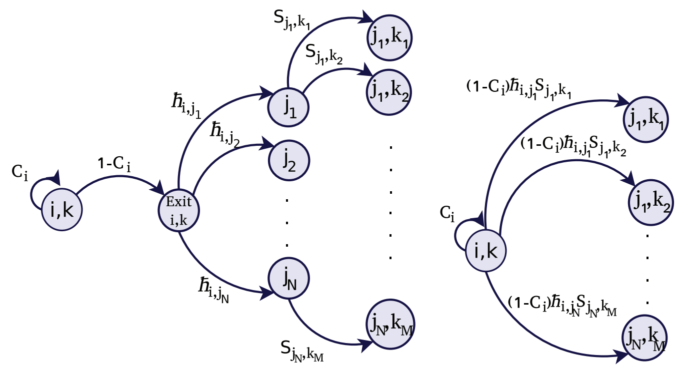
$$\forall i \in \mathcal{N}, M_i = \frac{\sum\limits_{\forall j \in \Upsilon} \mathcal{M}_{i,j}}{|\Upsilon|}$$

In our analysis, we assume that M_i, for $\forall i \in \Upsilon$, are independent and each M_i follows an exponential distribution of rate μ_i. $\frac{1}{|\mathcal{N}|} \sum\limits_{i \in \mathcal{N}} \alpha_i$ denotes the average arrival traffic of *UEs* in the network. Assuming that this traffic follows a Poisson process of rate λ, the inter arrival time between two consecutive calls is a random variable $\mathcal{T}$ that follows an exponential distribution of rate λ.



(a) (b)

Fig. 8. An illustrative example network used in the analysis.



(a) Embedded Markov chain (b) Embedded Markov chain with aggregated states

Fig. 9. The way to construct the embedded Markov chain used in the analysis.

These assumptions lead us to model the system using a Markov Chain $X = \{X_t, t \geq 0\}$ on the state space Θ defined by $\Theta = \{(i, k), \forall k \in \Gamma \wedge \forall i \in k \wedge \mathcal{S}_{ik} \neq 0\}$. In this model, $X_t = (i, k)$ indicates that at instant t, *TAL* k is assigned to *UEs* when visiting *TA* i. According to this description, it is obvious that we are dealing with a Continuous-Time Markov Chain (CTMC). In what follows, rather than the CTMC, we will use the corresponding Embedded Markov Chain (EMC), which is depicted in Fig. 9(a). From this figure, we notice two events that lead to leave a state (i, k) in EMC. The first one is when an incoming call arrives for a *UE* before it leaves its current *TA* i, whereas the second event is when the *UE* moves from its *TA* to another one before the incoming call arrives. As $M_i \sim Exp(\mu_i)$ and $\mathcal{T} \sim Exp(\lambda)$, the probability for the first and the second events to be occurred can be defined as follows:

- For an incoming call to arrive before the *UE* leaves its state i, the probability is $C_i = P(\mathcal{T} < M_i) = \dfrac{\lambda}{\lambda + \mu_i}$.
- For the *UE* to leave its *TA* i before the incoming call arrives, the probability is $1 - C_i = P(M_i \leq \mathcal{T}) = \dfrac{\mu_i}{\lambda + \mu_i}$.

Let $j_1, \cdots, j_N$ be the neighboring *TAs* of *TA* i. As depicted in Fig 9(a), when a *UE* exists its *TA* i, it has to move to its neighboring *TA* j according to the matrix H. Furthermore, when it moves to *TA* j, it has to select its *TAL* k according to the matrix $\mathcal{S}$. The EMC depicted in Fig. 9(a) can be reduced by grouping its states to a new EMC as shown in Fig. 9(b). Indeed, when a *UE*, assigned *TAL* l, moves from *TA* i to another *TA* j, two types of events can happen: (i) the first one corresponds to the case where *TA* j belongs to *TAL* k; (ii) the second one is when *TA* j does not belong to *TAL* k, in this case a *TA* update process should be accomplished to assign a new *TAL* k to the

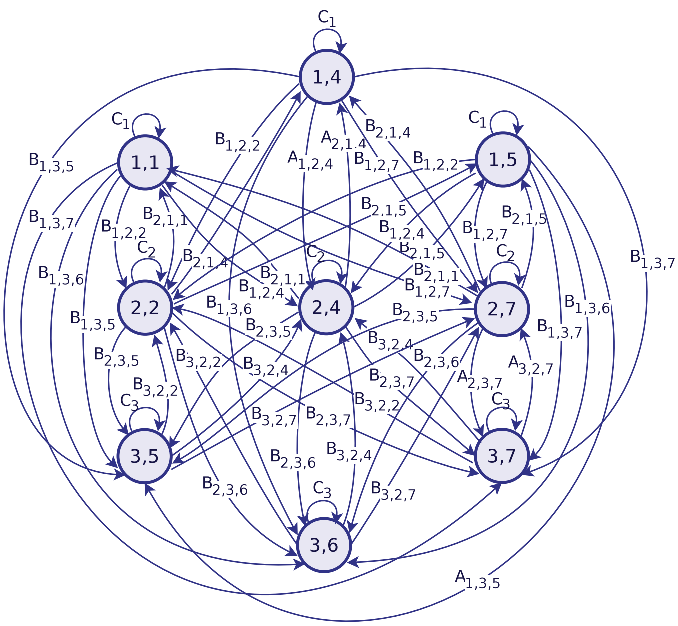


Fig. 10. An illustrative example of Embedded Markov Chain.

UE. Let denote by $A_{i,j,k}$ and $B_{i,j,k}$ the probability of the first and the second events, respectively. In fact, $A_{i,j,k}$ and $B_{i,j,k}$ represent the probabilities of moving from *TA* i to another *TA* j and then selecting *TAL* k.

$$\begin{cases} A_{i,j,k} = Pr(\mathcal{T} > M_i)\hbar_{i,j}\mathcal{S}_{jk}. \\ \quad \forall i, j \in \mathcal{N}, \forall k \in \Gamma, i \neq j \text{ and } i, j \in k \\ B_{i,j,k} = Pr(\mathcal{T} > M_i)\hbar_{i,j}\mathcal{S}_{jk}. \\ \quad \forall i, j \in \mathcal{N}, \forall k \in \Gamma, i \neq j, j \in k \text{ and } i \notin k \\ C_i = Pr(\mathcal{T} < M_i). \forall i \in \mathcal{N} \end{cases}$$

Hence,

$$\begin{cases} A_{i,j,k} = \dfrac{\mu_i}{\lambda + \mu_i}\hbar_{i,j}\mathcal{S}_{jk}. \\ \quad \forall i, j \in \mathcal{N}, \forall k \in \Gamma, i \neq j \text{ and } i, j \in k \\ B_{i,j,k} = \dfrac{\mu_i}{\lambda + \mu_i}\hbar_{i,j}\mathcal{S}_{jk}. \\ \quad \forall i, j \in \mathcal{N}, \forall k \in \Gamma, i \neq j, j \in k \text{ and } i \notin k \\ C_i = \dfrac{\lambda}{\lambda + \mu_i}. \forall i \in \mathcal{N} \end{cases}$$

Fig. 10 shows the corresponding Embedded Markov Chain of the network topology depicted in Fig. 8. The balance equations of EMC can be written according to the following formulas: $\forall (j,k) \in \Theta : \pi_{j,k} = C_j\pi_{j,k} + \sum_{i \in \mathcal{N} \wedge i \neq j \wedge \mathcal{S}_{ik}=0 \wedge \hbar_{i,j} \neq 0} (B_{i,j,k} \sum_{\ell \in \Gamma \wedge \mathcal{S}_{i\ell} \neq 0} \pi_{i,\ell}) + \sum_{i \in \mathcal{N} \wedge i \neq j \wedge \mathcal{S}_{ik} \neq 0 \wedge \hbar_{i,j} \neq 0} A_{i,j,k}\pi_{i,k}$ Where $\pi_{j,k}$ denotes the probability at steady state to assigning *TAL* k to *UEs* in *TA* i.

The following equations show the balance equations of the illustrative example shown in Fig 10:

$$\begin{cases} \pi_{1,1} = C_1\pi_{1,1} + B_{2,1,1}(\pi_{2,2} + \pi_{2,4} + \pi_{2,7}) \\ \pi_{1,4} = C_1\pi_{1,4} + B_{2,1,4}(\pi_{2,2} + \pi_{2,7}) + A_{2,1,4}\pi_{2,4} \\ \pi_{1,5} = C_1\pi_{1,5} + B_{2,1,1}(\pi_{2,2} + \pi_{2,4} + \pi_{2,7}) \\ \pi_{2,2} = C_2\pi_{2,2} + B_{1,2,2}(\pi_{1,1} + \pi_{1,4} + \pi_{1,5}) \\ \qquad + B_{3,2,2}(\pi_{3,5} + \pi_{3,6} + \pi_{3,7}) \\ \pi_{2,4} = C_2\pi_{2,4} + B_{1,2,4}(\pi_{1,1} + \pi_{1,5}) + A_{1,2,4}\pi_{1,4} \\ \qquad \times B_{3,2,4}(\pi_{3,5} + \pi_{3,6} + \pi_{3,7}) \\ \pi_{2,7} = C_2\pi_{2,7} + B_{1,2,7}(\pi_{1,1} + \pi_{1,4} + \pi_{1,5}) \\ \qquad \times B_{3,2,7}(\pi_{3,5} + \pi_{3,6}) + A_{3,2,7}\pi_{3,7} \\ \pi_{3,5} = C_3\pi_{3,5} + B_{1,3,5}(\pi_{1,1} + \pi_{1,4}) + A_{1,3,5}\pi_{1,5} \\ \qquad + B_{2,3,5}(\pi_{2,2} + \pi_{2,4} + \pi_{2,7}) \\ \pi_{3,6} = C_3\pi_{3,6} + B_{1,3,6}(\pi_{1,1} + \pi_{1,4} + \pi_{1,5}) \\ \qquad + B_{2,3,6}(\pi_{2,2} + \pi_{2,4} + \pi_{2,7}) \\ \pi_{3,7} = C_3\pi_{3,7} + B_{1,3,7}(\pi_{1,1} + \pi_{1,4} + \pi_{1,5}) \\ \qquad + B_{2,3,7}(\pi_{2,2} + \pi_{2,4}) + A_{2,3,7}\pi_{2,7} \end{cases}$$

Let N_{TAU} and N_{paging} denote the expected numbers of *TAU* and *paging* generated in the network, respectively. Their values are obtained as follows:

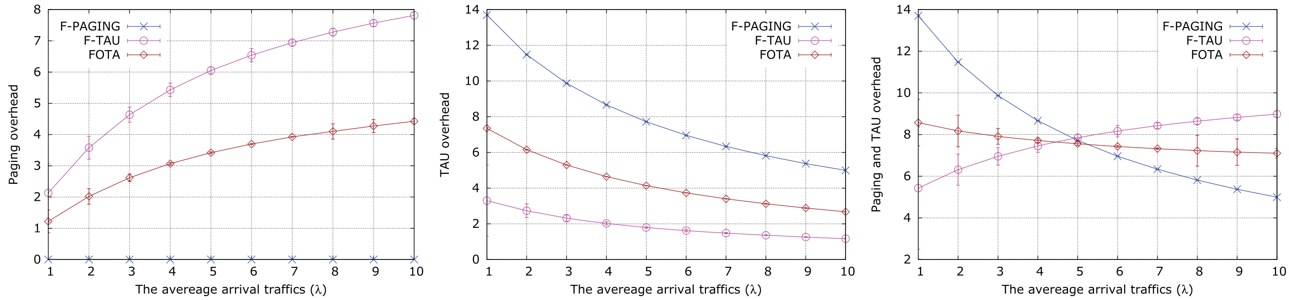
$$\begin{cases} N_{TAU} = \sum_{\text{i-k} \in \Theta} \left(\pi_{i,k} \sum_{\text{j-}\ell \in \Theta \wedge \ell \neq k \wedge i \neq j} B_{i,j,\ell} \right) \\ N_{paging} = \sum_{\text{i-k} \in \Theta} \pi_{i,k} C_i \sum_{j \in k \wedge i \neq j} \eta_j \end{cases}$$

VII. Performance Evaluation

In this section, we evaluate the performance of the three offline solutions *FOTA*, *F-TAU* and *F-PAGING*, by solving the analytical model. Then, we evaluate *ETAM* framework through simulation. Throughout this section, we fix the overhead of a single *TAU*, τ, to be ten times the value of ρ [14]. All solutions (i.e. *FOTA*, *F-TAU* and *F-PAGING*) are evaluated in terms of the following metrics:

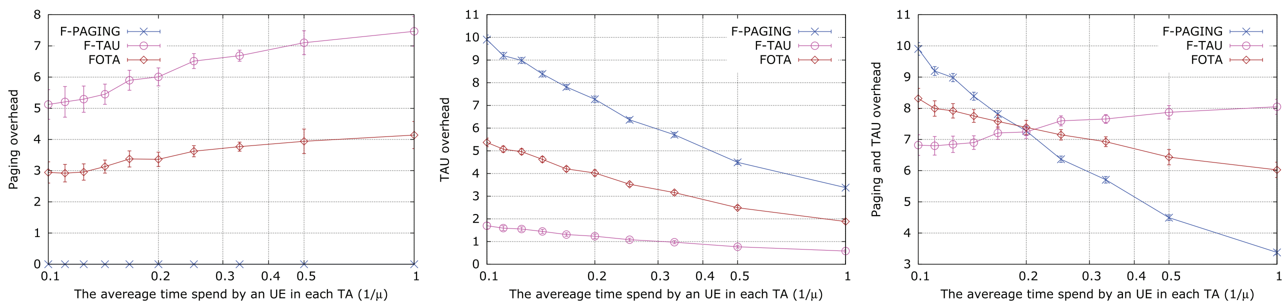
1) *TAU overhead:* the overhead of *TAU* messages (UP-Link) generated by *UEs* when visiting new *TALs*.
2) *Paging overhead:* the overhead of paging packets sent from *MME* to locate *UEs* during the call establishment.
3) *Total overhead:* the generated overhead due to both paging and *TAU*. The aim of this metric is to show the Pareto-efficiency between the *TAU* and paging overhead.

To evaluate *ETAM*, we divided the deployed area into a set of *TAs*, where each *TA* has a rectangular shape with a specific length and width. Note that *TA*s may have different surfaces according to their length and width. The mobility of UEs is modeled according to the Random Waypoint Mobility Model [18] with the pause-time sets to zero. Initially, we start the evaluation by placing each UE in a given *TA*. During the evaluation, each UE chooses a random destination (*TA*) in the deployed area and a speed that is uniformly distributed between $[avgSpeed - \Delta, avgSpeed + \Delta]$, where $avgSpeed$ is the average speed of different UEs and Δ is the variation in the speed between UEs. In the evaluation, we set Δ to 5 km/h. The UE then travels toward the newly chosen at the selected speed. This process is repeated until the evaluation time finishes. In the evaluation, we executed the online and the offline steps 10 times. The numerical results were obtained by solving



(a) Paging overhead. (b) *TAU* overhead. (c) Paging and *TAU* overhead.

Fig. 11. Performance of the proposed solutions as a function of λ.



(a) Paging overhead. (b) *TAU* overhead. (c) Paging and *TAU* overhead.

Fig. 12. Performance of the proposed solutions as a function of μ.

the Markov model corresponding to network model of Fig. 10, while the simulation were obtained through Matlab. Indeed, the simulator tool was implemented on top of Matlab and CVX (a package for disciplined convex optimization and geometric programming) [19]. In our evaluation, the sites (i.e., eNodeBs) are randomly deployed over the network. Without any loss of generality, we assume that the sites are already organized into *TAs* through any solution in the literature. The grouping of different sites into *TAs* is outside the scope of this paper.

A. Numerical Results

In this subsection, we present the numerical results, focusing on the impact of *TAU* and paging overhead on each solution by varying μ_i and λ. μ_i is the exponential distribution rate of the sojourn times of UEs in *TA* i, whereas λ is the average ratio of calls for a *UE* in the network. λ can be also defined as the exponential distribution rate of the inter arrival time between two consecutive calls for a UE. The latter refers to the percentage of time that a *UE* is called. Here, the term "call" refers not only to the classical voice call but also to data connection, such as VoIP and web applications. This parameter allows us to model the user activity in terms of active connections. Whereas, $\frac{1}{\mu_i}$ refers to the average time spent by a *UE* (i.e., sojourn time) in each *TA*. Increasing the values of μ_i corresponds to an increase in *UEs*' speeds and/or a decrease in the size of cells (micro-cell for $5G$ network) in the real world. Two scenarios are considered: (i) we vary λ from 1 to 10 while μ_i is fixed to 5; (ii) we vary μ_i from 1 to 10 while we fix λ to 5.

The *TAU*, *paging* and total overhead for each solution are evaluated using the following formulas:

$$\begin{cases} Overhead_{TAU} = \tau N_{TAU} \\ Overhead_{paging} = \rho N_{paging} \\ TotalOverhead = \tau N_{TAU} + \rho N_{paging} \end{cases}$$

Fig. 11 and Fig. 12 show the performance of the proposed solutions against increasing values of λ and $\frac{1}{\mu}$, respectively. As shown in Eq. 6, the increase of transitions probability of type "B" in EMC, reduces the sojourn time at each state in EMC. This results in a negative impact on *TAU* overhead and a positive impact on paging overhead, respectively. Whereas, the increase of transitions probability of type "C" in EMC, increases the sojourn time at each state in EMC. The latter has a positive impact on *TAU* overhead and a negative impact on paging overhead, respectively. The rise on λ values increases (resp., decreases) the transition probability of type "C" (resp., "B"), whereas the rise on μ values increases the transition probability of type "B" and decreases the transition probability of type "C".

For this reason, as depicted in Fig. 11(a) and Fig. 11(b), the increase of average arrival traffics (λ) has a negative impact on the paging overhead and a positive impact on the *TAU* overhead. Fig. 12(a) and Fig. 12(b) show that the increase of the sojourn time $(\frac{1}{\mu})$ in each *TA* has also a negative impact on the paging overhead and a positive impact on *TAU* overhead. Fig. 11(a) and Fig. 12(a) show that *F-PAGING* exhibits better performance than *FOTA* and *F-TAU* in terms of *TAU* overhead regardless the values of λ and $\frac{1}{\mu}$. This is attributable to the fact that the key objective of *F-PAGING* is to minimize paging overhead without tacking into account the *TAU* overhead. Whereas, Fig. 11(b) and Fig. 12(b) show that *F-TAU* exhibits better performance than *FOTA* and *F-PAGING* in terms of *TAU* overhead regardless the values of λ and $\frac{1}{\mu}$. This is obvious as *F-TAU* is designed to optimize the *TAU* overhead without tacking into account the paging overhead.

Fig. 11(c) and Fig. 12(c) show the total overhead due to both paging and *TAU* for different values of λ and $\frac{1}{\mu}$, respectively. *FOTA* achieves a tradeoff between the two conflicting objectives, i.e; reduction of both *TAU* and paging overhead. We observe from these figures that: (i) *F-TAU* has better performance in terms of total (i.e., paging and *TAU*) overhead when the values of λ and μ are below 5; (ii) *F-PAGING* has better performance when the values of λ and μ exceed 5. Indeed, the performance of *FOTA* is always between *F-TAU* and *F-PAGING*, whatever the values of λ and $\frac{1}{\mu}$. *FOTA* has performance similar to that of *F-TAU* when values of λ and μ are below 5 and similar to that of *F-PAGING* when values of λ and μ exceed 5. Thus, *FOTA* always finds an optimal tradeoff between *TAU* and paging overhead by maintaining the total overhead near to the optimal value regardless the *UEs*' behavior. This demonstrates that it successfully achieves the key objective of its design.

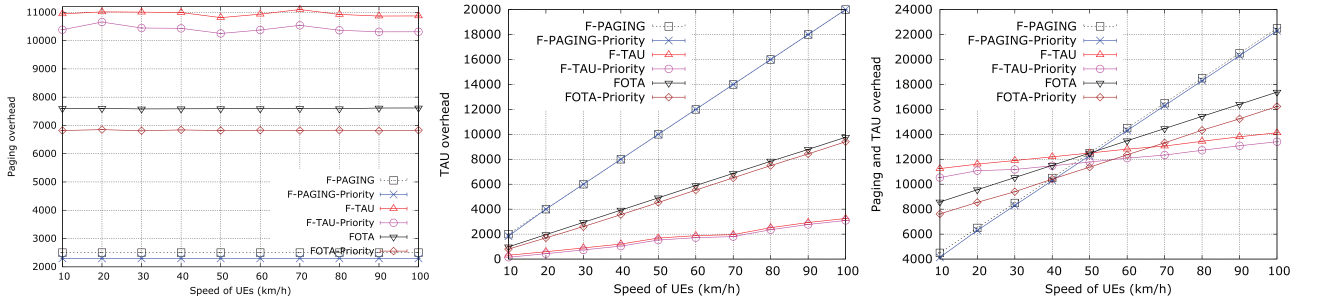
B. Simulation Results

In this subsection, the proposed schemes are evaluated through simulations. We used the proposed framework (ETAM) to evaluate through simulation the three solutions (*F-PAGING*, *F-TAU* and *FOTA*) of offline step and the two solutions of online step. Formally, we have six possible combinations of protocols. The same trajectory logs of UEs are used to evaluate the different combinations of protocols. The information of handover between different *TAs* is forwarded from the online to the offline step. During the movement of a UE, a *TAU* message is generated and sent to *MME* every-time a UE crosses a *TA* that does not belong to its *TAL* in the online step. The optimization problems are solved considering different values of the average speed $avgSpeed$ of *UEs* and the average ratio of calls of each *UE* in the network. The average speed of *UEs* shows the impact of *TAUs* signaling on the different optimization problems. In the simulation evaluation, we evaluate two scenarios: (i) we vary the average speed $avgSpeed$ of *UEs* and fix the average call ratio to 50 $calls/h$ for each *UE* in the network; (ii) we vary the average call ratio of *UEs* and fix the average speed $avgSpeed$ of *UEs* to 50 km/h. In contrast to the analysis part, the two solutions of online part are considered in the simulation evaluation: (i) *UEs* pick their *TAL* without any prioritization; (ii) each *UE* picks a *TAL* with prioritization, according to its behavior, to reduce the overhead of *TAU* and paging signaling.

Fig. 13 and Fig. 14 show the resilience of *FOTA*, *F-TAU* and *F-PAGING* against increase in *UEs*' speed and call ratio, respectively. We clearly observe that assigning *TALs* to *UEs* with prioritization (e.g., per *UE*'s activities - mobility and call ratio) has a positive impact on the performance of the three solutions. From Fig. 13(c), for the speed of *UEs* equals to 70 km/h, we observe that the selection of *TALs* with prioritization reduces the total overhead from 13060 to 12340 (an enhancement with more than 5.51%) for *F-TAU*, and for *FOTA* the total overhead is reduced from 14460 to 13321, which means an enhancement exceeding 7.87%. Meanwhile from Fig. 14(c), we observe that when the call ratio equals to 90 $call/h$, the selection of *TALs* with prioritization reduces the total overhead of *FOTA* from 18556 to 16336, which means an enhancement exceeding 11.96%.

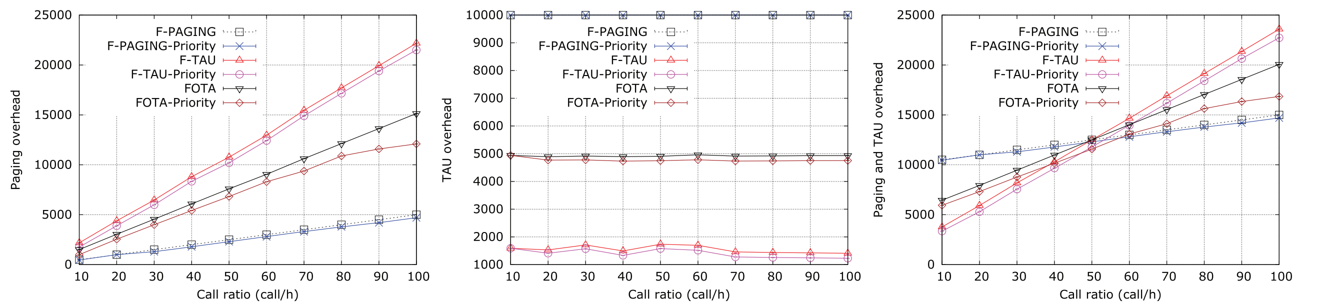
Fig. 13(b) and Fig. 13(c) show that the speed of *UEs* has a negative impact on *TAU* and total overhead, respectively. This behavior is expected as highly mobile users perform frequently handoff between *TAs* and ultimately generate high *TAU* messages. Thus, the higher the speed of *UEs* is, the higher the *TAU* overhead becomes. Further, we remark from Fig. 13(b) that *F-TAU* exhibits better performance than *FOTA* and *F-PAGING* in terms of *TAU* overhead regardless the speed of *UEs*. This is attributable to the fact that the key objective of *F-TAU* is to minimize *TAU* overhead without tacking into account the paging overhead. Whereas, Fig. 14(a) and Fig. 14(c) demonstrate that the call ratio has a negative impact on paging and total overhead, respectively. This is also predictable as highly active *UEs* (i.e., with high call ratios) cause high number of paging messages when they go in the idle mode and their locations are searched the network. Moreover, from Fig. 14(a), we observe that *F-PAGING* exhibits better performance than *FOTA* and *F-TAU* in terms of paging overhead regardless the call ratio. This is intuitively due to the fact that *F-PAGING* is designed to optimize the paging overhead without tacking into account the *TAU* overhead.

Fig. 13(c) and Fig. 14(c) illustrate the tradeoff achieved by *FOTA* between the two conflicting objectives, i.e; reduction of both *TAU* and paging overhead. They show the total overhead incurred in the three solutions and that is for different values of the *UE* speed and call ratio, respectively. We observe from these figures that: (i) *F-PAGING* exhibits better performance in terms of total (i.e., paging and *TAU*) overhead when the speed of *UEs* is below 50 km/h or when the call ratio exceeds 50 $calls/h$; (ii) *F-TAU* exhibits better performance when the average speed of *UEs* exceeds 50 km/h or when the call ratio does not exceed 50 $calls/h$; and (iii) *FOTA* has performance similar to that of *F-PAGING* when the speed of *UEs* is below 50 km/h or when the call ratio exceeds 50 $calls/h$. It is also observed that *FOTA* performs similarly to *F-TAU* when the call ratio does not exceed 50 $calls/h$ or the speed of *UEs* exceeds 50 km/h. Indeed, the performance of *FOTA* is always between *F-TAU* and *F-PAGING*, depending on the *UEs*' speed and their activity levels (i.e., call rate). For highly mobile *UEs*, *FOTA* performs similar to *F-TAU* (optimal) and better than *F-PAGING*, whilst for highly active *UEs*, *FOTA* performs similar to *F-PAGING* (optimal) and better than *F-TAU*. *FOTA* always finds an optimal tradeoff between *TAU* and paging overhead by maintaining the total overhead near to the optimal value regardless the *UEs*'



(a) Paging overhead. (b) *TAU* overhead. (c) Paging and *TAU* overhead.

Fig. 13. Performance of the proposed solutions as a function of speed of *UEs*.



(a) Paging overhead. (b) *TAU* overhead. (c) Paging and *TAU* overhead.

Fig. 14. Performance of the proposed solutions as a function of the call ratio.

behavior. This demonstrates that it successfully achieves the key objective of its design.

It is worth noting that we observe some differences between the simulation and the numerical results. In contrast to the simulations, varying the average of traffic arrival rate λ has an impact on *TAU* overhead and varying the average sojourn time $(\frac{1}{\mu})$ in each *TA* has an impact on the paging overhead. This is because in the analysis, the behavior of the network is shown as a ratio between λ and μ. Any increase in any of one of them has a negative impact on the other.

VIII. Conclusion

One key vision of the upcoming 5G is to support potential numbers of users connecting to the mobile networks. An important challenge is to cope with the amount of signaling to be generated by these mobile users, particularly signaling messages due to mobility (i.e., *TAU*) and for connection setup (i.e., paging). Particularly, the mentioned overhead could be exacerbated if small cells are deployed (as envisioned in the upcoming 5G) . To overcome this issue, we have devised the *ETAM* framework, which aims at mitigating the effect of *TAU* and paging signaling messages on the network. *ETAM* has two parts, one is executed online and another is executed offline. In the online part, we proposed two strategies to assign *TALs* to *UEs*, whereas in the offline part three solutions are proposed. Analysis and simulation results have proven the efficiency of each solution in achieving its key design objectives.

References

[1] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: Potential, challenges, and solutions," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 178–184, Mar. 2012.

[2] M. Bagaa, A. Ksentini, T. Taleb, R. Jantti, A. Chelli, and I. Balasingham, "An efficient D2D-based strategies for machine type communications in 5G mobile systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC'16)*, Apr. 2016.

[3] T. Taleb, "Towards carrier cloud: Potential, challenges, & solutions," *IEEE Wireless Commun. Mag.*, vol. 21, no. 3, pp. 80–91, Jun. 2014.

[4] T. Taleb et al., "EASE: EPC as a service to ease mobile core network," *IEEE Netw. Mag.*, vol. 29, no. 2, pp. 78–88, Mar. 2015.

[5] Y. Zhang and M. Fujise, "Location management congestion problem in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 942–954, Mar. 2007.

[6] Mitsubishi Electric, "Tracking areas size & tracking area list optimization," Technical report 3GPP TSG RAN WG3 meeting, R3-071931, 2007.

[7] S. Modarres Razavi and D. Yuan, "Mitigating mobility signaling congestion in LTE by overlapping tracking area lists," in *Proc. ACM Int. Conf. Model. Anal. Simul. Wireless Mobile Syst. (MSWIM'11)*, 2011, pp. 285–291.

[8] Y. W. Chung, "Adaptive design of tracking area list in LTE," in *Proc. IEEE 8th Int. Conf. Wireless Opt. Commun. Netw. (WOCN'11)*, May 2011, pp. 1–5.

[9] S. Razavi, D. Yuan, F. Gunnarsson, and J. Moe, "Exploiting tracking area list for improving signaling overhead in LTE," in *Proc. IEEE Veh. Technol. Conf. (VTC'10)*, May 2010, pp. 1–5.

[10] S. Razavi, Y. Di, F. Gunnarsson, and J. Moe, "Dynamic tracking area list configuration and performance evaluation in LTE," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2010, pp. 49–53.

[11] 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, 2012.

[12] H.-W. Kang, W.-J. Kim, S.-J. Koh, H.-G. Kang, and J.-B. Moon, "Configuration of tracking area code (TAC) for paging optimization in mobile communication systems," in *Ubiquitous Information Technologies and Applications*. New York, NY, USA: Springer, 2014, pp. 59–66.

[13] H.-W. Kang, H.-G. Kang, and S.-J. Koh, “Optimization of TAC configuration in mobile communication systems: A tabu search approach,” in *Proc. IEEE Int. Conf. Adv. Commun. Technol. (ICACT’14)*, 2014, pp. 5–9.

[14] K. Kyamakya and K. Jobmann, “Location management in cellular networks: Classification of the most important paradigms, realistic simulation framework, and relative performance analysis,” *IEEE Trans. Veh. Technol.*, vol. 54, no. 2, pp. 687–708, Mar. 2005.

[15] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[16] J. Nash, “The bargaining problem,” *Econometrica*, vol. 18, no. 2, pp. 155–162, Apr. 1950.

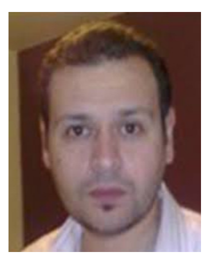
[17] Y. Zhao, S. Wang, S. Xu, X. Wang, X. Gao, and C. Qiao, “Load balance vs energy efficiency in traffic engineering: A game theoretical perspective,” in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 530–534.

[18] T. Camp, J. Boleng, and V. Davies, “A survey of mobility models for ad hoc network research,” *Wireless Commun. Mobile Comput.*, vol. 2, no. 5, pp. 483–502, 2002.

[19] M. Grant and S. Boyd, “Graph implementations for nonsmooth convex programs,” in Recent Advances in Learning and Control, V. Blondel, S. Boyd, and H. Kimura, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 95–110.

Miloud Bagaa received the Engineer’s, Master’s, and Ph.D. degrees from the University of Science and Technology Houari Boumediene (USTHB), Algiers, Algeria, in 2005, 2008, and 2014, respectively. From 2009 to 2015, he was a Researcher with the Research Center on Scientific and Technical Information (CERIST), Algiers, where he was a Member of the Wireless Sensor Networks Team, DTISI Division. From 2015 to 2016, he was granted a postdoctoral fellowship from the European Research Consortium for Informatics and Mathematics, and worked with the Norwegian University of Science and Technology, Trondheim, Norway. Currently, he is Senior Researcher with the Communications and Networking Department, Aalto University, Espoo, Finland. His research interests include wireless sensor network, Internet of Things, 5G wireless communication, security, and networking modeling.

Tarik Taleb (S’05–M’05–SM’10) received the B.E. degree in information engineering (with distinction), the M.Sc. and Ph.D. degrees in information sciences from GSIS, Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. He is currently a Professor with the School of Electrical Engineering, Aalto University, Espoo, Finland. Prior to his current academic position, he was a Senior Researcher and 3GPP Standards Expert with NEC Europe Ltd., Heidelberg, Germany. He was then leading the NEC Europe Labs Team working on R&D projects on carrier cloud platforms, an important vision of 5G systems. He was also serving as Technical Leader of the main work package, Mobile Core Network Cloud, in EU FP7 Mobile Cloud Networking project, coordinating among 9 partners including NEC, France Telecom, British Telecom, Telecom Italia, and Portugal Telecom. Before joining NEC and until March 2009, he worked as an Assistant Professor with the Graduate School of Information Sciences, Tohoku University, in a laboratory fully funded by KDDI, the second largest network operator in Japan. From October 2005 until March 2006, he worked as a Research Fellow with the Intelligent Cosmos Research Institute, Sendai, Japan. His research interests include architectural enhancements to mobile core networks (particularly 3GPPs), mobile cloud networking, network function virtualization, software-defined networking, mobile multimedia streaming, intervehicular communications, and social media networking. He has been also directly engaged in the development and standardization of the evolved packet system as a Member of 3GPP’s System Architecture working group. He is an IEEE Communications Society (ComSoc) Distinguished Lecturer. He is a Member of the IEEE Communications Society Standardization Program Development Board. In an attempt to bridge the gap between academia and industry, he founded the IEEE Workshop on Telecommunications Standards: From Research to Standards, a successful event that was recognized with the Best Workshop Award by the IEEE Communication Society (ComSoC). Based on the success of this workshop, he has also founded and has been the Steering Committee Chair of the IEEE Conference on Standards for Communications and Networking. He is the General Chair of the 2019 edition of the IEEE Wireless Communications and Networking Conference (WCNC’19) to be held in Marrakech, Morocco. He is/was on the Editorial Board of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the *IEEE Wireless Communications Magazine*, the IEEE JOURNAL ON INTERNET OF THINGS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE COMMUNICATIONS SURVEYS & TUTORIALS, and a number of Wiley journals. He is serving as the Chair of the Wireless Communications Technical Committee, the largest in IEEE ComSoC. He also served as the Vice Chair of the Satellite and Space Communications Technical Committee of the IEEE ComSoc (2006–2010). He has been on the technical program committee of different IEEE conferences, including Globecom, ICC, and WCNC, and chaired some of their symposia. He was the recipient of the 2009 IEEE ComSoc Asia-Pacific Best Young Researcher Award (June 2009), the 2008 TELECOM System Technology Award from the Telecommunications Advancement Foundation (March 2008), the 2007 Funai Foundation Science Promotion Award (April 2007), the 2006 IEEE Computer Society Japan Chapter Young Author Award (December 2006), the Niwa Yasujirou Memorial Award (February 2005), and the Young Researcher’s Encouragement Award from the Japan chapter of the IEEE Vehicular Technology Society (VTS) (October 2003). Some of his research works have been also awarded Best Paper Awards at prestigious conferences.



Adlen Ksentini (SM’14) received the M.Sc. degree in telecommunication and multimedia networking from the University of Versailles Saint-Quentin-en-Yvelines, Versailles, France, and the Ph.D. degree in computer science from the University of Cergy-Pontoise, Cergy-Pontoise, France, in 2005. From 2006 to 2015, he worked with the University of Rennes 1, Rennes, France, as an Associate Professor. During this period, he was a Member of the Dionysos Team with INRIA, Rennes, France. Recently, he joined the Mobile and Wireless Networking Department, EURECOM Institute, as an Associate Professor. He has been involved in several national and European projects on QoS and QoE support in future wireless, network virtualization, cloud networking, and mobile networks. He has coauthored over 100 technical journal and international conference papers. He has been acting as the TPC Symposium Chair for the IEEE ICC 2016 and 2017. He was a Guest Editor of the *IEEE Wireless Communications Magazine* the *IEEE Communications Magazine*, and two ComSoc MMTC letters. He has been on the Technical Program Committee of major IEEE ComSoc, ICC/Globecom, ICME, WCNC, and PIMRC conferences. He was the recipient of the Best Paper Award from the IEEE ICC 2012 and ACM MSWiM 2005.