This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2009 proceedings

The Spatial Effect of Mobility on the Mean Number of Handoffs: A New Theoretical Result

Wolfgang Bziuk, Said Zaghloul and Admela Jukan
Technische Universität Carolo-Wilhelmina zu Braunschweig, {bziuk, zaghoul, jukan}@ida.ing.tu-bs.de

***Abstract*— The mean number of handoffs is a fundamental performance measure in any mobile system, as it directly relates to the signaling load in the network as well as to the delivered QoS. As the mobile services are evolving from simple cellular voice calls towards media and data sessions, and as cellular providers are reinventing their businesses by incorporating third party services, the handoff rate will even play a more pivotal role. This is because future sessions are expected to last longer than voice calls and users are more likely to roam into other networks. Existing results provide an estimate of the mean number of handoffs in networks composed of an infinite number of access gateways and hence consider neither the topological arrangement of the gateways nor the mobility patterns between them. In this paper, we obtain a closed form solution for the mean number of handoffs under generic assumptions of two-dimensional Markovian mobility patterns, spatial arrangement of the access gateways, as well as generic session times and gateway residence times. Our solution unveils a new insight into mobility foundations as it shows that the consideration of the mobility pattern and the access gateways' layout simply transforms the known equation for the mean number of handoffs of an infinite network size from scalar to vector representation. We demonstrate that the mean number of handoffs is a non-linear function of the gateway spatial arrangement and user mobility.**

I. Introduction

Users' mobility has been and will continue to be a core design challenge to the wireless mobile systems. In these systems, the mean number of handoffs is a fundamental performance measure, as it directly relates to the signaling load in the network as well as to the delivered Quality of Service (QoS). Although handoffs in radio layers have been standardized and successfully implemented, the impact on higher layers is not well understood. The higher layer signaling and QoS provision are triggered when users move between the so-called IP Access Gateways (AGW), which typically serve a group of base stations (see Fig.1a). Examples of gateways in the current systems are GGSNs in 3GPP systems, ASNs in WiMAX systems, and PDSNs in 3GPP2 systems. As users move between gateways, MobileIP signaling is triggered to maintain IP connectivity, Diameter/RADIUS signaling is triggered for authentication, and SIP/QoS signaling [1] is triggered to authorize the session at the target gateway, as standardized within the IP Multimedia Sub-system (IMS) [2]. Therefore, in these systems the handoffs between gateways, resulting from handoffs between border base stations of each gateway area, are no longer limited to few SS7 exchanges, but rather extended to a melange of application layer signaling protocols, with impact on the signaling load, handoff latency, and session dropping probabilities.

Past established theories [3], [4], [5] in circuit switched cellular systems cannot be easily used to assess the mean number of handoffs generically. First, they majorly focused on aspects of call performance such as the estimation of the effective call duration and call blocking in macro and micro cellular systems. In such studies, mobility was only characterized by the mobility factor defined as the ratio of the mean session duration to the mean time a user spends in a cell (i.e., the cellular residence time). Second, due to the limited voice call durations such systems were assumed infinite in size and hence users are always served by one network for the entire call duration. Hence, neither the topology nor the mobility pattern were considered. In next generation systems, past assumptions on the session characteristics no longer hold as users are likely to experience long session durations and hence are more likely to change their serving network. Therefore, estimating the mean number of handoffs between gateways depends not only on the mobility factor but also on the mobility pattern among access gateways and their arrangement in the network. This requires revisiting the basics of past research and extending them to account for such effects.

In this article, we derive for the first time the closed form expression for the handoff rate under generic assumptions of two-dimensional arrangements of access gateways, session time and access gateway residence time distributions, and arbitrary mobility patterns. Our framework builds on established techniques from cellular networks[3], [4] and intertwines it with concepts of pixel based mobility behaviors used in realistic simulation studies [6] using transient Markov techniques and complex analysis. We obtain a highly interesting result which reveals the fact that the known equation for the mean number of handoffs in the literature simply translates from scalar to vector representation when considering mobility patterns and gateway arrangements. In fact, we show that it is no longer sufficient to represent mobility by the residence time, as in the current body of literature, but rather as a product of a scalar quantity representing the residence time and a spatial vector component corresponding to the topology and the mobility pattern. This paper is a major extension of the preliminary results we presented in [7], [8], as we here consider generic mobility patterns, session times, and two dimensional network layouts. The model is yet elegant, as its computational complexity is only a function of the number of gateways and does not depend on the residence- and session time statistical distributions.

This paper is organized as follows. In Section II, we present

978-1-4244-3435-0/09/$25.00 ©2009 IEEE

the analytical model. In Section III, we show numerical results. Section IV concludes the paper.

II. Analytical Model

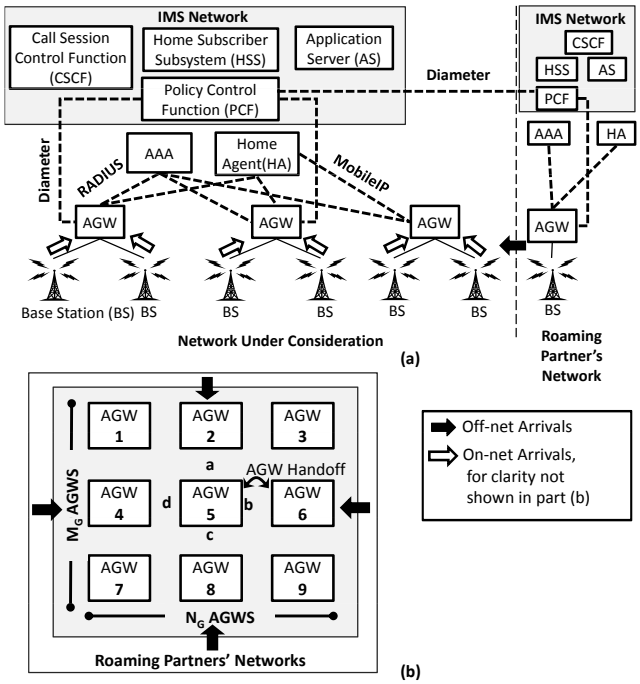

Fig. 1. (a) Network under consideration, and (b) sample topology with $n = 9$ Access Gateways (AGW); the borders of every gateway are marked a, b, c, d corresponding to north, east, south and west movements (e.g., AGW 5).

In this section, we derive the mean number of handoffs (MNH) for a session either entirely or partially served by a network of limited size (i.e., the network under consideration), which is characterized by a square area served by $M_g \times N_g$ gateways. The mobility model which describes movements between AGWs is based on the concepts of pixel based mobility modeling [6], developed for realistic mobility simulations. Our mobility model is versatile as it can represent a wide range of movement patterns ranging from directed patterns (e.g., highways) to completely random movements. In the analytical approach presented here, we start by jointly considering session time, residence time, and mobility patterns using a transient Markov chain with an infinite state space. Based on the structure of the state space, we then decompose the Markov chain, which results in a closed form solution incorporating the spatial component of the mobility, i.e., the gateways layout and network topology. As in our past work [8], to simplify the model description we distinguish between the so-called *on-net* and *off-net* traffic. On-net traffic refers to any session which start within the network under consideration. Off-net traffic, on the other hand, refers to sessions which start elsewhere, and roams into the network under consideration at some random point of time. To facilitate the discussion, we will first discuss the traffic starting from within the network (on-net traffic) and then briefly discuss the simplicity of incorporating off-net traffic at the end of Sec. II.

A. Assumptions

- The session duration, S, has density $f_S(t)$, a rational Laplace transform $f_S^*(s)$ and a mean of E_s.
- The gateway residence times, R, are independent and identically distributed. R is generally distributed with an existing Laplace transform $f_R^*(s)$ and mean of E_R.
- For simplicity, we assume that sessions are always resumed after handoffs. The inclusion of blocking is straightforward and can be carried out as in [3].

B. Handoff Probability Preliminaries

Since a new session starts inside an AGW coverage, the residence time in the first AGW, R_1, is different from the subsequent ones. Assuming that R_1 is given by the residual of the residence time, then its density and Laplace transform are given as $f_{R_1}(t) = \frac{1-f_R(t)}{E_R}$, $f_{R_1}^*(s) = \frac{1-f_R^*(s)}{sE_R}$ respectively. Let $R(k) = R_1 + \sum_{j=2}^{k} R$ denote the sum of residence times incurred since the session starting event until the k^{th} handoff event. *R(k)* has Laplace transform of $f_k^*(s) = f_{R_1}^*(s)\left(f_R^*(s)\right)^{k-1}$, $k \geq 1$. It follows that the probability that a call has already incurred *k*-1 handoffs and that its session duration is large enough to include at least one more handoff is [4],

$$P_H(k) = P\{R(k) \leq S \,|R(k-1) \leq S\} = \frac{G(k)}{G(k-1)} \quad (1)$$

where $G(0) = 1$. $G(k)$ is the probability that the session contains at least k handoffs ($k \geq 1$) (i.e., $P\{R(k) \leq S\}$) and is calculated as [4], [9],

$$G(k) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_{R_1}^*(s)\left(f_R^*(s)\right)^{k-1}}{s} f_S^*(-s)\,ds \quad (2)$$

Furthermore, if $f_S^*(s)$ is a rational function and if the set of poles, Ξ_{S-}, of $f_S^*(-s)$ only contains poles s_p in the right half side of the complex plane (i.e. $\Re\{s_p\} > 0$), the residue theorem can be applied to solve the integral (2). Let $\mathrm{Res}_{s=s_p}$ denote the residue at pole $s = s_p$, then we have [4],

$$G(k) = -\sum_{s_P \in \Xi_{S-}} \underset{s=s_p}{Res} \frac{f_{R_1}^*(s)}{s}\left(f_R^*(s)\right)^{k-1} f_S^*(-s) \quad (3)$$

Using (1)-(3), it can be shown that the mean number of handoffs during the whole session is given as [9], [12],

$$E\{HO\} = -\sum_{s_P \in \Xi_{S-}} \underset{s=s_p}{Res} \frac{f_{R_1}^*(s)}{(1-f_R^*(s))} \frac{f_S^*(-s)}{s} \quad (4)$$

C. Mobility Model

During a session, a mobile node may traverse multiple AGW areas. To simplify the discussion, let us first consider an area consisting of rectangularly arranged AGWs (i.e., $M_g \times N_g$ number of gateways within the network under consideration (see Fig.1b)). In the end of this section we show how to relax this assumption. When a mobile enters this area its future movement is described by a set of transition probabilities which depend on the entering and exit borders [6]. Thus each

AGW is described by 4×4 transition probabilities. Using the border labeling shown in Fig.1b for AGW 5, the transition probabilities are denoted as p_{jxy}, where j denotes the AGW, x and y define the entering and the exit borders respectively such that $(x, y) \in \{a, b, c, d\}$. The transition probabilities can be arranged into a set of two different one step transition matrices, one that defines the initial transition probabilities $\mathbf{P_{MI}}$ (i.e., when the session is first served by the network) and another, $\mathbf{P_M}$ that defines the transition probabilities afterwards (i.e., after the initial handoff) as,

$$\mathbf{P_{MI}} = \begin{pmatrix} \mathbf{Q_{MI}} & \mathbf{A_{MI}} \end{pmatrix}, \mathbf{P_M} = \begin{pmatrix} \mathbf{Q_M} & \mathbf{A_M} \end{pmatrix} \quad (5)$$

Assuming a network of $n = M_g N_g$ access gateways, the matrix $\mathbf{Q_M}$ describes the movement of a handoff session between AGWs and has $4n \times 4n$ elements describing movements between neighboring AGWs. For example, a session leaving AGW 5 in Fig.1b at border a enters AGW 2 at border c. Thus the exit border a of AGW j is linked to the entry border c of AGW $(j - N_g)$, where N_g is the number of AGWs in a row. Let us number the columns and rows of $\mathbf{Q_M}$ by the entry borders of the AGWs as $1a, 1b, 1c, 1d, 2a, 2b, 2c, 2d, ..., na, nb, nc, nd$. For a given entry border there are up to four possible transitions. For instance, entering from border a in the j^{th} AGW (i.e. $(ja)^{th}$ row of $\mathbf{Q_M}$), the transition probabilities (which sum to one) are $p_{jaa}, p_{jab}, p_{jac}, p_{jad}$. They have the column numbers $(j - N_g)\, c$, $(j + 1)\, d$, $(j + N_g)\, a$ and $(j - 1)\, b$, respectively, [6]. Additionally for an AGW at the network boundary, all transitions to the roaming partners (V) are listed in the $1 \times 4n$ matrix $\mathbf{A_M}$. An example for AGW 2 (rows $(2a)$ and $(2b)$) is shown for the matrixes $\mathbf{Q_M}$ and $\mathbf{A_M}$ below, where the state numbering is added on top and to the right side for clarity.

$$\mathbf{Q}_M = \begin{array}{c} \begin{matrix} (1a) & (1b) & \cdots & (3d) & \cdots & (5a) & \cdots \end{matrix} \\ \left(\begin{array}{ccccccc} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & p_{2ad} & \cdots & p_{2ab} & \cdots & p_{2ac} & \cdots \\ 0 & p_{2bd} & \cdots & p_{2bb} & \cdots & p_{2bc} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right) \end{array} \begin{matrix} (V) \\ \vdots \\ (2a) \\ (2b) \\ \vdots \end{matrix}, \quad \mathbf{A}_m = \begin{pmatrix} \vdots \\ p_{2aa} \\ p_{2ba} \\ \vdots \end{pmatrix}$$

The matrix $\mathbf{Q_{MI}}$ is of dimension $n \times 4n$ and contains the transition probabilities for a new session. A session starting in AGW j and leaving at border y corresponds to the jth row of $\mathbf{Q_{MI}}$ and has transition probabilities $\hat{p}_{ja}, \hat{p}_{jb}, \hat{p}_{jc}, \hat{p}_{jd}$. Again, all transitions to the roaming partners created by boundary AGWs are combined in the matrix $\mathbf{A_{MI}}$. Finally, it should be noted that our analysis is independent of the grid like arrangement of AGWs. For other AGW arrangements (e.g., irregular arrangements of AGW areas with more than 4 neighbors), we simply add rows and columns with the corresponding transition probabilities for each edge in the mobility matrices in (5).

D. Mean Number of Handoffs

In this subsection, we derive the mean number of handoffs, $E\{N_H\}$, for sessions partially served by a network comprised of $M_g \times N_g$ AGWs by majorly extending the transient Markov model in our previous work [8] to incorporate the mobility model described in Section II.C and to account for two dimensional AGW arrangements. In our context, a transient Markov chain contains transient and absorbing states. For a transient state, there is a chance of never being revisited after the initial visit. This is due to the existence of absorbing states that are never abandoned once entered. In our model, transient states represent the gateways inside the network, while the absorbing states represent departures from the network area to roaming partners' networks or session termination. Since it is practically infeasible to track sessions in roaming partners' networks, departing sessions return only as off-net arrivals (see [8] for more details). Let the transition probabilities $\mathbf{P}$ be,

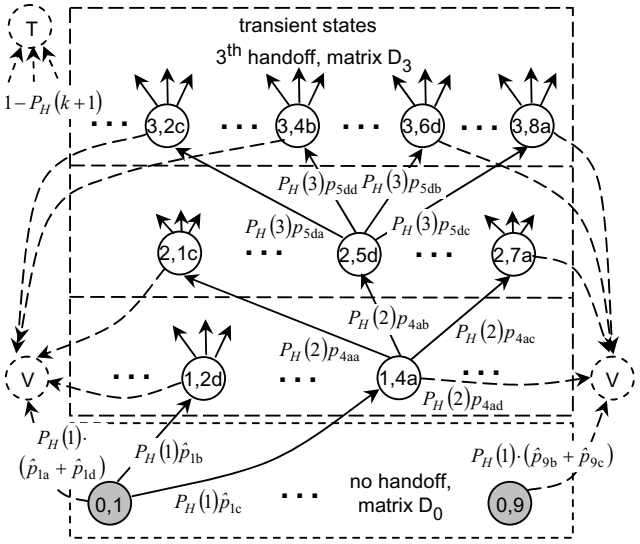

Fig. 2. Model diagram for the network of Fig.1b with selected states and transitions only (all transient states can reach state T, state V is drawn twice for clarity, initial states shaded, transient states dashed).

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{A} \\ \mathbf{0} & \mathbf{I} \end{pmatrix},$$

where $\mathbf{Q}$ contains only transitions between transient states, $\mathbf{A}$ contains only transitions to the absorbing states Λ_i, and $\mathbf{I}$ is the identity matrix with proper dimensions. Let the row vector $\mathbf{P_S}$ denote the initial probabilities for the transient chain. Using the fundamental matrix $\mathbf{M}$, given as $\mathbf{M} = (\mathbf{I} - \mathbf{Q})^{-1}$, the mean number of visits to transient states, excluding the first and before absorption is [8], [10],

$$E\{N\} = \mathbf{P_S M e}^T - 1, \quad (6)$$

where $\mathbf{e} = \{1, 1, ..., 1\}$ and $\mathbf{e}^T$ denotes the transpose. The -1 in (6) is subtracted because $\mathbf{P_S M e}^T$ includes the initial visit (i.e., the session start). The probability of being absorbed into state Λ_i is [10],

$$\beta_i = \mathbf{P_S M} \|\mathbf{A}\|_i \textit{ where } \|\mathbf{A}\|_i \textit{ is the } i^{th} \textit{column of } \mathbf{A} \quad (7)$$

Since we deal with generally distributed session times when calculating the likelihoods of making future handoffs, we must track the number of completed handoffs. Thus our transient chain consists of 2-tuple transient state definitions,

where $(0,j)$ represents a session starting inside the j^{th} AGW, while the state (k,jx) is assigned to a session entering the j^{th} AGW through a border $x \in \{a,b,c,d\}$ after completing k handoffs. We also define two absorbing states: V representing a session leaving the domain to the roaming partners and T representing the session termination. Fig.2 shows a short summary of the model, where the initial states are shaded and the transitions to the absorbing states are dashed. For a session starting in AGW j (i.e., state $(0,j)$), the transition probabilities are given by the joint event comprised of the transition probabilities $\hat{p}_{jy}$ for the initial movement (summarized in matrix $\mathbf{Q_{MI}}$), and the probability that the session contains at least one more handoff given that it made no handoffs, $P_H(1)$. For example in Fig.2, a session starting in AGW 1 (i.e., state $(0,1)$), leaving through border c and handing over to border a of AGW 4 (state $(1,4a)$) is described by the transition probability $P_H(1)\hat{p}_{1c}$. With (1) and (5) all transient transitions from $(0,j)$ to $(1,ix)$ are written in matrix form as $\mathbf{D_0} = P_H(1)\mathbf{Q_{MI}}$, where $\mathbf{D_0}$ is a $n \times 4n$ matrix. Similarly, the initial transitions from transient states $(0,j)$ to the absorbing state V are described by $\mathbf{A_0} = P_H(1)\mathbf{A_{MI}}$. Otherwise, if the session has terminated, the chain goes to the absorbing state T with a probability of $1 - P_H(1)$. Now consider the example, that a session enters AGW 5 through border d after the second handoff (i.e., state $(2,5d)$ in Fig.2). If the session leaves the AGW through border b, the transition probability to neighboring border d of AGW 6 is given by $P_H(3)p_{5db}$. Using (1) and (5) all transient transitions out of states (k,jx) to states $(k+1,iy)$ can be written in matrix form as $\mathbf{D_k} = P_H(k+1)\mathbf{Q_M}$, where $\mathbf{D_k}$ is a $4n \times 4n$ matrix. The transitions from transient states to the absorbing state V are similarly described by $\mathbf{A_k} = P_H(k+1)\mathbf{A_M}$. Ordering states lexographically as $(0,1),...,(0,n),(0,1a),...,(0,nd),(1,1a),...,(1,nd),...,$ $(k,1a),...,(k,nd)...$, the transient Markov chain is given as,

$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{D_0} & 0 & 0 & 0 & \cdots \\ 0 & 0 & \mathbf{D_1} & 0 & 0 & \cdots \\ 0 & 0 & 0 & \mathbf{D_2} & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \mathbf{A} = \begin{bmatrix} \mathbf{A_0} \\ \mathbf{A_1} \\ \mathbf{A_2} \\ \vdots \end{bmatrix} \quad (8)$$

where $\mathbf{Q}$ is a matrix of unlimited size since the number of handoffs k can go to infinity. The elements of (8) are,

$$\begin{aligned} \mathbf{D_0} &= P_H(1)\mathbf{Q_{MI}} \quad , \mathbf{A_0} = P_H(1)\mathbf{A_{MI}} \\ \mathbf{D_k} &= P_H(k+1)\mathbf{Q_M} \quad , \mathbf{A_k} = P_H(k+1)\mathbf{A_M} \ , \ k \geq 1 \end{aligned} \quad (9)$$

Let the initial state probabilities for new sessions be defined as $\mathbf{P_S} = [\mathbf{P_I}, 0, 0, ...]$ where $\mathbf{P_I} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]$ and ε_j represent the probability of starting a session from AGW j, then using (6) and (7), the mean number of handoffs before leaving the network, $E\{N\}$, to the roaming partners, $E\{V_H\}$, and in total, $E\{N_H\}$, are,

$$E\{N\} = \mathbf{P_S}\mathbf{M}\mathbf{e}^T - 1, E\{V_H\} = \beta_V, E\{N_H\} = E\{N\} + \beta_V \quad (10)$$

E. From Scalar to Vector Representation

Although (10) gives the solution for the MNH, the matrix $\mathbf{M} = (\mathbf{I} - \mathbf{Q})^{-1}$ may have considerable size resulting in computational and numerical issues. In the following proposition we avoid the inversion operation by studying (8) as,

Proposition 2.1: Let the matrix $\mathbf{Q}$ of (8) be truncated to an arbitrary finite number of handoffs k less than K (i.e. $\mathbf{Q}$ contains sub-matrices up to $\mathbf{D_K}$), then the matrix $\mathbf{M} = (\mathbf{I} - \mathbf{Q})^{-1}$ has the structure

$$\mathbf{M} = \begin{pmatrix} \mathbf{I} & \mathbf{D_0} & \mathbf{D_0D_1} & \mathbf{D_0D_1D_2} & \mathbf{D_0D_1D_2D_3} & \cdots \\ 0 & \mathbf{I} & \mathbf{D_1} & \mathbf{D_1D_2} & \mathbf{D_1D_2D_3} & \cdots \\ 0 & 0 & \mathbf{I} & \mathbf{D_2} & \mathbf{D_2D_3} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \quad (11)$$

Eq.(11) is simple to obtain by solving the linear system $\mathbf{M}(\mathbf{I} - \mathbf{Q}) = \mathbf{I}$ and applying principles of induction. The details of the proof are omitted due to space limitation.

Using (6) and (11), the mean number of handoffs before leaving the network $E\{N\}$ in (10) is given as,

$$E\{N\} = \mathbf{P_S}\mathbf{M}\mathbf{e}^T - 1 = \mathbf{P_I}\left(\mathbf{I} + \sum_{k=0}^{K}\prod_{j=0}^{k}\mathbf{D_j}\right)\mathbf{e}^T - 1 \quad (12)$$

Let us define $(\mathbf{Q_M})^0 = \mathbf{I}$, then utilizing (9) we get,

$$\begin{aligned} E\{N\} &= \mathbf{P_I}\mathbf{Q_{MI}}\left(\sum_{k=0}^{K}(\mathbf{Q_M})^k\prod_{j=0}^{k}P_H(j+1)\right)\mathbf{e}^T \\ &= \mathbf{P_I}\mathbf{Q_{MI}}\left(\sum_{k=0}^{K}G(k+1)(\mathbf{Q_M})^k\right)\mathbf{e}^T \end{aligned} \quad (13)$$

where we used the observation that $P_H(k) = G(k)/G(k-1)$ and $G(0) = 1$, (i.e., $\prod_{j=0}^{k}P_H(j+1) = \prod_{j=0}^{k}\frac{G(j+1)}{G(j)} = G(k+1)$). Using the complex integral representation (2) for $G(k)$ and letting the limit $K \to \infty$, we get,

$$\begin{aligned} E\{N\} &= \mathbf{P_I}\mathbf{Q_{MI}} \\ &\cdot\left(\sum_{k=0}^{K}\frac{\mathbf{Q_M}^k}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_{R_1}^*(s)f_S^*(-s)}{s}(f_R^*(s))^k ds\right)\mathbf{e}^T \\ &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_{R_1}^*(s)\mathbf{P_I}\mathbf{Q_{MI}}f_S^*(-s)}{s}\sum_{k=0}^{K}(f_R^*(s)\mathbf{Q_M})^k\mathbf{e}^T ds \\ &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_{R_1}^*(s)\mathbf{P_I}\mathbf{Q_{MI}}f_S^*(-s)}{s}\mathbf{M_R}(s)\mathbf{e}^T ds \end{aligned}$$

where we have defined $\mathbf{M_R}(s) = (\mathbf{I} - f_R^*(s)\mathbf{Q_M})^{-1}$. Using the residue theorem we obtain the final closed form solution as,

$$E\{N\} = -\sum_{s_P \in \Xi_{S^-}}\operatorname{Res}_{s=s_p}\frac{f_{R_1}^*(s)\mathbf{P_I}\mathbf{Q_{MI}}f_S^*(-s)}{s}\mathbf{M_R}(s)\mathbf{e}^T \quad (14)$$

Comparing the mean number of handoffs in (4) and our result in (14), we observe that the scalar residence time terms $f_{R_1}^*(s)$ and $f_R^*(s)$ reflecting the speed of the users and the AGW size are now multiplied by a spatial component which

corresponds to the mobility model. $f^*_{R_1}(s)$ is multiplied by the initial movement matrix and the initial probabilities (i.e., $\mathbf{P_I Q_{MI}}$) and $f^*_R(s)$ is multiplied by the movement matrix $\mathbf{Q_M}$. Such elegant result allows us to say that the consideration of the spatial aspects due to mobility patterns and AGW arrangements simply transforms the known solution for the mean number of handoffs from the scalar form in (4) to the vector representation in (14).

Finally, using (11) and assuming that $\mathbf{A_{K+1}} = \mathbf{0}$, the roaming probability in (10) is given as,

$$\beta_V = \mathbf{P_S M A} = \mathbf{P_I A_0} + \mathbf{P_I}\left(\sum_{k=1}^{K}\prod_{j=0}^{k-1}\mathbf{D_j A_k}\right) \tag{15}$$

$$= G(1)\,\mathbf{P_I A_{MI}} + \mathbf{P_I Q_{MI}}\left(\sum_{k=1}^{K} G(k+1)\,(\mathbf{Q_M})^{k-1}\right)\mathbf{A_M}$$

Using (2), the second term in (15) denoted as $\hat{\beta}$ is given as,

$$\hat{\beta} = \frac{\mathbf{P_I Q_{MI}}}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\left[\frac{f^*_{R_1}(s)\,f^*_R(s)\,f^*_S(-s)}{s}\right.$$
$$\left.\cdot\sum_{k=1}^{K}\left(f^*_R(s)\,\mathbf{Q_M}\right)^{k-1}\mathbf{A_M}\right]ds$$

Taking the limit $K\to\infty$ and applying the Residue theorem as shown in (3), we get a closed form expression as,

$$\beta_V = G(1)\,\mathbf{P_I A_{MI}} \tag{16}$$
$$-\sum_{s_P\in\Xi_{S-}}\operatorname*{Res}_{s=s_p}\frac{f^*_{R_1}(s)\,\mathbf{P_I Q_{MI}} f^*_S(-s)}{s}\mathbf{M_R}(s)\,f^*_R(s)\,\mathbf{A_M}$$

Example As an application we consider the case, where the residence time is generally distributed and the session time follows a hyper-Erlang distribution with Laplace transform,

$$f^*_S(s) = \sum_{j=1}^{J}\alpha_j\left(\frac{\mu_j}{s+\mu_j}\right)^{m_j}$$

The poles of $f^*_S(-s)$ are located at $s_{Pj} = \mu_j > 0$ and hence satisfies the Residue theorem in (3). Thus using (14) we get,

$$E\{N\} = -\sum_{j=1}^{J}\frac{\alpha_j(-\mu_j)^{m_j}\,\mathbf{P_I Q_{MI}}}{(m_j-1)!\,E\{R\}}\lim_{s\to\mu_j}\frac{d^{m_j-1}}{ds^{m_j-1}}\mathbf{V_N}(s)\,\mathbf{e}^T$$

where we have set,

$$\mathbf{V_N}(s) = \frac{1-f^*_R(s)}{s^2}\mathbf{M_R}(s) = \frac{1-f^*_R(s)}{s^2}\left(\mathbf{I}-f^*_R(s)\,\mathbf{Q_M}\right)^{-1}$$

For the practically interesting case of $m_j = 2$, where first and second moment matching is straightforward to calculate, the required derivative $d\mathbf{V_N}(s)/ds$ can be easily obtained, because in this case we have $d\mathbf{M_R}(s)/ds|_{\mu_j} = df^*_R(s)/ds|_{\mu_j}\,\mathbf{M_R}(\mu_j)\,\mathbf{Q_M}\mathbf{M_R}(\mu_j)$ [11].

Finally, let us discuss how to alter the analysis for off-net traffic. Since off-net sessions start outside the network and enter the network at an arbitrary time instant, the session duration S is now given by its residual, $\widetilde{S}$, (i.e. $f^*_{\widetilde{S}}(s) = \frac{1-f^*_S(s)}{sE_S}$). In addition, the residence time of the first AGW, R_1, is now equal to the full residence time R (i.e. $f^*_{R_1}(s) = f^*_R(s)$) because a roaming session can only enter a boundary AGW [8]. Thus, the initialization matrix $\mathbf{Q_{MI}}$ has only rows corresponding to entry borders of AGWs lying at the network edges (e.g. in Fig.1b rows $1a$ and $1d$ for AGW 1). Accordingly $\mathbf{P_I}$ now contains the initial roaming probabilities at the exterior gateway borders.

III. Numerical Results

We now show the main results which can be obtained from our model. Let us take an example of the mean signaling rate towards any of the systems in Fig. 1, such as Policy Control Function (PCF) server. The mean signaling rate at PCF due to the handoffs can be written as $\phi = \lambda E(N_H)$, where λ is the session arrival rate. In this case, we only need to consider the mean number of handoffs for a session, in particular as a function of the mobility ratio, defined as $\rho = \frac{E_S}{E_R}$ and the user mobility pattern.

The session duration has a mean of $E_S = 40$ min and is modeled by an Erlang distribution with a coefficient of variation of $c_S = \sqrt{0.5}$ or, alternatively, by an Hyper-exponential distribution with $c_S = 3$. Since the results mainly depend on the mobility ratio and not on the absolute value of E_S, we do not consider other mean durations. For the gateway residence time a Gamma distribution is assumed, as it is was shown in [3] that it can realistically approximate measured field data. The gateway residence time has its mean value defined by the mobility ratio and its coefficient of variation is set to $c_R = 3$. The mobility behavior is considered in two different topologies, linear and rectangular. For the linear topology, defined as $M_g = 1\times N_g = 9$ we consider two different mobility patterns, referred to as "Random Route" and "Directed Route". The mobility pattern 'Random Route' allows to change the direction at each AGW. Thus the matrix $\mathbf{Q_M}$ has nonzero entries $p_{1bb} = p_{9dd} = 0.5, p_{jbb} = p_{jbd} = p_{jdb} = p_{jdd} = 0.5, j = 2,..(n-1)$ and $\mathbf{A_M}$ is given by $[0, 0.5, 0, 0, ...0.5]^T$. In the mobility pattern 'Directed Route', on the other hand, the user cannot change the initial direction, thus $\mathbf{Q_M}$ has elements $p_{jbd} = p_{jdb} = 1, j = 2,..(n-1)$ and $\mathbf{A_M}$ is given by $[0, 1, 0, 0, ...1]^T$. For the square topology, defined as $M_g = 3\times N_g = 3$ based on the layout shown in Fig.1b, we consider the so-called "Random 3×3" and "Hotspot 3×3" mobility pattern. In "Random 3×3" mobility, all transition probabilities are set to 0.25. The initial probability $\mathbf{P_I}$ for a new session is equally distributed. The mobility model "Hotspot 3×3" is defined by a combined random and directed movement pattern, as shown in the top part of Fig. 3, where a handoff session can choose a random direction only in AGWs 6 and 7. A new session starts with probability of 0.8 in AGW 5 and with equal likelihoods in the remaining gateways.

In Fig. 3, we show the effect of different topologies and movement patterns. As expected, the mean number of handoffs $E[N_H]$, is always smaller than the mean number of handoffs for the whole session, i.e. for a network of unlimited size. The latter increases with increasing the mobility ratio as

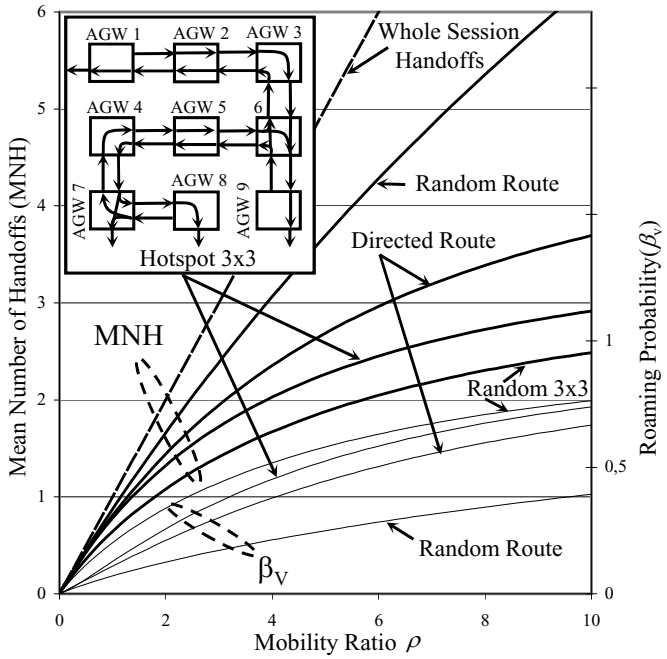

Fig. 3. Mean number of handoffs (MNH) and roaming probability β_V vs. mobility ratio for different mobility patterns (E_S=40 min, c_S = 0.5, c_R = 2).

was also shown in [9]. First, we observe that the 'Random Route' mobility pattern results in a much larger number of handoffs than the 'Directed Route'. Compared to the directed movement, the random movement incurs several direction changes inside the network and since the user can only leave the network at AGW 1 or 9, a higher mean number of handoffs is observed due to the low roaming probability (see Fig. 3). However, this behavior highly depends on the topology and the mobility patterns. This is evident by comparing the 'Random 3×3' with the 'Hotspot 3×3' mobility patterns. In this case, the random mobility behavior results in a much higher probability of roaming, β_V and thus the more directed Hotspot pattern achieves a higher mean number of handoffs. Fig. 3 clearly shows that only by joint consideration of the network topology and the mobility pattern can the mean number of handoffs for a session be accurately estimated.

In Fig. 4, we investigate the effects of different session time distributions. For all mobility patterns, a higher coefficient of variation c_S drastically reduces the mean number of handoffs. For the 'Random Route' mobility pattern, the the mean number of handoffs for an exponentially distributed session duration, widely used in literature, is also shown. Fig. 4 strongly suggests that the distribution of the session time has a great impact on the mean number of handoffs and deserves a separate and a more in-depth study.

IV. Conclusions

In this paper, we obtained a new result for the mean number of handoffs under generic assumptions of two-dimensional Markovian mobility patterns, spatial arrangement of the access gateways, as well as generic session times and gateway residence times. Our solution unveiled a new insight into mobility foundations as it shows that the consideration of the mobility pattern and the access gateways layouts simply transforms the known equation for the mean number of handoffs of an infinite network size from scalar to vector representation. We demonstrated that the mean number of handoffs is a non-linear function of the gateway spatial arrangement and user mobility and that its assessment strongly depends on the session time distribution.

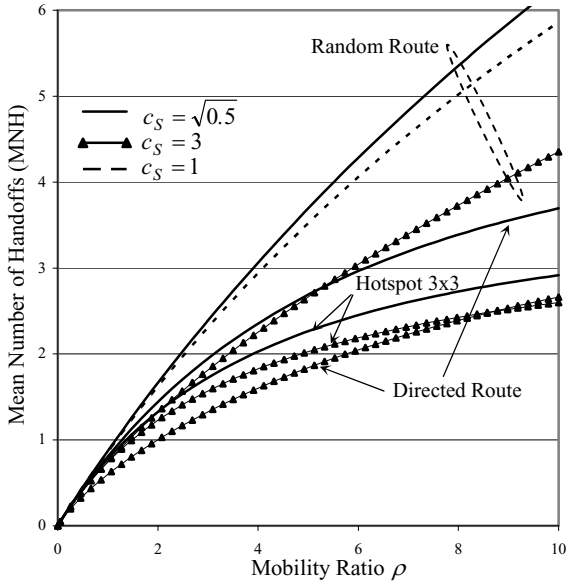

Fig. 4. Mean number of handoffs (MNH) vs. mobility ratio for different mobility patterns and session time distributions (E_S = 40 min, c_R = 2).

References

[1] 3GPP2 X.0013-12-0, 'All-IP Core Network Multimedia Domain - Service Based Bearer Control Stage 2,' Draft V. 0.21.0, July, 2006

[2] G. Camarillo, M. Garcia-Martin, The 3G IP Multimedia Subsystem (IMS), John Wiley and Sons, 2004.

[3] Y. Fang and I. Chlamtac, 'Call performance of a PCS network,' IEEE J. Selec. Areas Commun, vol. 15, no. 8, pp. 1568-1581, 1997.

[4] Y. Fang and I. Chlamtac, 'Analytical Generalized Results for Handoff Probabilities in Wireless Networks,' IEEE Trans. Communications, vol. 50, no. 3, pp. 396-399, 2002.

[5] Rodriguez-Dagnino et al, 'Counting Handovers in a Cellular Mobile Communication Network: Equilibrium Renewal Process Approach,' Performance Evaluation, vol. 52, no. 2, 2003, pp. 153-174.

[6] R. Perera, ete. al, 'Pixel oriented mobility modeling for UMTS network simulations,' in Proc. of IST Mobile and Wireless Telecommunications Summit 2002, Thessaloniki, Greece, pp 828-831, 2002.

[7] S. Zaghloul, W. Bziuk, A. Jukan, 'Relating the Number of Access Gateway Handoffs to Mobility Management: A Fundamental Approach,' submitted to IEEE INFOCOM 2009 Conference

[8] S. Zaghloul, W. Bziuk, A. Jukan, 'Signaling and Handoff Rates at the Policy Control Function (PCF) in IP Multimedia Subsystem (IMS),' IEEE Commun. Letters Journal, vol. 12, no. 7, pp. 526-528, 2008.

[9] W. Bziuk, S. Zaghloul, A. Jukan, 'A New Framework for Characterizingthe Number of Handoffs in Cellular Networks,' accepted for publication in PGTS 08, Berlin, Germany, 2008.

[10] U. Bhat, G. Miller, Elements of Applied Stochastic Processes, Wiley,2002, 3rd ed.

[11] W. Fischer, K. Meier-Hellstern, 'The Markov-modulated Poissonprocess cookbook,' Performance Evaluation, vol. 18, 1992.

[12] Y. Fang, 'Modeling and Performance Analysis for Wireless MobileNetworks: A New Analytical Approach,' IEEE Trans. Networking,vol. 13, no.5, pp. 989-1002, 2005.