Wireless Networks 11, 319–332, 2005
© 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

# Handoff Performance Improvement with Latency Reduction in Next Generation Wireless Networks

ÖZGÜR B. AKAN and BUYURMAN BAYKAL*
*Department of Electrical & Electronics Engineering, Middle East Technical University, Ankara/Turkey*

**Abstract.** The next generation (NG) wireless networks are expected to provide mobile users with the real-time multimedia services. High sensitivity to time constraints like delay and jitter is one of the important characteristics of the multimedia traffic. In order to maintain a certain quality of service (QoS) level, the handoff latency should be minimized. Furthermore, if the new cell is not ready at the actual handoff time, the handoff call may be even forced terminated. Hence, the handoff preparation latency directly affects the performance of the cellular networks in terms of QoS support and the handoff blocking probability. In this paper, we present the expected visitor list (EVL) method to achieve reduced handoff blocking probability and maintain a certain QoS level in the network by minimizing handoff preparation latency. The handoff signaling decomposition is introduced to make the neighbor cells aware of the resource demands and QoS requirements of the mobile terminal before the actual handoff time. The obtained information about the prospective active mobile terminal is stored in an EVL entry at the neighbor cells. The call admission control (CAC) with QoS-provisioning is run against each EVL entry. According to the CAC result, the network preparation algorithms are executed and the results are stored in the entry. No resource reservation or allocation is performed in advance, and the varying network conditions are reflected to validity and admission status of the entries. The results of handoff preparation algorithms stored in the EVL entry are activated at the actual handoff time and hence the handoff latency is minimized. Performance evaluation through mathematical analysis and extensive simulation experiments show that the EVL method reduces handoff latency and hence handoff call blocking probability significantly without introducing high overhead.

**Keywords:** handoff management, handoff latency, handoff call blocking, cellular networks, QoS provisioning

## 1. Introduction

The real-time multimedia is one of the new services which are expected to be provided by the next generation (NG) wireless networks. It requires timely delivery of the encoded multimedia streams such as video and audio. In order to maintain certain service quality, the QoS-oriented *DiffServ* [4] and *IntServ* [16] architectures are proposed for the Internet. In the NG wireless networks, however, a new delay element due to the mobility, i.e., handoff latency, should also be addressed to preserve agreed QoS-level throughout the connection. Any additional latency introduced during connection time degrades the service level received by the mobile terminal. Thus, the handoff latency is an important factor in the QoS performance of the NG wireless Internet.

The handoff latency also directly affects the overall network performance in terms of handoff call blocking probability. An active call is forced to termination during handoff due to two main latency factors. The first one is the handoff decision which is determined in accord with the signal strength measurements performed by the network or the mobile terminal. The delay in the handoff decision may cause signal loss during the cell-boundary crossing period. Conversely, early decision increases the unnecessary handoff probability leading to so-called *ping-pong* effect [12].

* Corresponding author.
E-mail: buyurman@metu.edu.tr

The second factor, which is primarily in the scope of this paper, is the handoff preparation delay for the prospective active mobile. After the handoff decision is made, network performs several tasks for the incoming mobile terminal. It first runs CAC algorithm against the mobile terminal and then prepares the new cell for it in case of acceptance. The network preparations for the admitted handoff call includes the determination of crossover switches (CoX), construction of paths for the newly transferred connection and applying a channel assignment strategy for the mobile in its new cell. The execution of CAC and then all preparation tasks require certain amount of time, i.e., the handoff preparation latency, which contributes to the overall handoff delay. If the mobile terminal is already in the new cell before its new environment is ready, then its connections will be terminated. For this reason, existing handoff strategies mostly try to use efficient handoff initiation control mechanisms and fast network preparation algorithms [1,11]. High accuracy in the handoff decision and fast preparation algorithms may be yet inadequate to provide seamless handoff support. If the incoming mobile has certain QoS requirements specified by its service-level-agreement (SLA) with the network, then network should also perform end-to-end QoS provisioning along the new routes of the mobile terminal. This amplifies the problem by increasing the handoff preparation latency.

There exist many studies in the literature to improve handoff performance. Most of them, however, aim to do that by mobility prediction and/or early resource reservation for incoming mobiles. In [2], a bandwidth allocation scheme is proposed to

guarantee QoS level by path prediction and preserving bandwidth along the path. An adaptive resource reservation for handoff requests based on the predefined mobile and location profiles is presented in [9]. Guard channel concept is introduced in [8] which allows new call initiation if $m + 1$ channels out of $M$ are available, reserving $m$ channels for handoff requests where $m + 1 < M$. The shadow cluster concept is introduced in [7] to perform resource estimation and call admission control by using active mobile probabilities to predict the future locations and resource demands to perform resource reservation. However, the reservation based approaches may result in severe degradation in overall utilization of the scarce wireless resources by performing unnecessary resource reservations. [2,5,15] depend on the characterization of the mobility patterns of active mobiles, which also may not hold for the next generation wireless networks. On the other hand, a distributed call admission control scheme is proposed in [10], which collects the number of call requests and average rates in target and neighbor cells and then performs admission control based on the bandwidth demands of current and prospective mobiles. This scheme may not provide each mobile terminal with seamless handoff support, since it lacks individual resource demands and QoS specifications of the mobile terminals to prepare new networking environment.

Instead of early resource reservation with mobility prediction, the overall handoff performance can be enhanced by handoff latency minimization. In this paper, we present expected visitor list (EVL) concept to reduce handoff blocking probability and provide certain QoS level in the network by reducing the handoff preparation latency. The conventional handoff signaling is decomposed into two parts so that the neighbor cells become aware of the resource demands and QoS requirements of their prospective active mobile terminals in advance. Each active mobile terminal is represented by an EVL entry in the EVL list of the neighbor cells of its current cell. The call admission control (CAC) with QoS-provisioning is run against each EVL entry. If the possible incoming active mobile is admitted, then necessary handoff preparation algorithms are executed and their results are stored in the proper fields of the EVL entry. Any change in the resource availability of the cell or the resource demands of the active mobile are reflected to the admission status of the EVL entry. In this method, no resource reservation or allocation is performed in advance. The results of the pre-execution of handoff preparation algorithms for the active mobile are only activated at the actual handoff time and hence the handoff latency is minimized. Performance evaluation through mathematical analysis and extensive simulation experiments shows that the proposed method reduces handoff call blocking probability significantly without introducing high overhead.

The remainder of this paper is organized as follows. Section 2 presents the decomposition of the conventional handoff signaling and the structure and operation of the EVL method. In Section 3, the analytical modeling of the EVL method and its mathematical performance analysis are presented. The performance evaluation of the EVL method and the simulation results are presented in Section 4. Finally, the paper is concluded in Section 5.

## 2. EVL: Expected visitor lists

The existing handoff strategies utilizes a handoff signaling which transfers all required information at the actual handoff time. This may significantly increase overall handoff latency. In this section, we present a decomposition of classical handoff signaling to avoid this problem. The structure of the expected visitor list (EVL) and how it operates in cooperation with the decomposed handoff signaling for improved handoff performance are introduced afterwards.

### *2.1. Handoff signaling decomposition*

In the conventional handoff, the new base station is informed about the resource demands of the incoming mobile whenever a mobile needs to change its network point of attachment. This is required for the network to run a call admission control algorithm against the incoming mobile. If the incoming mobile is admitted, then the network preparations for the admitted handoff call is initiated. These preparations include channel assignment and running some routing algorithms to construct a new path to new location. The target base station is provided with the information about the incoming mobile via classical handoff signaling among the mobile terminal, current base station and target base station as shown in figure 1. In this approach, all information about the mobile terminal is transferred to the target base station after the handoff decision is made.

The conventional handoff signaling leads to high handoff latency, $T$. This is basically because all information required for CAC and then network preparations can be transferred after this signaling occurs. If $T$ is not sufficiently low, then the handoff request is denied.

In order to minimize the latency, we propose to decompose the conventional handoff signaling. The entire signaling is divided into two parts as shown in figure 2. The resource demands and QoS specifications of the mobile terminal is sent to the new base station in advance, then only mobile identity

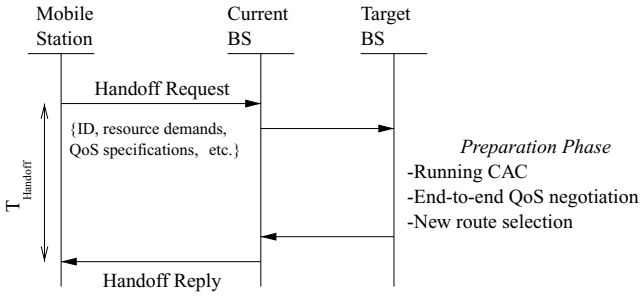


Figure 1. Conventional handoff signaling to inform the new base station for the incoming mobile.

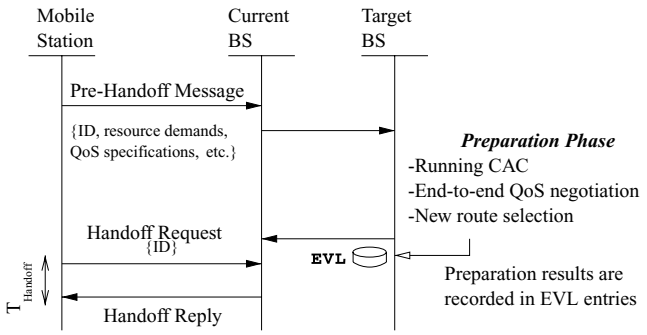


Figure 2. Decomposed handoff signaling to inform new base station about the incoming mobile in advance.

is sent at the actual handoff time. The objective of this is to enable the new base station to perform some time consuming tasks for possible handoff request from mobile terminal in advance. With the help of the decomposed handoff signaling, new base station learns resource demands and QoS specifications of the mobile terminal in advance. The CAC and then network preparation algorithms are executed for the prospective mobile terminal. The results are stored in EVL entry of the mobile terminal to be used at actual handoff.

### *2.2. EVL entry structure*

In the proposed method, once the information of an active mobile terminal is obtained by its current base station from decomposed handoff signaling, this is propagated to the all neighboring cells. This information is stored at the EVL of each neighbor base station. A typical EVL entry to store active mobile information is shown in figure 3.

An EVL entry for an active mobile terminal includes the information received from the decomposed handoff signaling described earlier. All obtained mobile specific information is entered into the fields described below:

- *Resource Demand:* The resource demands of the incoming active mobile terminal is entered into this field. The bandwidth and the number of channels required are some examples. The information entered in this field is used during call admission control for the incoming mobile.
- *QoS Specifications:* The QoS specifications of the mobile terminal are entered into this field. These QoS specifications can be a set of QoS variables which are already agreed on the SLA with the network. This information is used for end-to-end QoS-provisioning along the new connection paths of the incoming active mobile terminal.
- *Multicast Groups:* The multicast addresses of the groups that mobile terminal is a member of are entered into this field of the entry. This can be used to achieve IGMP query latency avoidance in Mobile IP multicast [3].

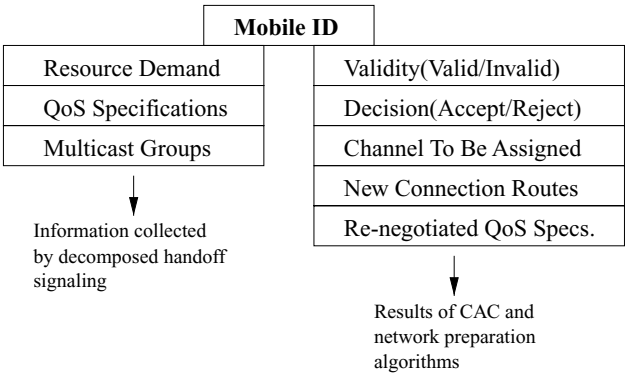


Figure 3. An EVL entry structure.

Once the entry is created, the call admission control is run against the mobile terminal and *Decision* field of the entry is set accordingly. At this point, no resource reservation is performed for any prospective active mobile terminal. If there is adequate resources for admission of a mobile terminal, then its admission is granted. If CAC result is accept, then the rest of the network preparation algorithms are executed for the mobile terminal and the results are stored in the appropriate fields described below:

- *Validity:* At the end of network preparations, this field is set valid indicating that the information in the fields of EVL entry is up-to-date. If any change occurs in network that may obsolete the entry, then the field is reset to invalid.
- *Decision:* The information in the resource demand field of the entry is used to run the call admission control for this entry. The result of the CAC is entered into this field.
- *Channel to be Assigned:* The channels that would be assigned at the actual handoff request are entered to this field so that the assignment would only be activated at the actual handoff time.
- *New Connection Routes:* The new routes for the existing connections of the active mobile terminal are determined and entered to this field. During the handoff, only these routes are switched for the connections.
- *Re-negotiated QoS Specifications:* During the end-to-end QoS-provisioning, if it is not possible to satisfy all QoS requirements of the mobile terminal, re-negotiation of the QoS variables may take place. In case of agreement on the adjusted QoS specifications, the new set of variables are entered into this field.

When all preparation is done for the mobile terminal, the *Validity* field of its entry is set valid. This field is important since it indicates the validity of the information stored in the entry. Any change in the resource status of the cell or the demands of the mobile terminal may affect the admission status of the entry. Therefore if such change occurs in the network, it may invalidate the entry to be reprocessed. The details of the EVL entry invalidation and processing are discussed in the next subsection.

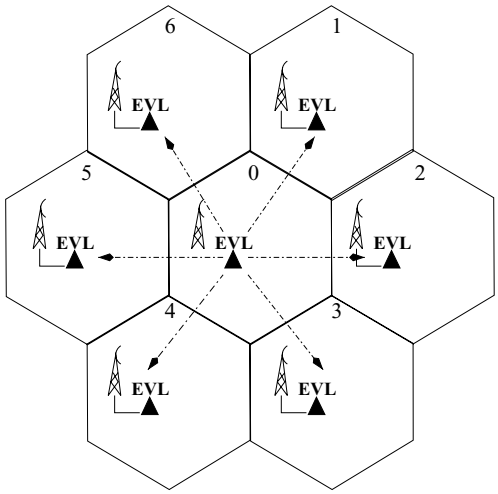


Figure 4. A sample cellular network topology with EVL employment.

### 2.3. EVL operation

The EVL method operation goes with cooperation of the decomposed handoff signaling and EVL entry processing at the base stations of each neighbor cell. The sample cell configuration for an EVL method is shown in figure 4.

In this scenario, the mobile specific information about the $i^{th}$ active mobile terminal, $MT_i$ in $cell_0$, is obtained via decomposed handoff signaling as explained before. This information is broadcast to the all neighbor cells, i.e., $cell_j$ for $1 \leq j \leq 6$. The information received is stored as an entry for the $MT_i$ in all $EVL_j$ in $cell_j$ for $1 \leq j \leq 6$. In each cell, call admission control is run for the entry. If the result is accept, then the network preparation algorithms are executed and their results are stored in the appropriate fields of the entry. At this point the entry of the $MT_i$ is valid, hence the information in the fields of this entry can be used at the actual handoff request. By this way, no preparation latency is experienced during actual handoff execution, and hence reduced handoff latency is achieved.

The EVL entries become obsolete due to varying resource status of the cell or the resource demands of the active mobile terminals. The decrease in the resource capacity of the cell may cause some accepted entries begin rejected. In the same way, some entries with reject decision may be admitted if the actual resource availability of the cell increases. The events, which can obsolete the EVL entries, are called *Invalidator Events*. The events which can obsolete entries with accept and reject decisions are different. Hence, the invalidator events can be categorized into *Accept Invalidator* (AI) and *Reject Invalidator* (RI) events as follows:

1. *Accept Invalidators*
   (a) A successful handoff completion to the target cell.
   (b) New call arrival in the target cell.
   (c) Active mobile terminal increases its resource demands.

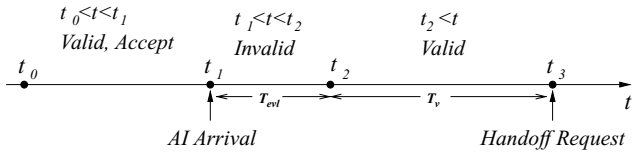


Figure 5. A typical validation timeline example.

2. *Reject Invalidators*
   (a) A successful handoff completion from the target cell.
   (b) Call termination in the target cell.
   (c) Active mobile terminal decreases its resource demands.

The events 1.(a) and 1.(b) invalidates all EVL entries with accept decision, while the event 1.(c) only invalidates the entry whose owner increases its demand. Similarly, the events 2.(a) and 2.(b) invalidates all rejected entries, and the event 2.(c) invalidates the entry itself. The invalidated entries are reprocessed to reflect the recent changes in network conditions or their resource demands to their admission status. After the reprocessing of the invalidated EVL entries, they become ready to be used in case of actual handoff request.

As an example, an entry is valid and accepted at $t = t_0$ in figure 5. At $t_1$, a mobile terminal in the target cell receives a new call and hence causes a decrease in the total resource allocation of the target cell. Since the earlier admission decision was made based on the handoff support of the cell at a prior time, $t = t_0$, the entry should be re-evaluated according to the latest conditions. Hence, this $AI$ invalidates the entry with accept decision at $t = t_1$. The EVL processing of the invalidated entries take a certain amount of service time ($T_{\text{evl}}$). At $t = t_2$, the entry becomes ready again and the handoff requesting active mobile finds the decision and network preparations ready at $t_3$ provided no other invalidator event arrived until $t_3$.

The performance of the proposed method is directly related to the validity of the entries throughout the time. The active mobile terminals, which find their entries valid and accept at the actual handoff request, benefit from the EVL method. Hence, the handoff preparation period is not a latency source, which may cause forced termination of active calls as in case of existing methods. Others can not take the advantage of the method and are served as in the case in classical approach experiencing higher handoff latency and hence higher probability of handoff call blocking.

## 3. Analysis of the EVL Method

In this section, we analytically investigate the performance of the EVL method in terms of handoff latency, probability of handoff call blocking, and the overhead incurred in Section 3.1, 3.2 and 3.3, respectively.

### 3.1. Handoff latency

The mobile terminals that have valid EVL entries with accept decision at the actual handoff time take the advantage of EVL

method and hence experience significantly reduced handoff latency. This, in turn, increases the probability of successful handoff completion as explained in Section 2.

The decrease in the handoff latency experienced by the mobile terminals due to the EVL method leads to decrease in the average handoff latency of the entire system. Let $T$ be the handoff latency due to the handoff initiation and the execution of preparation algorithms such as CAC, new route determination, and channel assignment. Let $\widehat{T}$ be the reduced handoff latency because of the EVL method. The average handoff latency, $\widetilde{T}$, of the system with the EVL deployment can be expressed by

$$\widetilde{T} = (1 - p_{av})T + p_{av}\widehat{T} \tag{1}$$

where $p_{av}$ is the probability that an active mobile terminal finds its EVL entry valid and accept at the actual handoff time. Note that $p_{av}$ depends on the processing capability of the EVL processors, the accept invalidator event rate, and the instantaneous resource availability of the target base station.

A valid EVL entry with accept decision becomes obsolete with an accept invalidator (AI) event as explained in Section 2.3. It is then re-processed and hence validated by the EVL processor; and its decision field is set according to the current resource availability of the target base station. Therefore, each EVL entry can be in any of the three states representing *accept valid* (AV), *reject valid* (RV), and *invalid* (I) as shown in figure 6.

Let $p_{rv}$, $p_{av}$, and $p_I$ be the probability that an EVL entry is valid with reject decision, valid with accept decision, and invalid, respectively. Then, the state equations of the Markov chain in figure 6 can be written as

$$\lambda_{RI}\, p_{rv} = \mu_{\text{evl}} p_r\, p_I \tag{2}$$

$$\lambda_{AI}\, p_{av} = \mu_{\text{evl}} p_a\, p_I \tag{3}$$

$$\lambda_{RI}\, p_{rv} + \lambda_{AI}\, p_{av} = \mu_{\text{evl}}(p_r + p_a) p_I \tag{4}$$

where

$\lambda_{RI}$ is the reject invalidator event arrival rate,

$\lambda_{AI}$ is the accept invalidator event arrival rate,

$\mu_{\text{evl}}$ is the service rate of the EVL processor,

$p_a$ is the probability that an EVL entry is accepted,

$p_r$ is the probability that an EVL entry is rejected.

Note that $p_{rv} + p_{av} + p_I = 1$. Therefore, it follows from (2), (3), and (4) that the state probability $p_{av}$, i.e., the probability of finding an EVL entry valid and accept, can be calculated as

$$p_{av} = \frac{\mu_{\text{evl}} p_a \lambda_{RI}}{\mu_{\text{evl}}\,[p_a(\lambda_{RI} - \lambda_{AI}) + \lambda_{AI}] + \lambda_{AI}\lambda_{RI}} \tag{5}$$

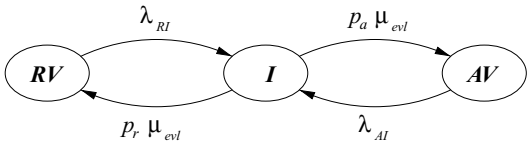


Figure 6. Three-state Markov chain for the state diagram of EVL operation.

Let $\lambda_{AI}^a, \lambda_{AI}^b, \lambda_{AI}^c$ be the arrival rates of the three accept invalidator events explained in Section 2, i.e., successful handoff rate, new call arrival rate, and resource demand change rate, respectively. Then, the accept invalidator arrival rate, $\lambda_{AI}$, is the total arrival rates of these events, i.e., $\lambda_{AI} = \lambda_{AI}^a + \lambda_{AI}^b + \lambda_{AI}^c$.

Given that $\lambda_h$, $\lambda_{CA}$, and $\lambda_{RC}$ are respective event rates of handoff request, call arrival, and resource demand change, then

$$\begin{aligned}\lambda_{AI}^a &= \lambda_h \frac{1}{n} N p_a n \\ &= \lambda_h N p_a,\end{aligned} \tag{6}$$

$\lambda_{AI}^b = \lambda_{CA}$, and $\lambda_{AI}^c = \lambda_{RC}$, where $n$ is the number of neighboring cells, $N$ is the average number of mobile terminals in each of the neighbor cells. Therefore, the AI arrival rate, $\lambda_{AI}$ can be expressed by

$$\lambda_{AI} = \lambda_h N p_a + \lambda_{CA} + \lambda_{RC} \tag{7}$$

Similarly, the reject invalidator event rate, $\lambda_{RI}$, is the sum of the arrival rate of the three reject invalidator events explained in Section 2, i.e., successful handoff rate, call termination rate, and resource demand change rate, respectively. Then, $\lambda_{RI}$ can be calculated by

$$\lambda_{RI} = \lambda_h N p_a + \lambda_{CT} + \lambda_{RC} \tag{8}$$

where $\lambda_{CT}$ is the call termination rate.

Consequently, by substituting (5) into (1), the average handoff latency with EVL deployment can be obtained as

$$\widetilde{T} = \frac{\mu_{\text{evl}}[T\lambda_{AI} - p_a(T\lambda_{AI} - \widehat{T}\lambda_{RI})] + T\lambda_{AI}\lambda_{RI}}{\mu_{\text{evl}}[\lambda_{AI} + p_a(\lambda_{RI} - \lambda_{AI})] + \lambda_{AI}\lambda_{RI}} \tag{9}$$

Here, $p_a$ depends on the resource capacity of the cell. Therefore, it can be approximately obtained by using *Erlang-B* formula for the probability of blocking in case of blocked calls clear situation in $M/M/m/m$ queue as

$$p_a = 1 - \frac{(\lambda/\mu)^m/m!}{\sum_{k=0}^{m}(\lambda/\mu)^k/k!} \tag{10}$$

where $m$ is the number mobile terminals that the cell can accommodate; $\lambda$ and $\mu$ are the mobile terminal arrival rate and service rate expressed as

$$\lambda = \lambda_{CA} + N\lambda_h \tag{11}$$

$$\mu = \mu_d + \lambda_h \tag{12}$$

where $\mu_d$, $\lambda_h$, and $\lambda_{CA}$ are respective rates of call duration, handoff, and call arrival; $N$ is the average number of mobile terminals in each of the neighbor cells.

In a practical scenario, for an EVL processor with very high processing capability, $\mu_{\text{evl}}$ can be reasonably assumed to very high. Consequently, $\widetilde{T}$ can be calculated by

$$\lim_{\mu_{\text{evl}} \to \infty} \widetilde{T} = \frac{T\lambda_{AI} - p_a(T\lambda_{AI} - \widehat{T}\lambda_{RI})}{\lambda_{AI} + p_a(\lambda_{RI} - \lambda_{AI})} \tag{13}$$

Hence, it follows from (7), (8), (11), (12), (10), and (13) that the average handoff latency in the system with EVL deployment, $\widetilde{T}$, can be calculated by (14). In order to analytically investigate the performance of the EVL method in terms

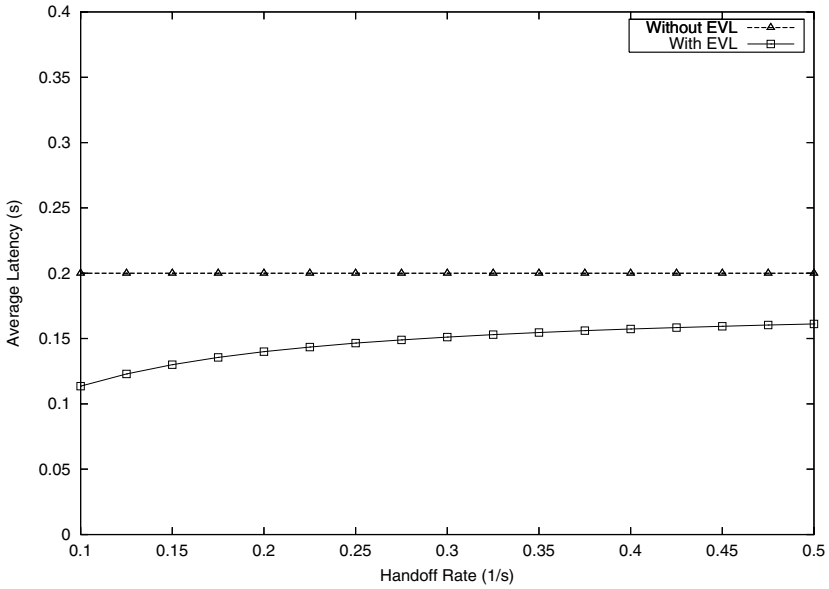


Figure 7. Average latency within coverage of a BS with/without EVL deployment for varying handoff rate.

of average latency reduction, we plot the average latency for varying handoff rate, $\lambda_h$, and the results are shown in figure 7. Here, we assume $m = 35$, $N = 50$, $\mu_d = 0.033$ (for exponential call duration), $T = 0.2$s, and $\widehat{T} = 0.05$s. We also assume that call arrival, call termination, and resource demand change events are following Poisson process with rates $\lambda_{CA} = 0.2$, $\lambda_{CT} = 0.2$, and $\lambda_{RC} = 0.5$, respectively.

$$\widetilde{T} = \frac{T(\lambda_{CA}+\lambda_{RC}) - \left[1 - \frac{\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^m \frac{1}{m!}}{\sum_{k=0}^{m}\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^k \frac{1}{k!}}\right]\left\{T\left(\lambda_{CA}+\lambda_{RC}-\lambda_h N\left[\frac{\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^m \frac{1}{m!}}{\sum_{k=0}^{m}\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^k \frac{1}{k!}}\right]\right) - \widehat{T}(\lambda_h N p_a + \lambda_{CT} + \lambda_{RC})\right\}}{\lambda_{CA}+\lambda_{RC}+\left[1 - \frac{\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^m \frac{1}{m!}}{\sum_{k=0}^{m}\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^k \frac{1}{k!}}\right](\lambda_h N + \lambda_{CT} - \lambda_{CA})} \tag{14}$$

As it is observed from figure 7, for wide range of handoff rate, the average latency with EVL deployment is acceptably lower than the one without EVL method although the latency increases with increasing mobility. The increase in the latency with increasing mobility is experienced as expected because the probability of finding EVL entry accept and valid decreases as the mobility increases. This analysis is further justified by the simulation experiment results presented in Section 4.2.

### *3.2. Probability of handoff call blocking due to latency*

A handoff request can be denied because of two main reasons as follows:

- *Resource Inadequacy:* The CAC algorithm examines the resource demands of the prospective mobile terminal that initiates a handoff request. If the new coverage area has available resources the request is granted. If there is not adequate resources, the handoff request is blocked and hence the active call is terminated as the mobile terminal moves into new cell. The handoff call blocking due to resource inadequacy is extensively studied in the literature [6,14] and is left beyond the scope of this paper.
- *Latency:* The handoff latency is composed of two main factors. The first is the delay experienced at the handoff decision process due to an inaccurate or late signal strength measurements performed by the mobile terminal or the network itself. The second is the latency experienced during the preparation for the prospective active mobile terminal after the handoff is initiated, i.e., handoff preparation delay. If the active mobile terminal is already in the new cell before its new environment is ready, then its connection is terminated.

Let $\widehat{T}$ and $T$ be the total handoff latency experienced by the active mobile terminal at the actual handoff time depending on if its EVL entry is valid accept or not, respectively. Let $t_m$ be the random variable denoting the time it takes for the active MT to cross cell-boundary and enter the new coverage area, i.e., handoff transition time. Then, if the EVL entry of the active MT is not accept valid at the handoff request time, the handoff request is blocked if $t_m < T$. Otherwise, the handoff request is blocked due to latency if $t_m < \widehat{T}$. Therefore, the probability, $p_b$, that the handoff call is blocked due to latency is expressed by

$$p_b = (1 - p_{av})\text{Prob}[t_m < T] + p_{av}\text{Prob}[t_m < \widehat{T}] \tag{15}$$

where $p_{av}$ is the probability that the EVL entry of the handoff requesting active MT is valid and has accept in its decision field as calculated by (5) in Section 3.1.

Note that $p_b$ decreases with increasing $p_{av}$ as $\widehat{T} < T$. Rearranging (15) yields

$$p_b = \text{Prob}[t_m < T] - p_{av}\text{Prob}[\widehat{T} < t_m < T] \quad (16)$$

Given that $f_m$ is the probability density function of the random variable $t_m$, $p_b$ can be calculated by

$$p_b = \int_0^T f_m(t)\mathrm{d}t - p_{av}\int_{\widehat{T}}^T f_m(t)\mathrm{d}t \quad (17)$$

The handoff transition time, $t_m$, is assumed to be random variable with gaussian density function [17] given by

$$f_m(t) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(t-\mu)^2/2\sigma^2}, \quad t > 0 \quad (18)$$

where $\mu$ and $\sigma^2$ are the mean and variance of $t_m$. Hence, $p_b$ can be restated as

$$p_b = \Phi\left(\frac{T-\mu}{\sigma}\right) - \Phi\left(\frac{0-\mu}{\sigma}\right) - \\ - p_{av}\left[\Phi\left(\frac{T-\mu}{\sigma}\right) - \Phi\left(\frac{\widehat{T}-\mu}{\sigma}\right)\right] \quad (19)$$

where $\Phi(x)$ is the cumulative distribution function of unit normal random variable $Z$, i.e., $Z = \frac{t_m-\mu}{\sigma}$.

$$p_b = \Phi\left(\frac{T-\mu}{\sigma}\right) - \Phi\left(\frac{-\mu}{\sigma}\right) - \frac{\left[1 - \frac{\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^m \frac{1}{m!}}{\sum_{k=0}^m \left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^k \frac{1}{k!}}\right]\left\{\lambda_h N\left[1 - \frac{\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^m \frac{1}{m!}}{\sum_{k=0}^m \left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^k \frac{1}{k!}}\right] + \lambda_{CT} + \lambda_{RC}\right\}}{\left[1 - \frac{\left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^m \frac{1}{m!}}{\sum_{k=0}^m \left(\frac{\lambda_{CA}+N\lambda_h}{\mu_d+\lambda_h}\right)^k \frac{1}{k!}}\right](\lambda_h N + \lambda_{CT} - \lambda_{CA}) + \lambda_{CA} + \lambda_{RC}} \\ \times\left[\Phi\left(\frac{T-\mu}{\sigma}\right) - \Phi\left(\frac{\widehat{T}-\mu}{\sigma}\right)\right] \quad (20)$$

It follows from (5), (7), (8), (10), (11), (12), and (19) that $p_b$ can be calculated by (20). To investigate the performance of EVL method in terms of probability of handoff blocking due to latency, we plot $p_b$ as a function of handoff rate. We assume $T = 0.2$s, and $\widehat{T} = 0.05$s as in Section 3.1; and $\mu = 0.24$ and $\sigma = 0.2$ as will be used for simulation experiments in Section 4. Using approximated values of $\Phi(x)$ from [13], $p_b$ is calculated and plotted in figure 8 as a function of handoff transition time mean, i.e., $\mu$. As observed from figure 8, $p_b$ increases with handoff transition time mean. The reason is the average latency experienced by mobile terminals also increases since the probability of finding EVL entry valid and accept degrades by mobility. However, the handoff blocking probability due to latency remains very low even for very high mobility, i.e., $p_b < 0.4$ for $\mu < 0.1$s.

### 3.3. Signaling overhead

The EVL method introduces the decomposed handoff signaling to inform the new base station about the incoming mobile in advance to reduce the overall handoff latency as described in Section 2.1. Furthermore, additional signaling between the base stations is also introduced by the EVL method to keep EVL entries updated. The conditions at which an additional message is sent are as follows:

(1) *New Call Arrival:* When a new call arrives in a base station (BS), this information is propagated to the neighboring BSs within decomposed handoff signaling. By this way, the neighboring BSs create an EVL entry for this active mobile terminal.

(2) *Call Termination:* When a call is terminated in a coverage area of a BS, this is also propagated to the neighboring BSs. Consequently, the neighboring BSs delete the EVL entry of the mobile terminal.

(3) *Resource Demand Change:* The resource demands of an active mobile terminal may change during its connection period. Similarly, other mobile specific information, e.g., QoS specifications and the multicast groups of which the mobile terminal is subscribed, may also change. Such event is transfered to the neighboring BSs to have them their EVL entries keep updated.

These additional messages incurred by the EVL method result in a signaling overhead.

Let $N$ be the number of mobile terminals in the coverage area of a base station. Given that $\lambda_{CA}$ and $\mu_d$ are the call arrival rate and the call service rate, then the number of mobile terminals that have active connections, $N_a$, and that have not, $N_i$, residing in a BS can be respectively expressed as

$$N_a = N\frac{\lambda_{CA}}{\mu_d + \lambda_{CA}} \quad (21)$$

$$N_i = N\frac{\mu_d}{\mu_d + \lambda_{CA}} \quad (22)$$

Any of the additional messages described above is transmitted to each of the neighboring BS. Let $b$ be the number of neighboring BSs, then the total number of additional messages sent from a BS to its neighbors due to three conditions state above, $C$, can be expressed by

$$C = bN_i\lambda_{CA} + bN_a\mu_d + bN_a\lambda_{RC} \quad (23)$$

where $\lambda_{RC}$ is the resource demand change rate.

It follows from (21) and (22) that

$$C = bN\frac{\mu_d}{\mu_d + \lambda_{CA}}\lambda_{CA} + bN\frac{\lambda_{CA}}{\mu_d + \lambda_{CA}}\mu_d +$$

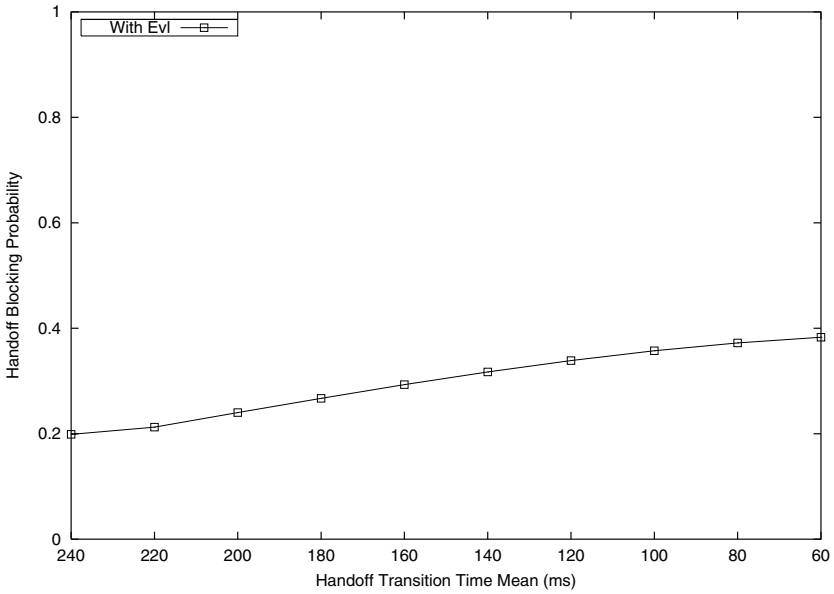


Figure 8. Probability of handoff blocking due to latency with/without EVL deployment for varying handoff rate.

$$+ bN \frac{\lambda_{CA}}{\mu_d + \lambda_{CA}} \lambda_{RC} \tag{24}$$

Consequently, by rearranging (24), the total signaling overhead, $C$, incurred by EVL method between a BS and its neighbors can be obtained by

$$C = bN \frac{(2\mu_d + \lambda_{RC})\lambda_{CA}}{\mu_d + \lambda_{CA}} \tag{25}$$

It is observed from (25) that the signaling cost, $C$, mainly depends on the invalidator event rates and the number of mobile terminals. This dependence will be further observed in the simulation experiments presented in Section 4.

## 4. Performance evaluation

In order to further investigate the performance of the EVL method, extensive simulation experiments are conducted. The average handoff latency, probability of handoff blocking due to handoff latency and resource inadequacy, and signaling overhead incurred by the EVL method are investigated in this section.

### *4.1. Simulation environment*

To explore the efficiency of our methodology, a discrete event simulator is implemented. In the simulations, one central cell neighbored by 6 adjacent cells as shown in figure 4. The call arrival, call duration, call residence time, handoff transition time, and resource demand change arrival are distributed with poisson, exponential, gamma, gaussian, and exponential, respectively. For the rest of the simulations parameters, we use $T = 0.2$s, and $\widehat{T} = 0.05$s as in Section 3.1; $\mu = 0.24$ and $\sigma = 0.2$ for gaussian handoff transition time random variable; $\lambda_{CA} = 0.2$, $\lambda_{CT} = 0.2$, and $\lambda_{RC} = 0.5$ for the call arrival rate, the call termination rate, and the arrival rate for the incremental changes in the mobile resource demand, respectively. The exponential call duration rate, $\mu_d$, is set as 0.033 seconds.

EVL processors in each cell are assumed to have very high processing speed. All invalidated entries are reprocessed upon their arrival to the EVL processor. The experiments are performed for two cases, i.e., with and without EVL deployment, to assess the performance improvement by the EVL method. All the simulations are run for varying mobility, i.e., handoff transition time mean, from 240 ms to 60 ms; and for varying number of active mobile terminals per cell, i.e., from 7 to 50.

### *4.2. Handoff latency*

To investigate the handoff latency reduction achieved by EVL method, the simulations are performed for varying mobility, i.e., handoff transition time mean, number of mobile terminals, and for varying handoff rate, i.e., $\lambda_h$. The results are plotted in figures 9–11. The average latency increases with mobility since number of mobile terminals that find their EVL entries valid and accept decreases for higher mobility. The probability that mobile terminal finds its EVL entry valid and accept, $p_{av}$, is also plotted in figure 12(a) for varying mobility. As observed in figure 12(a), $p_{av}$ decreases with mobility which, in turn, increases the average latency. Despite such behavior, EVL method achieves up to 40% latency reduction compared to the conventional handoff procedure.

Note also that the simulation results given in figure 10 for varying handoff rate and the corresponding analysis results given in figure 7 are, in fact, consistent, which justifies the analysis and the performance improvement achieved by the EVL method.

On the other hand, the average latency also increases with the number of mobile terminals as shown in figure 11. This

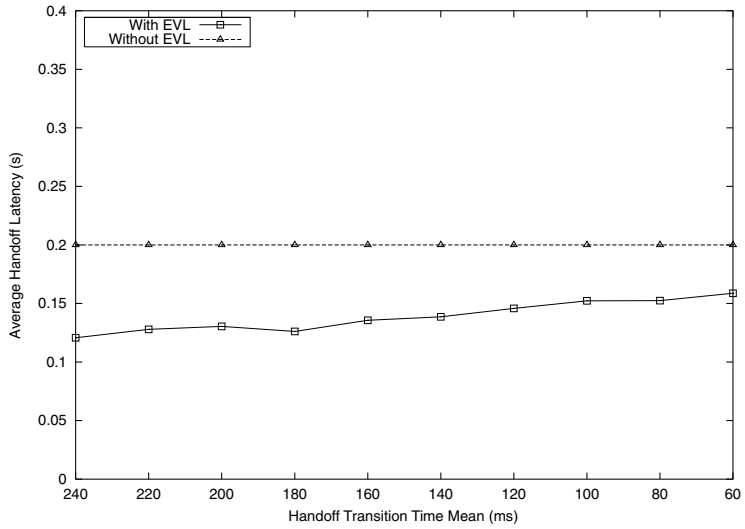


Figure 9. Average latency experienced by an MT within coverage of a BS with/without EVL deployment for varying mean mobility.

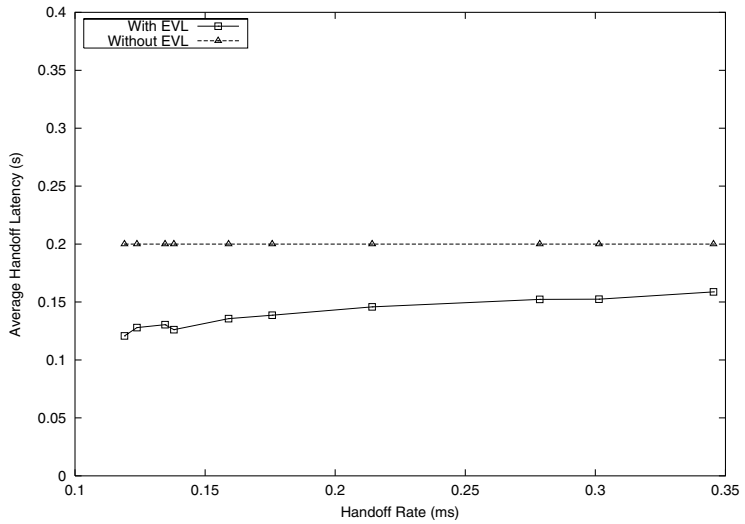


Figure 10. Average latency experienced by an MT within coverage of a BS with/without EVL deployment for varying handoff rate.

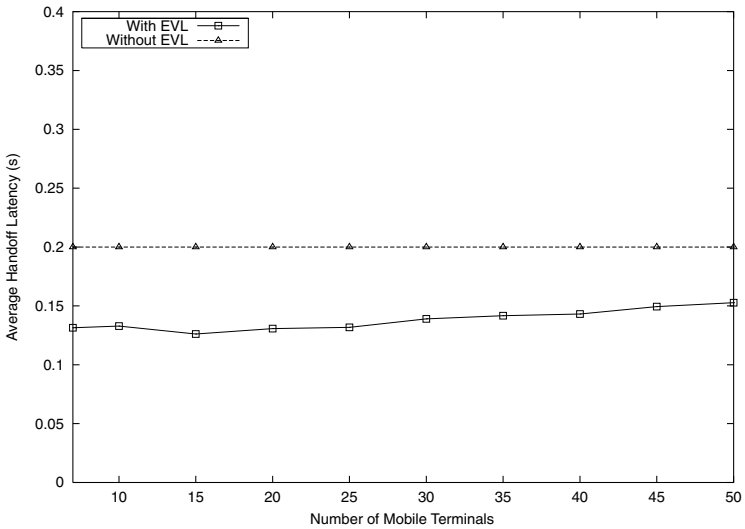


Figure 11. Average latency experienced by an MT within coverage of a BS with/without EVL deployment for varying number of mobile terminals.

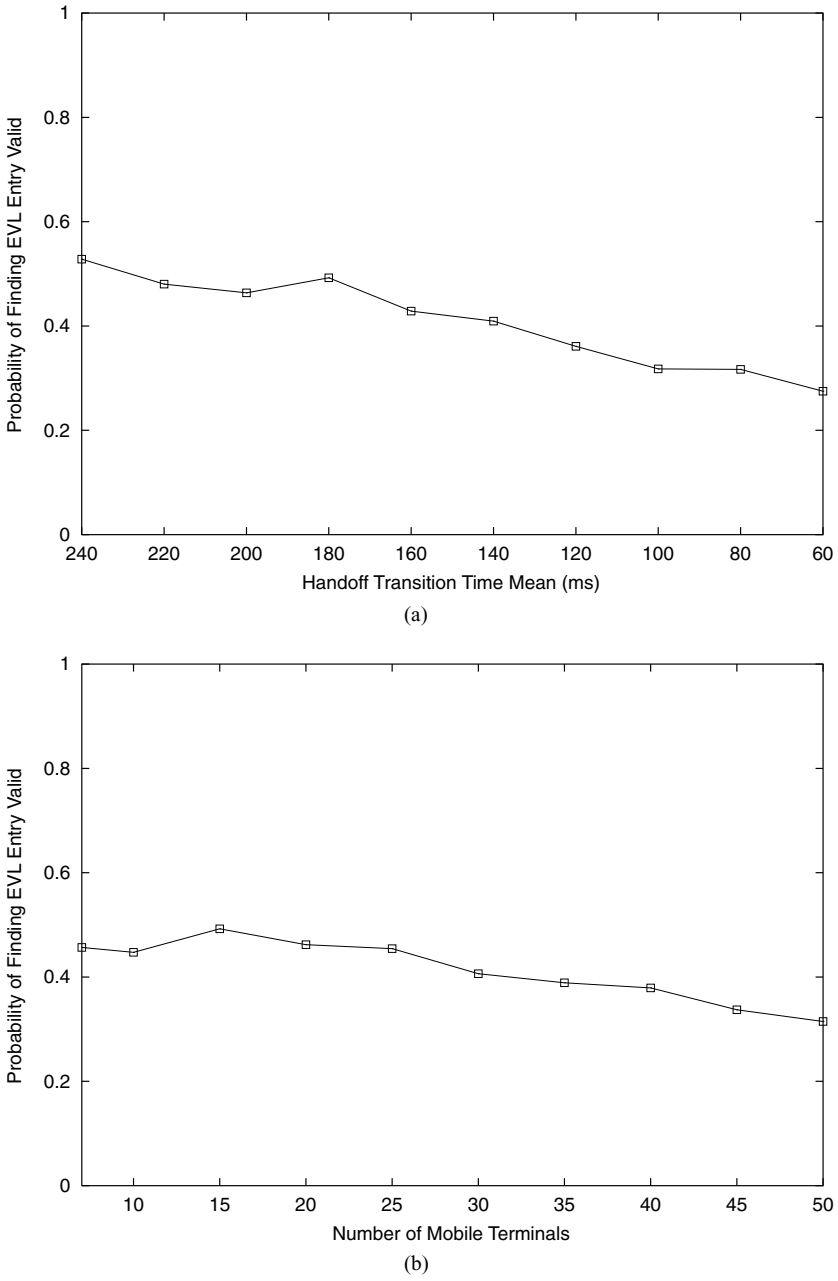


Figure 12. Probability of finding EVL entry valid for varying number of mobile terminals (a) varying mean mobility, (b) varying number of mobile terminals.

is because the acceptance probability, $p_a$, decreases as the number of mobile terminals in the cell increases. This also decreases the probability of finding EVL entry valid and accept as shown in figure 12(b). However, for a wide range of number of mobile terminals, EVL method significantly decreases the average handoff latency.

### *4.3. Probability of handoff call blocking*

The main objective of EVL method is to achieve lower handoff blocking probability by handoff latency reduction without any resource reservation for an incoming mobile terminal. The probability of handoff blocking due to latency is shown in figures 13 and 14 for varying mobility and the number of mobile terminals, respectively.

As shown in figure 13, the probability of handoff blocking due to latency increases with mobility due to the decrease in handoff transition time. However, EVL method reduces the probability up to 55% by reducing the average handoff latency experienced by the mobile terminals.

Note also that this result is consistent with the results of the analytical investigation in Section 3.2. Although there exist slight difference between the average latency values estimated by the mathematical analysis and the simulation results, this could be mainly because of the error introduced due to the Erlang approximation used for the probability of

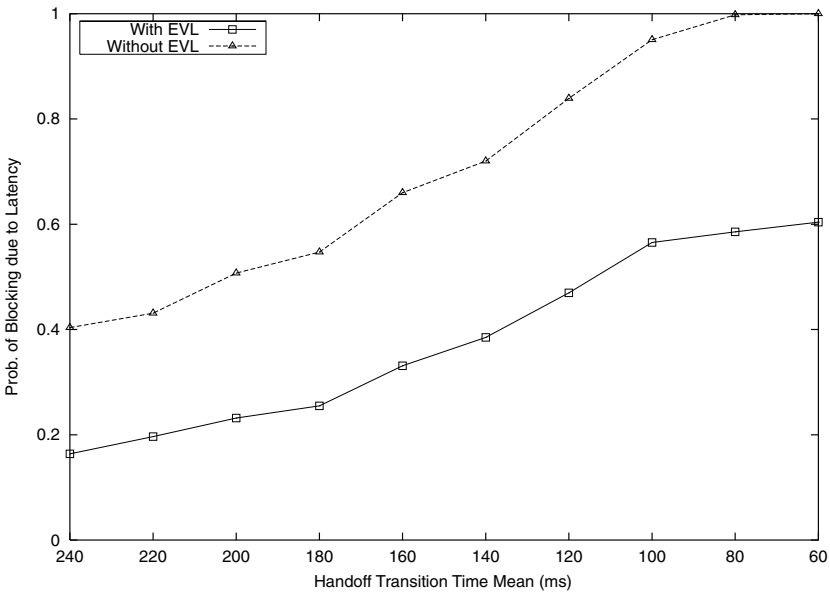


Figure 13. Probability of handoff blocking due to latency with/without EVL deployment for varying mean mobility.

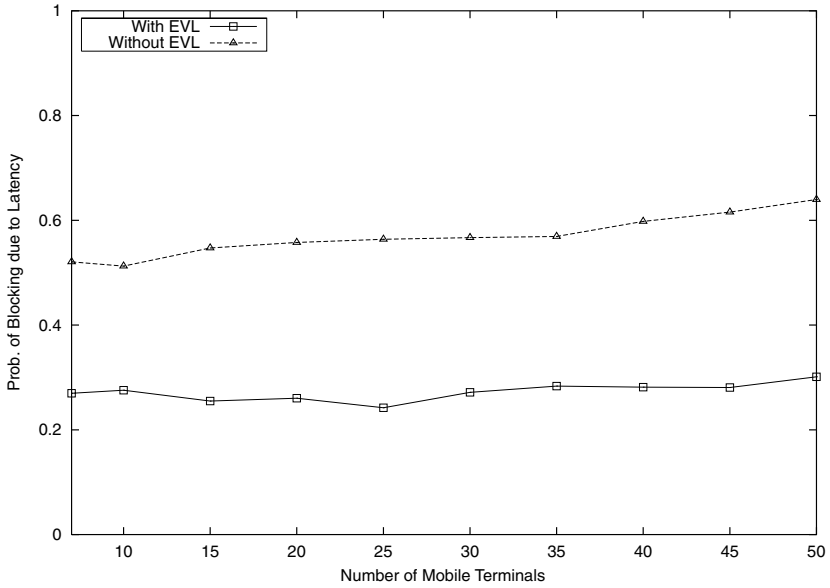


Figure 14. Probability of handoff blocking due to latency with/without EVL deployment for varying number of mobile terminals.

handoff call blocking. Despite this slight deviation, the results obtained by both the analysis and the simulation experiments are significantly close to each other justifying the analysis and the performance improvement achieved by the EVL method.

As shown in figure 14, the probability of handoff blocking due to latency is also significantly reduced for varying number of mobile terminals with EVL method. Although the probability increases with the number of mobile terminals, this increases is not as severe as it is for varying mobility case in figure 13. EVL method achieves up to 51% performance improvement in terms of probability of handoff blocking due to latency compared to conventional handoff approach.

The reduction achieved by the EVL method in the probability of handoff blocking due to latency directly affects the total handoff blocking probability of the system. The handoff blocking probability for varying mobility and number of active mobile terminals per cell are shown in figures 15 and 16. Although in both cases handoff blocking probability increases with mobility and number of mobile terminals as explained before, EVL method significantly improves the network performance by reducing blocking probability up to 50% in average.

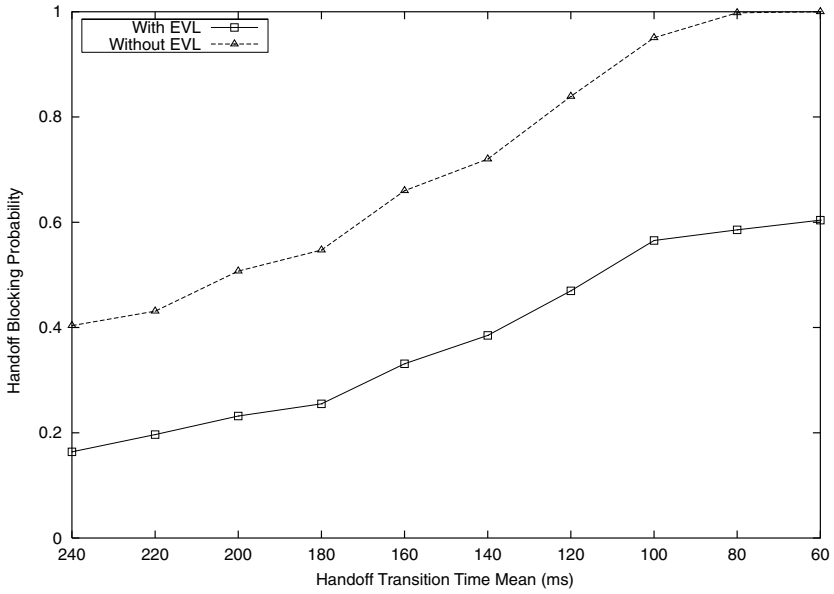


Figure 15. Probability of handoff blocking with/without EVL deployment for varying mean mobility.

### *4.4. Signaling overhead*

EVL method reduces the handoff latency and hence handoff blocking probability in an expense of signaling overhead due to additional messaging that takes place in the decomposed handoff signaling. Signaling overhead incurred by EVL method for varying mobility and number of mobile terminals are shown in figure 17(a) and 17(b), respectively.

In this case, the overhead is the average number of additional packets sent from a base station to each of its neighbors throughout the simulation time, which is set to 1000 seconds. As shown in figure 17(a), the signaling overhead does not vary with increasing mobility. This is consistent with the signaling cost function obtained in (25), which does not depend on the mobility. The overhead remains around approximately 15 packets between base stations for all handoff transition time mean values. This means that only 0.105 packets/s, i.e., 105 bytes/s for a packet size of 1000 bytes, bandwidth is used for decomposed handoff signaling.

Overhead increases with increasing number of mobile terminals as shown in figure 17(b). Recall that (25) is also proportional to the number of mobile terminals. However, even for 50 mobile terminals, the additional number of packets sent is 343, which means that 0.343 packets/s, i.e., 343 bytes/s for a packet size of 1000 bytes, bandwidth overhead is incurred by the EVL method. This overhead is very low compared to the

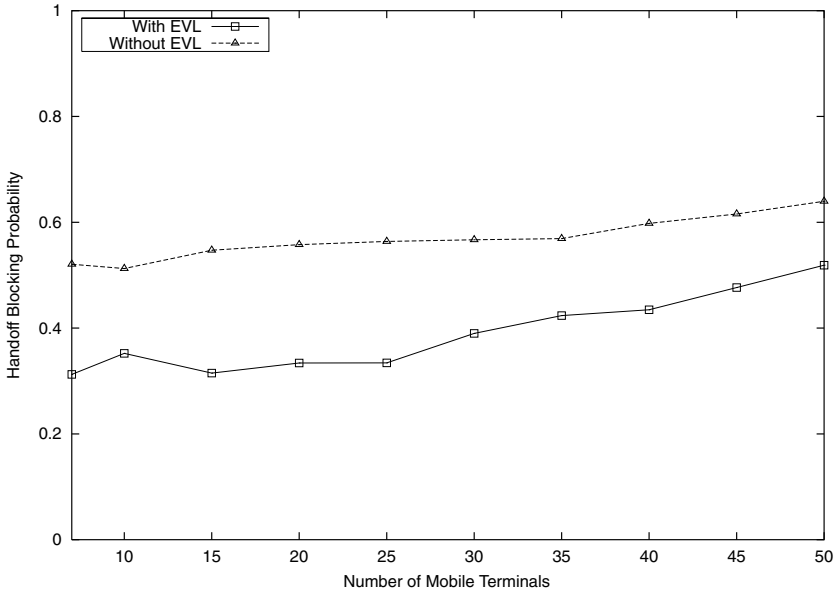


Figure 16. Probability of handoff blocking with/without EVL deployment for number of mobile terminals.

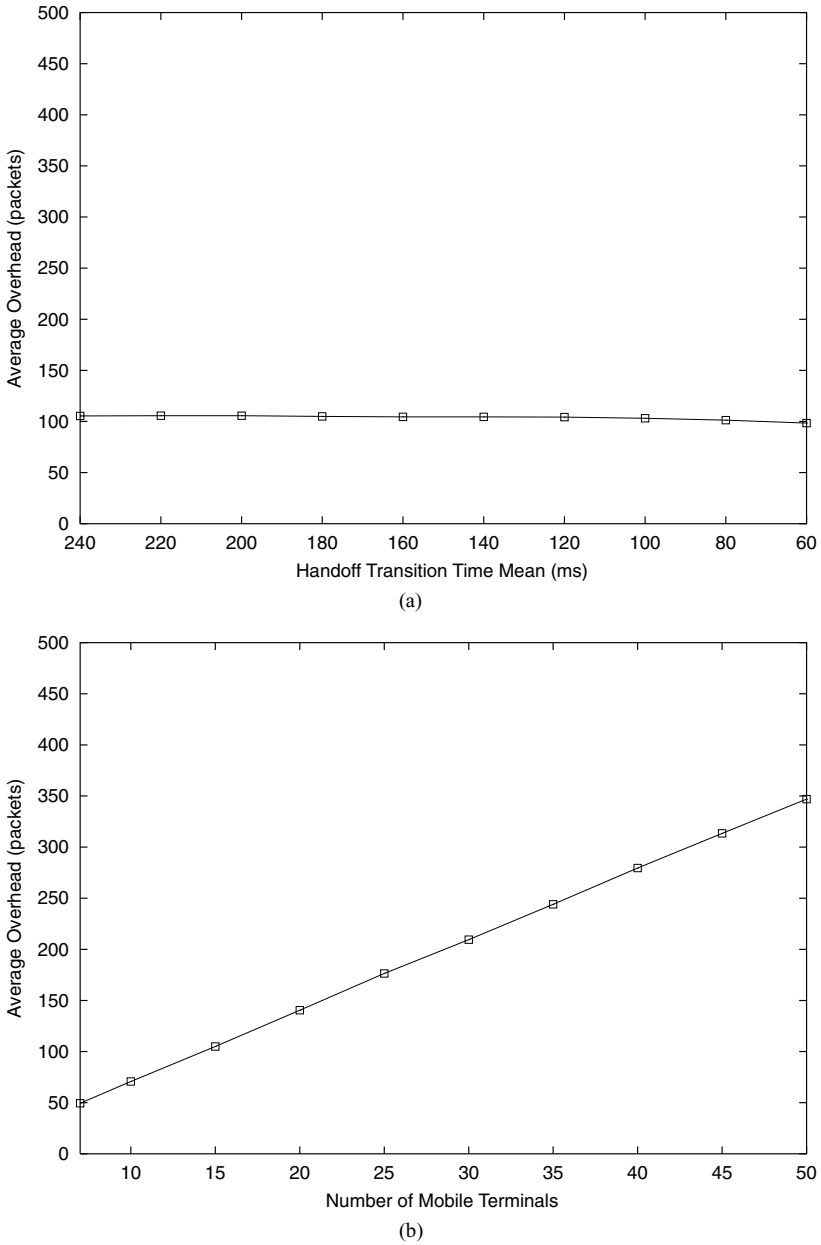


Figure 17. Average number of additional packets sent to each neighbor base station due to new decomposed handoff signaling for EVL operation with (a) varying mean mobility, (b) varying number of mobile terminals.

significant performance improvement in terms of up to 40% latency reduction and 55% handoff blocking probability reduction achieved by EVL method. Furthermore, the overhead incurred by the EVL method can be further reduced by adopting accurate trajectory prediction methods into the scheme.

## 5. Conclusions

In this paper, we presented expected visitor list (EVL) method to achieve reduced handoff blocking probability and maintain certain QoS level in the network by minimizing handoff preparation latency. The decomposed handoff signaling is introduced to enable the neighbor cells to obtain the resource demands and QoS requirements and store them in an EVL entry for the active mobile terminal. The call admission control (CAC) with QoS-provisioning is run against each EVL entry. According to the CAC result, the network preparation algorithms are executed and the results are stored in the entry. No resource reservation or allocation is performed in advance, and the changing network conditions are reflected to validity and admission status of the entries. The results of handoff preparation algorithms stored in the EVL entry are activated at the actual handoff time and hence the handoff latency is minimized. Performance evaluations via mathematical analysis and simulation experiments show that the EVL method reduces handoff latency and hence significantly reduces handoff call blocking probability without introducing high overhead.

## References


[1] I.F. Akyildiz, J. McNair, J.S.M. Ho, H. Uzunalioglu and W. Wang, Mobility management in next generation wireless systems, Proceedings of IEEE 87(8) (1999) 1347–1381.

[2] A. Aljadhai and T. Znati, A predictive bandwidth allocation scheme for multimedia wireless networks, in: *Conference on Communication Networks and Distributed Systems Modeling and Simulation (CNDS '97)* Phoenix, Arizona (1997) pp. 95–100.

[3] B. Baykal, O.E. Akdemir and O.B. Akan, An IP multicast handoff scheme with focus on IGMP sourced latency, in: *Proc. IEEE HSNMC 2002* (Korea, 2002) pp. 361–364.

[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, An architecture for differentiated services, in: *RFC 2474* (Dec. 1998).

[5] S. Choi and K. Shin, Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks, in: *Proc. ACM SIGCOMM '98* (Vancouver, British Columbia, Aug. 1998).

[6] S. Kundu and S. Chakrabarti, Call blocking in a mobile radio system with directed retry and priority handoff, in *Proc. IEEE Conf. on Personal Wireless Communications 2002*. (2002), pp. 163–167.

[7] D. Levine, I.F. Akyildiz and M. Naghshineh, A Resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept, IEEE/ACM Trans. Networking 5(1) (1997) 1–8.

[8] W. Li and A.S. Alfa, Channel reservation for handoff calls in a pcs network, *IEEE Trans. Veh. Technol.* 49(1) (2000) 95–104.

[9] S. Lu and V. Bharghavan, Adaptive resource management algorithms for indoor mobile computing environments, in: *Proc. ACM SIGCOMM '96* (Aug. 1996).

[10] M. Naghshineh and M. Schwartz, Distributed call admission control in mobile/wireless networks, IEEE J. Select. Areas Commun. 14(4) (1996) 711–717.

[11] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J.-P. Makela, R. Pichna, and J. Vallström, Handoff in hybrid mobile data networks, IEEE Pers. Commun. 87(8) (2000) 34–47.

[12] G. P. Pollini, Trends in handover design, IEEE Commun. Mag. (March 1996) 82–90.

[13] S. M. Ross, Introduction to probability and statistics for engineers and scientists, *Wiley Series in Probability and Mathematical Statistics* (1987).

[14] M. Sidi and D. Starobinski, New call blocking versus handoff blocking in cellular networks, in: *Proc. IEEE INFOCOM 1996*, (March 1996) vol. 1, pp. 35–42.

[15] A.K. Talukdar, B.R. Badrinath and A. Acharya, On accomodating mobile hosts in an integrated services packet network, in: *Proc. IEEE INFOCOM 97* (April 1997).

[16] J. Wroclawski, The use of RSVP with IETF integrated services, *RFC 2210* (Sept. 1997).

[17] M.M. Zonoozi and P. Dassanayake, User mobility modeling and characterization of mobility patterns, *IEEE J. Select. Areas Commun.* 15(7) (1997).



**Özgür B. Akan** received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University and Middle East Technical University, Ankara, Turkey, in 1999 and 2001, respectively. He received the Ph.D. degree in electrical and computer engineering from the Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, in 2004. He is currently an Assistant Professor with the Department of Electrical and Electronics Engineering, Middle East Technical University. His current research interests include sensor networks, next-generation wireless networks, and deep space communication networks.
E-mail: akan@eee.metu.edu.tr



**Buyurman Baykal** received his B.Sc. (High Hons.) degree in Electrical and Electronics Engineering from Middle East Technical University in 1990; M.Sc. (Distinction) and Ph.D. degrees in 1992 and 1995 from Imperial College of Science, Technology and Medicine. Dr. Baykal has research and teaching interests in speech processing, signal processing for telecommunications, and communication networks. He has extensive experience both in the theory and applications of adaptive signal processing techniques to communication applications such as acoustic echo cancellation, noise reduction, channel equalization and digital receiver design through self-conducted research and industry-funded research projects. He conducts research and implementation work on low bit rate speech coding and content based indexing of audio signals. He is also involved in communication network research with particular interest in ATM/IP design aspects, wireless networks and network management issues. Dr. Baykal is an Associate Editor of Computer Networks (Elsevier Science), Sensor Letters (American Scientific Publishing), a past Associate Editor of the IEEE Transactions on Circuits and Systems Part II—Analog and Digital Signal Processing (TCAS-II). He authored and co-authored over 50 technical papers.
E-mail: buyurman@metu.edu.tr