A Novel Analytical Framework for Mobility Modeling in All-IP Wireless Systems

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

***Abstract*— In all-IP wireless networks, groups of cells are served by IP access gateways. When users move between regions served by different gateways, IP and application layer mobility signaling is triggered. The key performance indicator to assess a given mobility protocol is the mean number of gateway handoffs. In this paper, we propose a new fundamental approach, which majorly extends past work in cellular communications, to evaluate the number of access gateway handoffs based on their size and the mobility pattern between cells during the session lifetime. Using transient Markov chain and complex analytic techniques, we obtain the mean number of handoffs for a session, be it entirely or only partially served by the network under consideration, for directed and random mobility. The main set of results indicates that the number of handoffs in the network is a non-linearly increasing function of the session duration and the number of access gateways, for both mobility patterns. To demonstrate the practical relevance of the approach we show results for MobileIP airlink load as a function of handoff rate.**

I. Introduction

With the increasing demand for new services and wireless networks offering high data rates and uninterrupted realtime content streams, mobility poses significant design challenges to the network design and to the performance of IP signaling protocols. Mobility is particularly a challenge due to the expected increase in the number of handoffs between regions served by different IP access gateways (AGW) which serve multiple cells. The increase in the handoff rate is due to the introduction of new network architectures, new business models for roaming, and new types of services with longer session durations. In this regard, new IP based architectures are expected to become flatter and composed of smaller radio network components supporting a lower number of base stations with smaller coverage areas [1][2]. Furthermore, roaming traffic between networks is expected to increase due to the success of mobile virtual network operator models, the incorporation of small sized operators (e.g., rural alliances), and the envisioned inter-operability between heterogeneous networks such as between WiMAX and LTE. Clearly, ignoring mobility in the design of future networks can potentially lead to excessive handoffs causing significant growth in the signaling load in the airlink and within the network, and hence degrading the QoS of sessions due to the increased likelihoods of packet losses and session dropping. Therefore, it is pivotal to revisit the past cellular handoff models and extend them to cope with designs based on AGW serving areas composed of multiple cells and potential chances for roaming between different wireless network operators.

Past research in cellular telephony addressed various challenging issues in micro and macro cellular environments [3][4][5][6] and between cellular and WiFi systems [7], most notably call dropping likelihoods and channel utilization. Common to most of the past established work is that the mean number of handoffs (MNH) was estimated based on cellular residence time distributions which are directly obtained from measurements or by fitting simulated mobility traces. However, to evaluate the number of handoffs during a session in next generation networks, the AGW residence time distribution needs to be first calculated from the measured cellular residence times. This necessitates the consideration of the mobility patterns between cells which was not used in traditional cellular environments. Although other work relevant to designing paging areas introduced mobility effects, e.g., the Markovian analytic model and the Gauss-Markov simulation model [8][9][10], only the mean residence times in the paging areas were usually needed. Further complexities arise in the analysis of handoffs in next generation systems due to the increased likelihood to roam between networks due to new roaming models and longer session durations. As such this results in partially served sessions by the network and hence affecting the estimate for the observed MNH.

In this paper, we develop a fundamental approach to evaluate the mean number of handoffs between IP access gateways in next generation systems for a given number of cells, session duration, cellular residence time distribution, and mobility pattern. To incorporate roaming, we evaluate the MNH for:

(i) The entire session duration, irrespective of location, and
(ii) Partially served sessions, including new sessions in the network until they leave the network and sessions arriving from other networks until they leave the network,

Using transient Markov chain and complex analytic techniques, we develop a hierarchical analytical model which entails two significant contributions, (a) the derivation of the access gateway residence time from the (measured) cellular residence times and a given mobility pattern, (b) the derivation of the MNH during IP sessions as function of the derived residence times, the mobility pattern, and the number of AGWs in the network. Our approach majorly extends the seminal results from [4][9] and yields the handoff rate by considering the number of cells per AGW and mobility patterns among them for any session be it entirely or only partially served by the network. Our main set of results shows that the MNH in the network is a non-linearly increasing function of the session

duration and the number of access gateways for different mobility patterns. To demonstrate the practical relevance of the approach we show results on the effect of the number of cells per AGW on the residence time and show an exemplary metric of the *base station airlink load* as a function of the MNH. The rest of the paper is organized as follows. Section II provides the necessary background. Sections III and IV present our analytical model. In Section V, we discuss the results. In Section VI, we conclude the paper.

II. Background

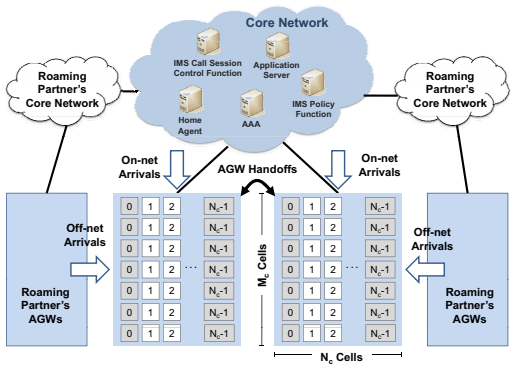

Fig. 1. Access Gateway Areas in the Network, $N_g = 2$

Fig.1 illustrates an operator's network (i.e., network under consideration, or simply, network) which consists of two access gateways (AGWs) each serving a square gateway area of $M_c \times N_c$ cells. Exemplary AGW elements include ASN-GW in WiMAX networks, PDSN in 3GPP2 networks, and GGSN in 3GPP networks. Every handoff between a pair of AGWs contributes to the mobility management signaling in higher layers. Note that for clarity and due to our scope, we will use the term "handoff" to refer to access gateway handoffs only, unless explicitly stated otherwise, e.g., *cellular* handoff. Upon every handoff (i.e., access gateway handoff) the higher layer signaling is triggered towards the core network and affects the design and architecture of multiple components, including the AAA system, the home agent supporting MobileIP, the IMS policy function which authorizes and enforces certain QoS levels, etc. Relevant to our model, we refer to sessions initiating from within the network under consideration as *on-net* sessions while we refer to sessions that have been already initiated in other operators' networks as *off-net* sessions. Off-net sessions are first observed and given a service by the border cells (shaded cells in Fig.1, e.g., AGW_0 and AGW_{N_g-1}), while on-net sessions start from within the network. Common to both session types is the feature that they may only be partially served by the network. The longest time an on-net session spends in the network is the whole session duration or else this session leaves due to mobility. On the other hand, the off-net traffic can only spend up to the remaining session time (session *residual*) in the network.

As previously noted, the number of handoffs affects the mobility management protocols, and thus the QoS mobile users observe. For instance, if no packet bicasting or buffering is enabled, every gateway handoff results in loss of IP packets and hence QoS degradation. Also the airlink load due to signaling in the border cells is proportional to the number of handoffs due to mobility signaling over the air. For some users, typically subscribers or home users, the whole signaling pertaining to the session is always received by the home network irrespective of the user location, either directly from the access gateways or indirectly through proxy systems (e.g., in case of AAA). From the modeling point of view, this is equivalent to the infinite network size where no matter how the mobile node moves, it never leaves the network under consideration; we referred to this as case (i) in the Introduction. Otherwise, it can happen that sessions are partially served by the network under consideration, i.e., proportionally to the time they spend in the network, where only the number of handoffs incurred while being served is important (case *ii* in the Introduction). Whether the sessions will leave the network before they terminate depends not only on session statistics but also on mobility. To this end, we analyze two extreme mobility patterns, depending on how fast a session can leave the network: directed mobility (similar to fluid flow) and random mobility. In the directed mobility case, users pick a random direction (i.e., east or west) and keep it until their session terminates or they leave the network. With random mobility, users move randomly between cells resulting in a Markovian mobility behavior between gateways as we will show in Section IV.

III. The Gateway Residence Time

In this section, we use the cellular residence time distribution (e.g., from measurements) to derive the gateway residence time statistics depending on the mobility model and network size. Based on this result, we obtain the MNH as will be shown in Section IV. Let us now start by explaining the directed mobility pattern which clearly illustrates our analysis and then proceed to a more general analysis for random mobility.

A. Assumptions

- The session arrival process for on-net and off-net traffic is Poissonian with mean rates of λ_Ω and λ_Φ, respectively.
- The session duration, S, has a rational Laplace transform and a mean of E_s.
- The cellular residence times, R_c, are independent and identically distributed. R_c is generally distributed with an existing Laplace transform and mean of E_{R_c}.
- The gateways are arranged linearly with roaming partners located east and west. There are $M_c \times N_c$ cells per gateway arranged rectangularly (Fig.1). The generalization for arbitrary movements and AGW arrangements can be performed based on our work in [16] using pixel based mobility models [14] which were used to fit realistic movement traces from cellular networks.

B. Directed Mobility

In this model, users move in a single direction, either east or west, with equal likelihoods and throughout their whole

session duration. As such users move in the shortest possible path towards the boundaries of the AGWs. For the directed user movement the gateway residence time is composed of the sequence of cellular residence times R_c, which are assumed to be iid. When the session hands off from a neighboring gateway, the mobile node may cross all N_C cells until it leaves the gateway, which yields the residence time R_g. On the other hand, when the session starts inside the gateway, then the gateway residence times depends on the starting cell. The average overall possibilities give us the residence time R_{g1}.

1) Residence time for sessions starting inside the gateway: Let us first consider the case that the session starts in a cell that belongs to column $j-1, j \geq 1$ of a chosen AGW. Then if we first assume that the user moves to west he leaves the AGW after the j'th cellular handoff. For this case, the sequence of cellular residence time is given by the residual of the cellular residence time $\tilde{R}_c$ incurred in the first cell and $j-1$ subsequent cellular residence times. After j handoffs, the gateway residence time in the first AGW is then,

$$R_{g_1}(j)=\tilde{R}_c+\sum_{k=2}^{j} R_c,\ f^*_{R_{g_1}}(s,j)=f^*_{\tilde{R}_c}(s)\left(f^*_{R_c}(s)\right)^{j-1} \quad (1)$$

Assuming uniformly distributed session arrivals per cell and taking into account the symmetry for going west or east, the Laplace transform of R_{g_1} is

$$\begin{aligned} f^*_{R_{g_1}}(s) &= \frac{1}{N_c}\sum_{j=1}^{N_c} f^*_{\tilde{R}_c}(s)\left(f^*_{R_c}(s)\right)^{j-1} \\ &= \frac{1-f^*_{R_c}(s)}{sN_cE\{R_c\}}\frac{1-\left(f^*_{R_c}(s)\right)^{N_c}}{1-f^*_{R_c}(s)} = \frac{1-\left(f^*_{R_c}(s)\right)^{N_c}}{sN_cE\{R_c\}} \end{aligned} \quad (2)$$

2) The gateway residence time: The gateway residence time R_g for a session that handoffs from a neighbor gateway can only start in cell 0 and leave in cell N_C-1 or vice-versa. Due to the symmetry it follows

$$R_g=\sum_{k=1}^{N_c} R_c, \quad f^*_{R_g}(s)=\left(f^*_{R_c}(s)\right)^{N_c} \quad (3)$$

Interestingly, we can observe here that eq.2 is equivalent to the residual of the AGW residence time (i.e., $R_{g_1}=\tilde{R}_g$).

C. Random Mobility

In this section we assume that users move randomly between cells. To characterize the users' movements between cells, we use transient Markov chains. The transient states are used to model the cells inside the access gateway while the absorbing states represent departures from an AGW coverage area. Before we proceed, let us first summarize known results on transient Markov chains[13]. Let the transition probabilities matrix $\mathbf{P}$ be given as,

$$\mathbf{P}=\begin{bmatrix} \mathbf{Q} & \mathbf{A} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad (4)$$

where states are reordered such that $\mathbf{Q}$ contains only transitions between transient states, $\mathbf{A}$ contains only transitions to absorbing states V_i, and $\mathbf{I}$ is the identity matrix with proper dimensions. Let the row vector $\mathbf{f}$ denote the initial starting probabilities for the transient Markov chain and $\mathbf{A_i}$ be the ith column of $\mathbf{A}$. Then the joint probability of being absorbed into state V_i after the j^{th} transition is given as

$$P\{N_{(i)}=j\}=\mathbf{f}\ \mathbf{Q}^{j-1}\ \mathbf{A_i}, j=1,2,... \quad (5)$$

Using the fundamental matrix $\mathbf{M}$ defined as $\mathbf{M}=[\mathbf{I}-\mathbf{Q}]^{-1}$[13], the probability of being absorbed into state V_i is

$$\beta_i=\mathbf{f}\ \mathbf{M}\ \mathbf{A_i} \quad (6)$$

The mean number of all visits to transient states, including the first and before absorption is,

$$E\{N\} \quad = \quad \mathbf{f}\ \mathbf{M_p}\ ,\ \mathbf{M_p}=\sum_{j=1}^{k}\|\mathbf{M}\|_{i,j} \quad (7)$$

In our context, as shown in Fig.1, we view the access gateway as a collection of columnar groups of cells. After each cellular handover, the user may move to another cell within the same column i, leave the current column i and go east to column $i+1$, or go west to column $i-1$. Note that for simplicity of explanation we only consider mobility east-west between AGWs; the north-south mobility can be incorporated similar to [8]. From the geometry, the probabilities of going east and west are equal and are $\alpha=0.25$, while the probability of staying in the same column is $\zeta=0.5$. We will model the user movement inside the AGW using a transient Markov chain. The access gateway area consists of N_c zones which represent the transient states. Only through two departure areas (i.e., shaded zones 0 and N_c-1) the user can leave the AGW. This is modeled by two absorbing states, G_E and G_W, representing departure to the east or west. The transition probabilities between the transient states are characterized by the $N_c \times N_c$ matrix $\mathbf{Q_g}$ using the zone departure and stay probabilities α and ζ as,

$$\mathbf{Q_g}=\begin{bmatrix} \zeta & \alpha & \dots & & \\ \alpha & \zeta & \alpha & 0 & \dots \\ & & \vdots & & \\ & & & \alpha & \zeta \end{bmatrix} \quad (8)$$

The transitions to the absorbing states (i.e., departure east or west) are given by the $N_c \times 1$ column vectors $\mathbf{A_{gE}}$ and $\mathbf{A_{gW}}$

$$\begin{aligned} \mathbf{A_{gW}} &= [\ \alpha \quad 0 \quad \dots \quad 0\ ]^T \\ \mathbf{A_{gE}} &= [\ 0 \quad 0 \quad \dots \quad \alpha\ ]^T \end{aligned} \quad (9)$$

1) Residence time for sessions starting inside the gateway: Following our previous notation, when the session starts inside the AGW the residence time is R_{g1}. The new sessions can start with equal probability within any cell of the AGW. The initial state probabilities for the transient Markov chain are given as

$$\mathbf{f_{gI}}=[1\dots1]/N_c \quad (10)$$

Now the number of cellular handoffs until departure can be seen as equal to the number of transitions between transient

states until absorbtion. Thus with (5) we can define the joint probability that the departure occurs with the j'th cellular handoffs to the east (or west) for new sessions as,

$$P\{N_{G(I)}=j\}=P\{N_{G(I,y)}=j\}=\mathbf{f_{gI}}\mathbf{Q_g}^{j-1}\mathbf{A_{gy}}, \quad (11)$$

where (I,y) specifies the starting state I and the departure side $y\in\{E,W\}$. Due to the symmetry we get the same distribution for both departure sides. Because $\mathbf{A_{gy}}$ contains only one absorbing state, eq.11 defines a discrete phase-type distribution, where j gives the number of time steps until absorption. Due to symmetry the probability of being absorbed into state G_W or G_E is given by $\beta_I=\beta_{IW}=\beta_{IE}=0.5$. Finally to derive the residence time we need the probability that the departure occurs with the j'th cellular handoff conditioned on the departure to the east or west, which simply follows from (11) to

$$P\{N_G=j|I\}=P\{N_{G(I)}=j\}/\beta_I \quad (12)$$

Each step in the transient Markov chain represent a cellular residence time, where the first step has a duration of the residual cellular residence time due to the session start, followed by $j-1$ steps with duration of the cellular residence time. Similar to (2) the Laplace transform of the residence time is,

$$\begin{aligned} f^*_{R_{g1}}(s)&=\sum_{j=1}^{\infty}f^*_{\tilde{R}_c}(s)\left(f^*_{R_c}(s)\right)^{j-1}P\{N_G=j|I\}\\ &=f^*_{\tilde{R}_c}(s)\mathbf{f_{gI}}\sum_{j=1}^{\infty}\left(f^*_{R_c}(s)\mathbf{Q_g}\right)^{j-1}\mathbf{A_{gW}}/\beta_I\\ &=f^*_{\tilde{R}_c}(s)\mathbf{f_{gI}}\mathbf{M_g(s)}\mathbf{A_{gW}}/\beta_I \end{aligned} \quad (13)$$

where we have defined

$$\mathbf{M_g(s)}=[\mathbf{I}-f^*_{R_c}(s)\mathbf{Q_g}]^{-1} \quad (14)$$

The mean residence time can be derived from (13) and (14) and is given as $E\{R_{g1}\}=df^*_{R_{g1}}(s)/ds|_{s=0}=E\{\tilde{R}_c\}+E\{R_c\}(E\{N_{G(I)}\}-1)$. From (7) we know that $E\{N_{G(I)}\}$ is the mean number of visits before absorption. The first visit has to be excluded, because it is not associated with a transition, i.e. a cellular handoff. The probability to leave is 1, thus $1+(E\{N_{G(I)}-1\})$ is equal to the mean number of cellular handoffs until leaving the AGW. Thus the mean residence time is composed of the residual cellular residence time $E\{\tilde{R}_c\}$ for the first cell and $(E\{N_{G(I)}-1\})$ cell residence times $E\{R_c\}$.

2) The gateway residence time ["Short" and "Long"]: Let us now analyze the details the gateway residence times for a handoff session entering the gateway region at any cell in edge columns (0 or N_c-1) of the gateway AGW_k, $k=0,...,N_g-1$. These sessions are particularly important because a handoff session staring in cell 0 will either handoff to the west AGW_{k-1} and thus present very short residence times R_{ga} within the gateway AGW_k ("short residence"), or on another extreme, the session may cross the whole gateway area, by moving east and experience very long residence times R_{gb}. The corresponding initial state probabilities are given as

$$\begin{aligned} \mathbf{f_{gW}}&=[\,1\quad 0\quad \ldots\quad 0\quad 0\,]\\ \mathbf{f_{gE}}&=[\,0\quad 0\quad \ldots\quad 0\quad 1\,] \end{aligned}$$

Let us also denote the joint probability that the departure occurs with the j'th cellular handoff to the east (or west) given their initial starting edge east (west) as $P\{N_{G(E,E)}=j\}=P\{N_{G(W,W)}=j\}=P\{N_{G(a)}=j\}$ and $P\{N_{G(E,W)}=j\}=P\{N_{G(W,E)}=j\}=P\{N_{G(b)}=j\}$ due to symmetry. Thus, the distributions of the number of handoffs are,

$$\begin{aligned} P\{N_{G(a)}=j\}&=\mathbf{f_{gW}}\mathbf{Q_g}^{j-1}\mathbf{A_{gW}}\\ P\{N_{G(b)}=j\}&=\mathbf{f_{gW}}\mathbf{Q_g}^{j-1}\mathbf{A_{gE}} \end{aligned} \quad (15)$$

Finally, the departure probabilities for a session starting at an edge zone and leaving at the same edge towards the neighboring gateway, β_{ga}, and the opposite case, β_{gb}, are

$$\beta_{ga}=\mathbf{f_{gE}}\mathbf{M_g}\mathbf{A_{gE}}=\frac{N_c}{N_c+1}\ ,\ \beta_{gb}=1-\beta_{ga} \quad (16)$$

where $\mathbf{M_g}=[\mathbf{I}-\mathbf{Q_g}]^{-1}$. The result for β_{ga} follows from the the last element of $\mathbf{M_g}$ derived by simple backward substituation. Similar to eq.12 we have to condition the probability distributions (15) by the departure probabilities, i.e.,

$$P\{N_G=j|x\}=P\{N_{G(x)}=j\}/\beta_{gx}, x\in\{a,b\} \quad (17)$$

For j cellular handoffs the access gateway residence time is given by $R_g(j)=\sum_{k=1}^{j}R_c$. Using (14), we get the Laplace transform for "short" and "long" residence times as

$$\begin{aligned} f^*_{R_{ga}}(s)&=\sum_{j=1}^{\infty}\left(f^*_{R_c}(s)\right)^j P\{N_G=j|a\}\\ &=f^*_{R_c}(s)\mathbf{f_{gW}}\mathbf{M_g(s)}\mathbf{A_{gW}}/\beta_{ga}\\ f^*_{R_{gb}}(s)&=f^*_{R_c}(s)\mathbf{f_{gW}}\mathbf{M_g(s)}\mathbf{A_{gE}}/\beta_{gb} \end{aligned} \quad (18)$$

IV. The Mean Number of Handoffs (MNH)

In this section, we derive the mean number of handoffs (MNH) between AGWs for the whole sessions and for sessions only partially served by the network. We consider directed and random mobility using the gateway residence times developed in the previous section. Before we start, let us shortly discuss the probability of making k handoffs for on-net and off-net sessions. Let us first define the effective session duration, S_E, as the time a session spends in the network under consideration. For on-net sessions, i.e., new sessions starting inside the network, the effective session duration is then equal to the session duration, $S_{E\Omega}=S$. On the other hand, an off-net session, Φ, starts in other operator networks before entering into the network and hence we only observe the residual session duration (i.e., $S_{E\Phi}=\tilde{S}$). The effective session duration has Laplace transform of $f^*_{S_{Ex}}(s)$ where $x\in\{\Omega,\Phi\}$. The corresponding residence times in the first AGW is R_{g1} for on-net and R_g for off-net sessions (see Section III). All subsequent handoffs occur after a residence time duration

R_g elapses. Let $R_{g\Omega}(k) = R_{g1} + \sum_{j=1}^{k-1} R_g$ and $R_{g\Phi}(k) = \sum_{j=1}^{k} R_g$ denote the sum of residence times since the session starting event until the k^{th} handoff event for on-net and off-net sessions respectively. The Laplace transforms for $R_{g\Omega}(k)$ and $R_{g\Phi}(k)$ are given as,

$$f^*_{R_{g\Omega}(k)}(s) = f^*_{R_{g1}}(s)\Big(f^*_{R_g}(s)\Big)^{k-1},\ f^*_{R_{g\Phi}(k)}(s)=\Big(f^*_{R_g}(s)\Big)^{k} \quad (19)$$

Denoting $x \in \{\Omega, \Phi\}$, it can be shown that the probability of making k handoffs for is given as [15], [16],

$$P\left\{N_{H_x} = k\right\} = G_x\left(k\right) - G_x\left(k+1\right), \quad k \geq 1 \quad (20)$$

where using (19), $G_x\left(k\right)$ is given as,

$$G_x\left(k\right) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f^*_{R_{gx}(k)}\left(s\right) f^*_{S_{Ex}}\left(-s\right)}{s} ds$$

Using (20), it can also be shown that the MNH for the whole session (i.e., the complete session duration) is given as [15],

$$E\left\{N_H\right\} = E\left\{N_{H_\Omega}\right\} = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f^*_{R_{g1}}(s)\, f^*_S(-s)}{s\left(1 - f^*_{R_g}\left(s\right)\right)} ds \quad (21)$$

When R_{g1} is equal to the residual of the gateway residence time $\tilde{R}_g$, then the MNH is simply given as $E\left\{N_H\right\} = \frac{E\{S\}}{E\{R_g\}}$.

A. Directed Mobility

1) MNH during the whole session: In this case, the effective session duration is equal to the whole session time. Since R_{g1} in (2) is the residual of R_g in (3), the MNH is given as,

$$E\{N_{H_\Omega}\} = \frac{E\{S\}}{E\{R_g\}} = \frac{E\{S\}}{N_C E\{R_C\}} \quad (22)$$

2) MNH for sessions partially served by the network: We now assume a network of finite size, with N_g AGWs and consider the case of on-net traffic first. When the session starts in AGW$_j, j = 0, 1, ..., N_g - 1$, and moves east (west), then using (20) the conditioned MNH within the network is,

$$E\{N^{West}_{H\Omega}(j)\}=\sum_{k=0}^{j} kP\{N_{H_\Omega} = k\}=E\{N^{East}_{H\Omega}(N_g-j-1)\} \quad (23)$$

Starting from cell j and moving west the user leaves after $j+1$ handoffs, thus using symmetry the probability of leaving the network for on-net traffic from either side is given as,

$$\beta^{West}_\Omega(j) = (j+1)\sum_{k=j+1}^{\infty} P\{N_{H_\Omega}=k\} = \beta^{East}_\Omega(N_g-j-1) \quad (24)$$

If $f_\Omega(j)$ denotes the probability that the session starts in AGW$_j$, the mean number of handoffs until departure is given by the sum of the MNH within the network ($E\{N^{West}_{H_\Omega}(j)\} + E\{N^{East}_{H_\Omega}(j)\}$) and the last handoff given by the exit probability ($\beta^{West}_\Omega(j) + \beta^{East}_\Omega(j)$) as,

$$E\{N^{NET}_{H_\Omega}\} = \frac{1}{2}\sum_{j=0}^{N_g-1} f_\Omega(j)\Bigg[E\{N^{West}_{H_\Omega}(j)\} + \beta^{West}_\Omega(j) + E\{N^{East}_{H_\Omega}(j)\} + \beta^{East}_\Omega(j)\Bigg] \quad (25)$$

Since off-net traffic starts either from east or west and crosses the network to the opposite edge, its MNH is then,

$$E\{N^{NET}_{H_\Phi}\}= \sum_{j=0}^{N_g-1} jP\{N_{H_\Phi} = j\}+N_g \sum_{j=N_g}^{\infty} P\{N_{H_\Phi} = j\} \quad (26)$$

The second term in (26) gives the probability of departing from the network for off-net sessions.

B. Random Mobility

1) MNH during the whole session: To derive the MNH for the whole session, we use (19)-(21). To use (19), we note that after k handoffs, the sequence of the k residence times includes R_{g1}, n R_{ga}, and (k-n-1) R_{gb} where $0 \leq n < k$. Hence the Laplace transform for the sum of the residence times conditioned on n is,

$$f^*_{R_{\Omega}(k,n)}(s) = f^*_{R_{g1}}(s)\left(f^*_{R_{ga}}(s)\right)^n\left(f^*_{R_{gb}}(s)\right)^{k-n-1} \quad (27)$$

After the kth handoff, any occurrence of R_{ga} is equally likely, which means that we can weight the Laplace transform of $f_{R_{\Omega}(k,n)}$ using the binomial distribution,

$$\begin{aligned} f^*_{R_{\Omega}(k)}(s) &= \sum_{n=0}^{k-1}\binom{k-1}{n}\beta^n_{ga}(1-\beta_{ga})^{k-n-1} f^*_{R_{\Omega}(k,n)}(s) \\ &= f^*_{R_{g1}}(s)\left(\beta_{ga} f^*_{R_{ga}}(s) + (1-\beta_{ga}) f^*_{R_{gb}}(s)\right)^{k-1} \end{aligned} \quad (28)$$

Comparing (28) with (19), we observe that the mean of the "short and long" residence times for sessions is given by,

$$f^*_{R_g}(s) = \beta_{ga} f^*_{R_{ga}}(s) + (1-\beta_{ga}) f^*_{R_{gb}}(s) \quad (29)$$

Thus under the assumption given above, the MNH can be calculated for a generally distributed session time by substituting $f^*_{R_g}(s)$ and $f^*_{R_{g1}}(s)$ into (21). For an exponentially distributed session duration we simply get,

$$E\{N_H\}=\frac{f^*_{R_{g1}}(E_s^{-1})}{1-\beta_{ga} f^*_{R_{ga}}(E_s^{-1})-(1-\beta_{ga}) f^*_{R_{gb}}(E_s^{-1})} \quad (30)$$

2) MNH for the sessions partially served by the network: For a network of finite size, the binomial model used in (28) does not hold due to the possibility of network departure. Hence, for tractability, we assume an exponential session duration and characterize the users' movement among access gateways using a transient Markov chain at the gateway levels, as shown in Fig.2. In this model, transient states represent the serving gateways during the session lifetime, while absorbing states represent session termination (state T) or departure from the network area under consideration (states V_W and V_E). Note that cell related statistics do not play a role here. We represent each AGW after the first handoff by two transient states: one representing sessions entering AGW i from the eastern (i, E) and another representing sessions entering from the western border (i, W). Since on-net sessions may start at anywhere within the gateway, an extra state representation is needed (shaded states in Fig.2). Here the transition probabilities a' and b' define the initial choice of a direction. In this regard,

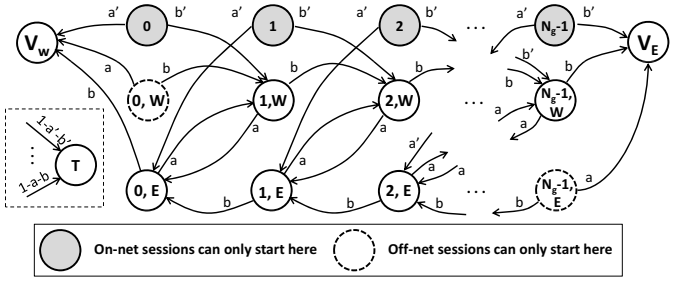

Fig. 2. Mobility model among access gateways

off-net sessions are a special case of the "short" and "long" residence traffic as they only start at the border AGWs in states (0, W) or $(N_g - 1, \text{E})$. Once the user makes the first handoff, all subsequent handoff to the next gateways are determined by the initial border (i.e., east or west) where the user entered the region. The transition probability b is related to a user who crosses the AGW ("long"), that means the user has entered from the western side and leaves through the eastern side, or vice versa. Similarly, the transition probability a is related to a user who has entered and left the AGW from through the same side ("short"). These transition probabilities represent the probabilities of making at least one more handoff and depend on session duration, the cellular residence time, and direction of movement. Finally, each transient state can reach the absorbing state T with the probabilities $1 - a - b$ and $1 - a' - b'$, respectively as derived later.

As shown in Fig.2, for on-net traffic we need two types of transient states: G_0 that is only visited once and represents the initial AGW for the session (shaded states), and G_1 that represent the states where sessions entering from the eastern and western borders. Let us order the states lexographically as 0,1,..,$N_g - 1$, $(0, W)$, $(0, E)$, $(1, W)$, $(1, E)$,..., $(N_g - 1, W)$, $(N_g - 1, E)$, then the probability transition matrix among transient states for on-net and off-net traffic, $\mathbf{Q_\Omega}$ and $\mathbf{Q_\Phi}$ are,

$$\mathbf{Q_\Omega} = \begin{bmatrix} \mathbf{0} & \mathbf{\Upsilon_0} \\ \mathbf{0} & \mathbf{\Upsilon_1} \end{bmatrix}, \mathbf{Q_\Phi} = \mathbf{\Upsilon_1} \tag{31}$$

where $\mathbf{\Upsilon_0}$ is a $N_g \times 2N_g$ matrix representing the transitions to the (i, W) and (i, E) states from the initial states belonging to group G_0. $\mathbf{\Upsilon_1}$ is a $2N_g \times 2N_g$ matrix representing the transitions among states belonging to the group G_1, and $\mathbf{0}$ is an all zeros matrix with the proper size, i.e.,

$$\mathbf{\Upsilon_0} = \begin{bmatrix} 0 & 0 & b' & 0 & \dots & & \\ 0 & a' & 0 & 0 & b' & 0 & \dots \\ 0 & 0 & 0 & a' & \dots & & \\ & & \vdots & & & & \end{bmatrix}$$

$$\mathbf{\Upsilon_1} = \begin{bmatrix} 0 & 0 & b & 0 & & 0 & \dots \\ 0 & 0 & a & 0 & & 0 & \dots \\ 0 & a & 0 & 0 & b & 0 & \dots \\ 0 & b & 0 & 0 & a & 0 & \dots \\ & & \vdots & & & & \end{bmatrix}$$

The transition probability matrices for on-net and off-net sessions towards the absorbing states V_W and V_E are,

$$\mathbf{A_\Omega} = \begin{bmatrix} \mathbf{\Psi_0^{(W)}} & \mathbf{\Psi_1^{(W)}} \\ \mathbf{\Psi_0^{(E)}} & \mathbf{\Psi_1^{(E)}} \end{bmatrix}^T, \mathbf{A_\Phi} = \begin{bmatrix} \mathbf{\Psi_1^{(W)}} \\ \mathbf{\Psi_1^{(E)}} \end{bmatrix}^T$$

where $\mathbf{A_\Omega}$ is a $3N_g \times 2$ matrix and $\mathbf{A_\Phi}$ is a $2N_g \times 2$ matrix. The $1 \times N_g$ row vectors $\mathbf{\Psi_0^{(W)}}$ and $\mathbf{\Psi_0^{(E)}}$, and the $1 \times 2N_g$ row vectors $\mathbf{\Psi_1^{(W)}}$ and $\mathbf{\Psi_1^{(E)}}$ are given as,

$$\mathbf{\Psi_0^{(W)}} = \begin{bmatrix} a' & 0 & \dots & 0 \end{bmatrix}, \mathbf{\Psi_0^{(E)}} = \begin{bmatrix} 0 & \dots & 0 & b' \end{bmatrix}$$
$$\mathbf{\Psi_1^{(W)}} = \begin{bmatrix} a & b & 0 & \dots & 0 \end{bmatrix}, \mathbf{\Psi_1^{(E)}} = \begin{bmatrix} 0 & \dots & 0 & b & a \end{bmatrix}$$

Let the initial state probabilities for on-net sessions be defined using the $1 \times 3N_g$ row vector as $\mathbf{f_\Omega} = \begin{bmatrix} \epsilon_0 & \epsilon_1 & \dots & \epsilon_{N_g-1} & 0 & \dots & 0 \end{bmatrix}$ where ϵ_k represent the probability of starting a session from AGW_k. Let the initial state probabilities for off-net sessions be defined using the $1 \times 2N_g$ row vector $\mathbf{f_\Phi} = \begin{bmatrix} \varepsilon_W & 0 \dots & 0 & \varepsilon_E \end{bmatrix}$ where ε_W and ε_E represent the probabilities of arrival from the western roaming partner V_W towards the western border of AGW_0 and from the eastern roaming partner V_E towards the eastern border of AGW_{N_g-1}.

The transition probabilities a' and b' for on-net traffic are then evaluated as the joint probability of departure from the eastern side or the western side and the event, that the session has not terminated in the initial AGW. For exponentially session durations and since both movement directions are chosen with equal probability, we have

$$\begin{aligned} a' = b' &= \frac{1}{2} P\{S > R_{g1}\} = \int_0^\infty P\{S > t\} f_{R_{g1}}(t) dt \\ &= \int_0^\infty e^{\frac{-t}{E_s}} f_{R_{g1}}(t) dt = f^*_{R_{g1}}(s) \mid_{s=E_s^{-1}} \end{aligned} \tag{32}$$

where the Laplace transform of R_{g1} in the first AGW has been derived in (13) and E_s is the mean session duration.

For subsequent handoffs from both on-net and off-net sessions, the transition probability a is given by the probability to enter and leave the AGW through the same border, given by the departure probability β_{ga} defined in (16) and the probability that the session makes at least one more handoff, $P\{S > R_{ga}\}$. For this case we have to take the related residence time R_{ga}, whose Laplace transform has been derived in (18). For the transition probability b we use β_{gb} and R_{gb} instead. Due to the memoryless property of the exponential session duration, we have,

$$a = \beta_{ga} P\{S > R_{ga}\}, b = \beta_{gb} P\{S > R_{gb}\} \tag{33}$$

The probabilities $P\{S > R_{ga}\}$ and $P\{S > R_{gb}\}$ are then,

$$\begin{aligned} P\{S > R_{ga}\} &= \int_0^\infty e^{\frac{-t}{E_s}} f_{R_{ga}}(t) dt = f^*_{R_{ga}}(s) \mid_{s=E_s^{-1}} \\ P\{S > R_{gb}\} &= f^*_{R_{gb}}(s) \mid_{s=E_s^{-1}} \end{aligned} \tag{34}$$

As can be seen from (34), the probabilities are calculated by evaluating the Laplace transform of the gateway residence times at a real value, which is simply done using eq.(18).

We are finally close to obtain the MNH using the described transient Markov chain. From eq.(6) we get the probability to leave the network, which reflects the last handoff, as

$$\beta_x = \mathbf{f_x}\ \mathbf{M_x}\ \mathbf{A_x}\ \mathbf{e},\ \ x \in \{\Omega, \Phi\}\ ,\ \mathbf{M_x} = [\mathbf{I} - \mathbf{Q_x}]^{-1} \quad (35)$$

where $\mathbf{e} = [1, 1]^T$. The MNH for on-net $E[N_{H_\Omega}^{NET}]$ and off-net $E[N_{H_\Phi}^{NET}]$ sessions are than given using (7) and (35) as,

$$E[N_{H_x}^{NET}] = \beta_x + \mathbf{f_x}\ \mathbf{M_{p_x}} - 1\ , x \in \{\Omega, \Phi\} \quad (36)$$

$$\mathbf{M_{p_x}} = \sum_{j=1}^{k} \|\mathbf{M_x}\|_{i,j}\ .$$

The -1 in (36) reflects the fact that the mean number of handoffs corresponds to the number of transient state *revisits* until departure, and since $\mathbf{f_x}\mathbf{M_{p_x}}$ includes also the first visit (the session start), it has be subtracted from the final equation.

V. Numerical Results

In this section, we show results on the MNH during a session as a function of the number of access gateways and the number of cells served by each gateway. We then use our model for three case studies relevant to the effect of the number of cells per AGW on mobility, the effect of expanding the network by adding AGWs, and the airlink load of the MobileIP protocol for both mobility patterns. Notice that in all figures, two axes are used for comparison and better clarity.

1) *The mean number of handoffs:* In Fig.3, we show the MNH per session as function of the number of AGWs in the network for both directed (left y-axis) and random mobility models (right y-axis). While random mobility is more suitable for pedestrians who move slowly in the system and directed mobility is more often encountered for fast moving users, we study two cell residence times 4 and 40 mins for random mobility to allow comparison with the directed mobility with residence time of 4 mins. The mean session duration is set to 40 mins. We show the MNHs for the entire session as well as for partially served sessions in the network as function of the size of the number of access gateways in the network with each serving 10×10 cells. First, we observe that directed mobility results in larger number of handoffs than random mobility in all cases which conforms with our intuition. We also observe that the mean number of handoffs during the whole session (dashed lined) serves as the asymptote for the mean number of handoffs in the network. This also conforms with our expectations and can rather be very useful for assessing the largest impact of AGW handoffs on the QoS perceived by users for realtime services. Finally, in both cases for random mobility, we observe that since random movers are often more "localized" than the directed movers, they are unlikely to leave the network to other operators and hence the MNH in the whole session (eq. 30) can be used as a good approximation for the MNH the random movers incur.

2) *The number of cells per AGW:* This case is of particular importance due to the trend to have flatter networks with smaller sized equipment (a.k.a AGWs)[1]. We consider a network that consists of a fixed number of cells as 10×120 and study the AGW residence time as function of the number of AGWs for directed and random mobility. For comparison purposes, we study both movement patterns under at the same cellular residence time ($E[R_c]$= 4 min). We show results for the residence time incurred by new sessions R_{g_1} and for handoff sessions R_g derived in (1), (3), (13), and (29). We also show values for the two possible residence times for handoff sessions R_{g_a} and R_{g_b} derived in (18) as observed by random movers. As shown in Fig. 4, when the whole 10×120 cells are served by one gateway, the gateway residence times are very large. Directed movers always observe a relatively short residence time as they never change their movement direction, while random movers experience very long residence times as their movement behavior is relatively localized. We also show that for random users who make at least one handoff, they are most likely to incur R_{g_a} with a likelihood of $N_c/(N_c + 1)$ after their first handoff and hence dominate the gateway residence time R_g.

3) *The number of AGWs in the network:* This case is of particular importance to operators who expand their coverage and want to assign the new cells to their existing AGWs. Since in our model we only consider horizontal mobility (east-west) between gateways, we vary the number of cell columns (N_c) per AGW. We study a network consisting of 5 AGWs and observe that as the gateway coverage increases by adding more cells, the MNH decreases non-linearly. We observe that the MNH for directed movers is highly affected by the number of cells in the gateway and reduces drastically compared to random movers (from 2.6 to 1.2 for the whole session and 1.4 to 0.8 within the network). This is because directed users do not change their movement direction and hence has to cross N_c cell columns after making their first handoff.

4) *MobileIP signaling rate:* Finally, we use our model to investigate the effect of the handoff signaling on the airlink load in border cells. We consider protocol exchanges consisting of N_{ext} request plus response messages (including overhead) of sizes M_{RQ} and M_{RS} bits, respectively. Assuming a packet error rate of p_e and infinite number of retransmissions, the airlink load per message L_M is $L_M = N_{ext}(1 - p_e)^{-1}\ (M_{RQ} + M_{RS})$. The airlink signaling load in border cells belonging to the interior gateways, $\wp_B$, is given by the product of the session arrival rates, the number of handoffs per AGW ($\frac{E[N_{Hx}^{NET}]}{N_g}$), and the load per message (L_M) in bits divided by the number of interior border cells, $(2M_c)$. Assuming a uniformly distributed load among gateways, the airlink signaling load is $\wp_B$=$\sum_{x\in\{\Omega,\Phi\}} \frac{\lambda_x E[N_{Hx}^{NET}] L_M}{2M_c N_g}$. In Fig.6, we assume a 1xEVDO network running MobileIP with signaling overhead of 354 bytes for requests and responses including EVDO airlink overhead, i.e., 116 MobileIP registration req + 104 MobileIP registration response + UDP(8 bytes), IP(20 bytes), PPP(4 bytes), and EVDO(40 bytes). Different session durations represent type of services, such as Video (40 mins) and data service (240 mins). We observe that the airlink load per border cell grows in non-linear fashion with the number of gateways and reaches an asymptote as the number of

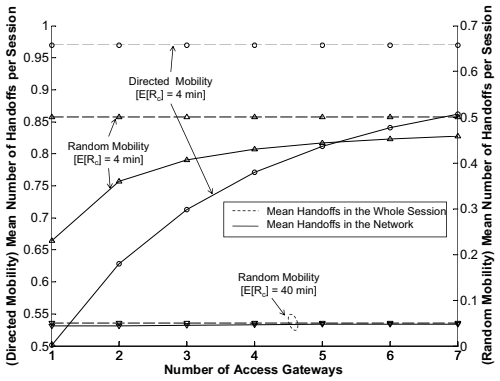

Fig. 3. The mean number of handoffs vs. number of gateways [$E[S] = 40$ min, $M_c = 10, N_c = 10$, $E_R = 4$ min, $C_R = 2$, 5% offnet traffic]

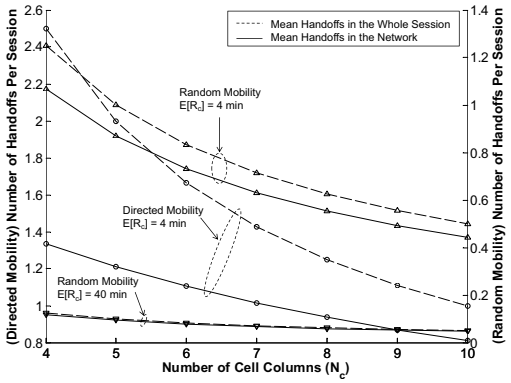

Fig. 5. The mean number of handoffs vs. number of cell columns (N_c) [$N_g = 5, E[S] = 40$ min, $M_c = 10$, $E_R = 4$ min, $C_R = 2$, 5% offnet]

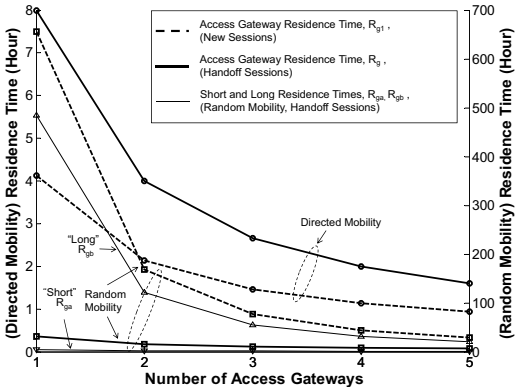

Fig. 4. Mean gateway residence time vs. number of access gateways in the fixed area [$E[S] = 40$ min, $M_c = 10$, $E_R = 4$ min, $C_R = 2$, 5% offnet]

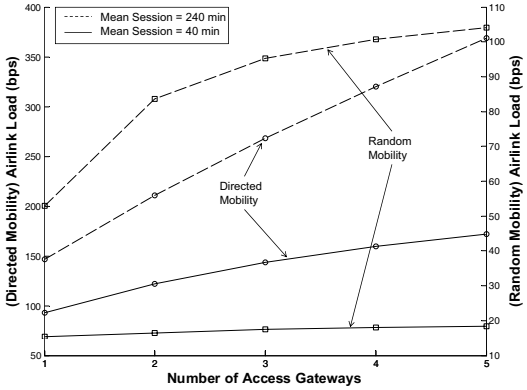

Fig. 6. Airlink load in border cells as a function of the number of AGWs

gateways (i.e., coverage) increases which is proportional to the maximum number of handoffs within the whole session.

VI. Conclusion

In this paper, we proposed a fundamental approach to obtain the mean number of handoffs for a session, for directed and random mobility. The main set of results indicates that the number of handoffs in the network is a non-linearly increasing function of the session duration and the number of access gateways, for both mobility patterns. We demonstrated the practical relevance of the approach by showing the results for the airlink load as a function of handoffs and for a range of session statistics. Our future work includes generalizing the mobility patterns and the topologies, investigating the error introduced by the exponential assumption, and applying our results to study the performance of other pivotal protocols such as SIP and Diameter in the context of next generation IP based wireless network architectures such as LTE/SAE.

References

[1] E. Andrews, 'Migrating to Flatter, All-IP Wireless Networks,' Converge Network Digest Online Magzine Jan, 2008

[2] 'A Flexible All-IP Wireless Network', product datasheet, IPWireless, 2004 http://www.ipwireless.com/pdfs/network.pdf

[3] R. Rodriguez-Dagnino et al, 'Counting Handovers in a Cellular Mobile Communication Network: Equilibrium Renewal Process Approach,' Performance Evaluation, v52(2), 2003

[4] Y. Fang, 'Modeling and Performance Analysis for Wireless MobileNetworks: A New Analytical Approach,' IEEE Trans. Networking, v13(5), 2005.

[5] P. Orlik, S. S. Rappaport, 'A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions,' IEEE JSAC, June 1998.

[6] K. Yeo, C. Jun, 'Modeling and Analysis of Hierarchical Cellular Networks With General Distributions of Call and Cell Residence Times,' IEEE Transactions on Vehicular Tech, Vol. 51, No. 6, Nov 2002

[7] A. Zahran et al, "Mobility Modeling and Performance of Heterogeneous Wireless Networks," Trans. Mobile Computing, v7(8), Aug 2008

[8] I. Akyildiz, W. Wang "A Dynamic Location Management Scheme for Next-Generation Multitier PCS Systems,"IEEE Trans. Wireless Communications, Vol. 1, No. 1, Jan 2002

[9] A. Bar-Noy, I. Kessler and M. Sidi, "Mobile users: To update or not to update", in Proc. INFOCOM'94, June 1994

[10] B. Liang et al, "Predictive distance-based mobility management for multi-dimensional PCS networks," Trans. on Networking, 11(5), 2003.

[11] S. Nanda, 'Teletraffic models for urban and suburban microcells: Cell sizes and handoff rates,' IEEE Trans. on Vehicular Tech. 42(4) (Nov. 1993)

[12] S. Zaghloul, W. Bziuk, A. Jukan, "Signaling and Handoff Rates at the Policy Control Function (PCF) in IMS," IEEE Comm. Letters, July, 2008.

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

[14] R. Perera et 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.

[15] W. Bziuk, S. Zaghloul, A. Jukan, 'A New Framework for Characterizing the Number of Handoffs in Cellular Networks,' 5th Polish-German Teletraffic Symposium, Berlin, 2008

[16] W.Bziuk, S. Zaghloul, A. Jukan, 'The Spatial Effect of Mobility on the Mean Number of Handoffs: A New Theoretical Result,', accepted for publication in ICC'09, Germany, 2009