WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2005; **5**:867–885
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/wcm.352

Performance evaluation of wireless cellular networks under more realistic assumptions

Yuguang Fang*,†

Department of Electrical and Computer Engineering, University of Florida, 435 Engineering Building, P.O. Box 116130, Gainesville, Florida 32611-6130, U.S.A.

Summary

In wireless cellular networks, performance evaluation is an important part in modeling and designing effective schemes to utilize the limited resource. In the past, performance evaluation was carried out either under restricted assumption on some time variables such as exponential assumption or via simulations. In this paper, we present a survey on a new analytical approach we have developed in the last few years to evaluate the performance of wireless cellular networks under more realistic assumptions. In particular, we apply this approach to the analysis of call connection performance and mobility management under assumptions that many time variables such as call holding time, cell residence time, channel holding time, registration area (RA) residence time, and inter-service time are assumed to be generally distributed and show how we can obtain more general analytical results. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: wireless cellular networks; performance evaluation; call blocking probability; mobility management

1. Introduction

The future telecommunications networks (such as the 3G or B3G wireless networks) target to provide integrated services such as the voice, data, and multimedia via inexpensive low-powered mobile computing devices over the wireless environments [1,34,70,71]. The demand for multimedia services over the air has been steadily increasing over the years, and the wireless multimedia networking has been a very active research area. In order to support various integrated services with certain Quality of Service (QoS) requirements in these wireless networks, teletraffic analysis, resource provisioning study, and mobility management are necessary [34,35,38].

In wireless cellular networks, call connection premature termination (call dropping) is possible due to an unsuccessful handoff when the user moves out the current cell coverage while the target cell does not have resource to serve the call connection [14,21,48,51]. In order to reduce such forced call

*Correspondence to: Yuguang Fang, Department of Electrical Engineering, University of Florida, 435 Engineering Building, P.O. Box 116130, Gainesville, FL 32611-6130, U.S.A.
†E-mail: fang@ece.ufl.edu

Contract/grant sponsor: National Science Foundation Faculty Early Career Development Award; contract/grant number: ANI-0093241.
Contract/grant sponsor: Office of Naval Research Young Investigator Award; contract/grant number: N000140210464.

termination, call arrivals (new calls and handoff calls) have to be treated differently, which leads to the call admission control (CAC) and resource management in wireless cellular networks. Various priority-based CAC schemes have been proposed [26,45,73] in the literature, they can be classified into two broad categories: (1). *Guard Channel (GC) Schemes*: Some channels are reserved for handoff calls. Three different schemes can be identified. The first scheme, *the cutoff priority scheme*, is to reserve a portion of channel for handoff calls, whenever a channel is released, it is returned to the common poll of channels [38,61]. The second scheme, *the fractional GC schemes* is to admit a new call with certain probability (which depends on the number of busy channels). This scheme was first proposed by Ramjee *et al.* [68] and shown to be more general than the cutoff priority scheme. The last scheme is to divide all channels allocated to a cell into two groups, one is for the common use for all calls while the other is for handoff calls only (the rigid division-based CAC scheme [48]). (2). *Queueing Priority (QP) Schemes*: In this scheme, calls are accepted whenever there are free channels. When all channels are busy, either new calls are queued while handoff calls are blocked [36], or new calls are blocked while handoff calls are blocked [17,78], or all arriving calls are queued with certain rearrangements in the queue [14,51]. Various combinations of the above schemes are possible depending on specific applications [14,51]. Performance evaluation of these schemes are important in choosing the appropriate parameters. Moreover, under a specific CAC scheme, the call connection performance is also crucial to meet the design requirements. With the given call blocking probabilities, we need to find the call dropping probability (for the premature call termination), handoff probability, handoff rate, and actual call holding times. It is desirable that all these quantities can be computed under more general assumptions.

Another important problem in wireless mobile networks is mobility management. In order to effectively deliver a call connection service to a mobile user, the location of a called mobile user must be determined within a certain time limit (before the service is blocked). When the call connection is in progress while the mobile is moving, the network has to follow the mobile users and allocate enough resource to provide seamless continuing service without user awareness that in fact the network facility (such as base station) is changing. *Mobility management* is used to track the mobile users that move from place to place in the coverage of a wireless network or in the coverage of multiple communications networks working together to fulfill the grand vision of ubiquitous communications. Thus, *mobility management* is a key component for the effective operations of wireless networks to deliver wireless services (see Reference [1] and references therein). In wireless cellular systems, two operations are usually used to carry out the mobility management: *location update* and *terminal paging*. Location update is a process for a mobile user to inform the network where it is, while terminal paging is a process that the network attempts to locate a mobile user in the area it was last reported, which is called *uncertainty area*. Both processes will invoke signaling traffic in the signaling networks. The more frequent the location update, the higher the signaling traffic for location updates, while the smaller the uncertainty area, hence the lower the signaling traffic for terminal paging. On the other hand, the less frequent the location update, the lower the signaling traffic for the location updates, while the larger the uncertainty area, hence the higher the signaling traffic for terminal paging. Hence, there is a tradeoff between the location update and terminal paging. How to balance these two kinds of signaling traffic is important for the mobility management. To come up with an optimal design for mobility management, we need to quantify the total signaling traffic analytically so that optimization can be carried out.

To obtain analytical results for performance evaluation in wireless cellular networks, and even in general communications networks, time variables are usually assumed to be independent and exponentially distributed. For example, in the study of call blocking performance for the traditional cellular networks and PCS networks, the following assumptions are commonly used in order to obtain some analytical results: the interarrival time of cell traffic, the call holding time, and the channel holding time are all assumed to be exponentially distributed [21,37,38,79]. However, field data and simulation study showed that exponential assumption is not appropriate. In References [8,37,42–44], it has been shown that the channel holding time is not exponentially distributed for many wireless and cellular systems. In Reference [31], under certain assumptions, we showed that channel-holding time is exponentially distributed if and only if the cell residence time is exponentially distributed, where the cell residence time is the time a mobile user stays in a typical cell. The study for common-channel signaling (CCS) networks [11] demonstrated that the call holding time couldnot be accurately modeled by exponential distribution and showed that the mixed-type probability distribution model is much more appropriate. In Reference [67], the authors showed that the cell traffic is smooth (which implies that the interarrival-times for the cell traffic cannot be modeled by Poisson process). Thus, it is imperative for us to seek more general distribution models for the time variables needed to carry out more effective performance evaluation.

Distribution models, such as the exponential distribution, the lognormal distribution, the Erlang distribution, the (generalized) Gamma distribution, and Coxian distribution have been used to approximate the distributions of the channel holding times in the past [30–33,42,61,62,82]. It is well-known that exponential distribution can be used for one-parameter approximation of the measured data, while the Gamma distribution can be used for two-parameter approximation. Although the exponential and Erlang distribution models have simple good properties for queueing analysis, however, they are not general enough to fit the field data. The (generalized) Gamma and lognormal distributions are more general and has been shown to provide good fits to field data, however, application of these models will lead to the loss of the Markov property required in the queueing analysis [46]. Two new models were proposed for the mobility modeling for wireless cellular networks. One is the so-called Sum of Hyper-exponential (SOHYP) models [66], which has been used to model the channel holding time for cellular systems with mixed platforms and various mobility. The other model, called the mixed Erlang model (Hyper-Erlang model as we used in the past) [30], is used to model the cell residence time for wireless cellular networks. It has been demonstrated that the mixed-Erlang models and SOHYP models can approximate any distribution of nonnegative random variables [6,30,66]. In fact, both models belong to a more general class of distribution models called *phase-type distribution*, which has been extensively applied in performance evaluation in computer systems [6]. The popularity of phase-type distribution is because such distribution model leads to tractable queueing analysis using the quasi-birth-death (QBD) process [50]. Due to the generality of phase-type distribution, we can use it to model many time variables in the performance evaluation of wireless cellular networks. Fang *et al.* [22,30–33] developed a new unifying analytical approach to evaluating the call connection performance under more realistic distribution models for call holding time and cell residence time. Li *et al.* [4,5,54,55] proposed more general Markov correlated arrival process to characterize the new calls and handoff calls, and obtained many analytical performance results under general channel holding time distribution such as phase-type distribution. To capture the effect of link unreliability, Zhang *et al.* [80,81] and references therein generalized many aforementioned analytical results recently. The performance of hierarchical cellular systems under generally distributed call holding times and cell residence times were also carried out lately (see References [63,77]). On a significantly different research front, applying the mathematical technique we developed in the past (see Reference [22]), we also analyzed a few mobility management schemes [23,24].

In this paper, we present the analytical approach for performance evaluation of wireless cellular networks we have developed in the last few years, and demonstrate how a simple mathematical technique can be applied to obtain so many elegant analytical results for many performance metrics in the evaluation of wireless cellular networks under more realistic distribution models for the involved random time variables. The interesting observation is that many seemingly unrelated problems can be solved in a unified framework.

2. Mathematical Tools

In this section, we present some preliminary mathematical results which will be frequently used in this paper.

Many analytical results for the evaluation of wireless network performance are boiled down to the calculation of the following type of probability $Pr(\xi_1 \leq \xi_2)$ for random variables ξ_1 and ξ_2. This probability can then further be reduced to the computation of the following type of integral:

$$C = \frac{A}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} X(s)Y(-s)\mathrm{d}s \qquad (1)$$

where A is a constant, $X(s)$ and $Y(s)$ are analytic over the set $\mathcal{D}_\sigma = \{s \mid \Re(s) \geq \sigma\}$ in the complex plane ($\Re$ denotes the real part of a complex number). If $X(s)$ and $Y(s)$ are known, then the techniques used to find the inverse Laplace transforms [57] can be used to find C. In particular, if $X(s)$ and $Y(s)$ are rational functions in s, then the well-known Residue Theorem [57] can be applied to easily find C, in which case, the partial fractional expansion technique [46,57] can be used to obtain easily computable formula. Let σ_p denote the poles of $Y(-s)$ and let $\mathrm{Res}_{s=\mathrm{p}}$ denote the residue at pole $s = \mathrm{p}$, then we can easily show.

Lemma 1: *If $X(s)$ and $Y(s)$ are proper rational functions and σ_p is the set of poles of $Y(-s)$, then we have*

$$\begin{aligned} C &= \frac{A}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} X(s)Y(-s)\mathrm{d}s \\ &= -A\sum_{\mathrm{p}\in\sigma_{\mathrm{p}}}\mathop{\mathrm{Res}}_{s=\mathrm{p}}[X(s)Y(-s)] \end{aligned} \tag{2}$$

In the subsequent development, we often encounter the probability expression $Pr(X_0 + X_1 + X_2 + \cdots + X_k \le X)$, where X_0 and X are random variables whose probability density functions (pdf) have Laplace transforms $f^*_{X_0}(s)$ and $f^*_X(s)$, respectively, and $X_1, X_2, \ldots, X_k$ are random variables, independent and identically distributed (iid), whose pdf has a Laplace transform $f^*_{X_i}(s)$. The following result gives a method to compute the aforementioned probability [22].

Lemma 2: *Assume that the random variables X_0, $X_1, X_2, \ldots, X_k$, X are independent, and $X_1, X_2, \ldots, X_k$ are independent and identically distributed (iid), whose pdf has a Laplace transform $f^*_{X_i}(s)$. If $f^*_{X_0}(s)$, $f^*_{X_i}(s)$, and $f^*_X(s)$ are analytic in $\mathcal{D}_\sigma$ for a real number σ, then*

$$\begin{aligned} &Pr(X_0 + X_1 + X_2 + \cdots + X_k \le X) \\ &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f^*_{X_0}(s)[f^*_{X_i}(s)]^k}{s}f^*_X(-s)\mathrm{d}s \end{aligned} \tag{3}$$

Proof: Let $\xi = X_0 + X_1 + X_2 + \cdots + X_k$. Let $f_\xi(t)$ and $f^*_\xi(s)$ denote the pdf and the Laplace transform of ξ, respectively. From the independence of $X_0, X_1, X_2, \ldots, X_k$, we have

$$f^*_\xi(s) = E[\mathrm{e}^{-s\xi}] = E[\mathrm{e}^{-sX_0}]\prod_{i=1}^{k}E[\mathrm{e}^{-sX_i}] = f^*_{X_0}(s)(f^*_{X_i}(s))^k$$

So the pdf of ξ is given by

$$f_\xi(t) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} f^*_{X_0}(s)(f^*_{X_i}(s))^k\mathrm{e}^{st}\mathrm{d}s$$

Notice that the Laplace transform of $Pr(\xi \le t)$ (the distribution function) is $f^*_\xi(s)/s$. Thus, we have ($f_X(t)$ is the pdf of X)

$$\begin{aligned} &Pr(X_0 + X_1 + X_2 + \cdots + X_k \le X) \\ &= \int_0^\infty Pr(\xi \le X \mid X = t)f_X(t)\mathrm{d}t \\ &= \int_0^\infty \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f^*_{X_0}(s)[f^*_{X_i}(s)]^k}{s}\mathrm{e}^{st}\mathrm{d}s f_X(t)\mathrm{d}t \\ &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f^*_{X_0}(s)[f^*_{X_i}(s)]^k}{s}f^*_X(-s)\mathrm{d}s \end{aligned}$$

This completes the proof. □

The following probability is also very useful in performance evaluation:

$$\begin{aligned} &\beta(0) = Pr(X \le X_0), \\ &\beta(k) = Pr(X_0 + X_1 + X_2 + \cdots + X_{k-1} \le X \\ &\le X_0 + X_1 + X_2 + \cdots + X_k), k > 0 \end{aligned}$$

From Lemma 2, we can easily obtain

Lemma 3: *Assume that the random variables X_0, $X_1, X_2, \ldots, X_k, \ldots$, X are independent, and $X_1, X_2, \ldots, X_k, \ldots$ are independent and iid, whose pdf has a Laplace transform $f^*_{X_i}(s)$. If $f^*_{X_0}(s)$, $f^*_{X_i}(s)$, and $f^*_X(s)$ are analytic in $\mathcal{D}_\sigma$ for a real number σ, then*

$$\begin{aligned} \beta(0) &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{1 - f^*_{X_0}(s)}{s}f^*_X(-s)\mathrm{d}s \\ \beta(k) &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f^*_{X_0}(s)[1 - f^*_{X_i}(s)][f^*_{X_i}(s)]^{k-1}}{s}f^*_X(-s)\mathrm{d}s \end{aligned}$$

Let

$$S(n) = \sum_{k=1}^{n-1}\beta(k)$$

then, from Lemma 3, we can obtain

Lemma 4: *Under assumptions of Lemma 3, we have*

$$S(n) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f^*_{X_0}(s)[1 - (f^*_{X_i}(s))^{n-1}]}{s}f^*_X(-s)\mathrm{d}s$$

3. Performance Metrics for Call Connection

We first study the call connection performance to demonstrate the use of our technique. Before we do this, we present the notation we will use in this

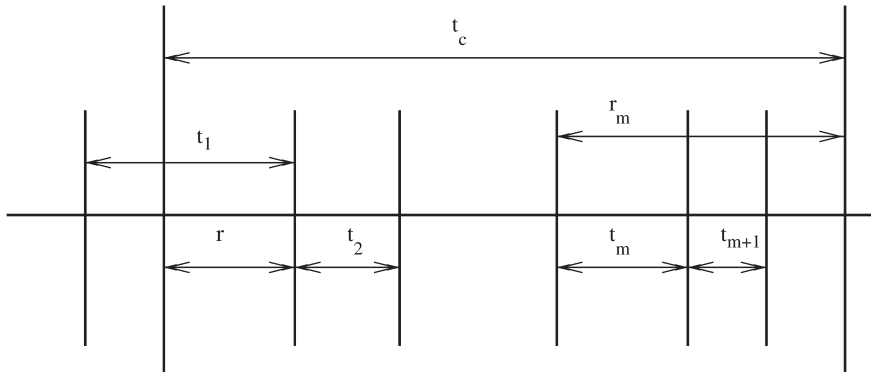


Fig. 1. The time diagram for call holding time and cell residence time.

section. In a wireless mobile network, a mobile user moves from cell to cell and engages call connection when he/she moves. Figure 1 shows the time diagram for a typical mobile user. Let t_c be the *call holding time* (the time of the requested connection to a wireless network, also known as *unencumbered call holding time*) for a typical new call with the mean value $1/\mu$. Let t_m denote the *cell residence time* in the mth cell a mobile user travels during a call life with the mean value $1/\eta$. Let r be the time between the time instant the new call is initiated at and the instant the new call moves out of the cell if the new call is not completed (we call it the *residual cell residence time*), let r_m $(m > 1)$ be the *residual call holding time* when the call finishes mth handoff successfully. Let λ and λ_h denote the arrival rates for new calls and handoff calls, respectively. Let $f_c(t)$, $f(t)$, and $f_r(t)$ denote, respectively, the probability density functions of t_c, t_m, and r with their corresponding Laplace transforms $f_c^*(s)$, $f^*(s)$, and $f_r^*(s)$, respectively. We assume throughout the paper that all distributions are nonlattice, that is, they do not contain the discrete singular components. In what follows, we will use $h^{(i)}(s)$ to denote the ith derivative of a function $h(s)$ at the point s. When $i = 0$, it gives the function itself.

In the current literature, most time variables here are assumed to be exponentially distributed [38,40,61,62,64], so that closed-form solution for call blocking performance can be obtained. Under certain assumptions, we observe that the channel holding time is exponentially distributed if and only if the cell residence time is exponentially distributed (see Reference [31] and reference therein). However, as we mentioned before, field study showed that channel holding time is not exponentially distributed, thus the above exponential assumption will not be valid in general, and demonstrated that many time variables can be determined by the characterization of the cell residence time, a time variable which can be used to characterize the mobility. In Reference [29], we propose to use the hyper-Erlang distribution (i.e., the mixed Erlang distribution) to model the cell residence time. In Reference [22], under the assumption that $r, t_c, t_2, t_3, \ldots$ are independent, and $t_2, t_3, \ldots$ are identically distributed with the pdf $f(t)$, we study a few performance metrics for call connection performance. Specifically, we derived analytical results for the aforementioned performance metrics: handoff probability, handoff rate, call dropping probability, and actual call holding times for complete calls and incomplete calls, which will be illustrated in this section. More details can be found in Reference [22].

3.1. Handoff Probability

Handoff probability is defined as the probability that a call connection needs at least one more handoff during its remaining lifetime. Depending on whether a call connection is a new call or a handoff call, we call the probability the *handoff probability for a new call* or the *handoff probability for a handoff call*.

We first study the handoff probability for a new call. Let P_n denote the handoff probability for a new call. We observe that a new call needs at least one handoff if and only if the call holding time t_c is greater than the residual cell residence time r, from Lemma 1 and Lemma 2 we obtain

$$P_n = Pr(r \leq t_c) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s} f_c^*(-s)\mathrm{d}s$$
$$= -\sum_{p\in\sigma_p} \operatorname*{Res}_{s=p} \frac{f_r^*(s)}{s} f_c^*(-s)$$

where σ_p is the set of poles of $f_c^*(-s)$.

Next we derive the handoff probability for a handoff call. Let $P_h(k)$ denote the probability that a handoff call connection needs at least one more handoff in its remaining lifetime after kth handoff. From the time diagram in Figure 1, we observe that this quantity can be expressed as the following conditional probability:

$$P_h(k) = Pr(r + t_2 + \cdots + t_k + t_{k+1} \leq t_c | r + t_2 + \cdots + t_k \leq t_c). \tag{4}$$

Since $r, t_2, \ldots, t_{k+1}, t_c$ satisfy all the assumptions in Lemma 1 and Lemma 2, we obtain

$$\begin{aligned} P_h(k) &= \frac{Pr(r + t_2 + \cdots + t_k + t_{k+1} \leq t_c)}{Pr(r + t_2 + \cdots + t_k \leq t_c)} \\ &= \frac{\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[f^*(s)]^k}{s} f_c^*(-s)\mathrm{d}s}{\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[f^*(s)]^{k-1}}{s} f_c^*(-s)\mathrm{d}s} \\ &= \frac{\sum_{p\in\sigma_p} \operatorname*{Res}_{s=p} \frac{f_r^*(s)[f^*(s)]^k}{s} f_c^*(-s)}{\sum_{p\in\sigma_p} \operatorname*{Res}_{s=p} \frac{f_r^*(s)[f^*(s)]^{k-1}}{s} f_c^*(-s)} \end{aligned}$$

3.2. Handoff Rate

Handoff rate, defined as the average number of handoffs undertaken during the actual call connection in the wireless cellular network, is a very important parameter for network design and traffic characterization. Nanda [64] and Lin *et al.* [61] obtained some analytic results for handoff rate for the case when the call holding time is exponentially distributed. There are no results for handoff rate when call-holding time is not exponentially distributed in the open literature except our prior works [22,29,31]. In this subsection, we present a formula for general cases where the call holding time and cell residence time are generally distributed.

Let H be the number of handoffs of a typical admitted call (either completed or forced to terminate) during the call connection. We first study the distribution of H under general call holding time and cell residence time distributions. Figure 1 shows the time diagram. We observe the following: $H = 0$ if and only if the call is not blocked and the call holding time t_c is shorter than the residual life r, that is, the call completes before the mobile moves out of the cell; $H = k$ (an admitted call experiences k handoffs during its call connection life) if and only if the call is either successfully having k handoffs and finishing its service in the following cell without any forced termination, or successfully having $(k-1)$ handoffs and then failed at the kth handoff. If the call blocking probability for a new call is p_o (call blocking probability) and the blocking probability for a handoff call (the handoff blocking probability) is p_f, which is also called forced termination probability [38], then we can easily obtain

$$\begin{aligned} Pr(H = 0) &= (1 - p_o)Pr(r \geq t_c) \\ Pr(H = 1) &= (1 - p_o)Pr(r < t_c \leq r + t_2)(1 - p_f) \\ &\quad + (1 - p_o)Pr(t_c \geq r)p_f \\ &\vdots \\ Pr(H = k) &= (1 - p_o)Pr(r + t_2 + \cdots + t_k < t_c \\ &\quad \leq r + t_2 + \cdots + t_{k+1})(1 - p_f)^k \\ &\quad + (1 - p_o)Pr(t_c \geq r + t_2 + \cdots + t_k) \\ &\quad \times (1 - p_f)^{k-1}p_f \\ &\vdots \end{aligned} \tag{5}$$

When $k > 0$, applying Lemma 2 in Equation (5), we obtain

$$Pr(H = k) = \frac{(1 - p_o)(1 - p_f)^{k-1}}{2\pi j} \times \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1 - (1 - p_f)f^*(s)][f^*(s)]^{k-1}}{s} f_c^*(-s)\mathrm{d}s$$

Thus, we can obtain the handoff rate as follows

$$\begin{aligned} E[H] &= \sum_{k=1}^{\infty} kPr(H = k) = \sum_{k=1}^{\infty} kPr(H = k) \\ &= \frac{1 - p_o}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s[1 - (1 - p_f)f^*(s)]} f_c^*(-s)\mathrm{d}s \end{aligned}$$

Applying the Residue Theorem, we obtain

$$\begin{aligned} E[H] &= \frac{1 - p_o}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s[1 - (1 - p_f)f^*(s)]} f_c^*(-s)\mathrm{d}s \\ &= -(1 - p_o)\sum_{p\in\sigma_p} \operatorname*{Res}_{s=p} \frac{f_r^*(s)}{s[1 - (1 - p_f)f^*(s)]} f_c^*(-s) \end{aligned}$$

where σ is a sufficiently small positive number.

The handoff call arrival rate, a very important quantity for call blocking analysis, can be computed from the handoff rate for the homogeneous wireless mobile networks. We observe that for each admitted new call, there will be on the average $E[H]$ number of handoff calls induced in the overall network, so the handoff call traffic will have arrival rate $\lambda_h = \lambda E[H]$. This is the major reason why we are interested in handoff rate.

3.3. Call Dropping Probability

Call dropping probability (call incompletion probability) is the probability that a call connection is prematurely terminated due to an unsuccessful handoff during the call life. Customers are more sensitive to call dropping than to call blocking at call initiation. Wireless service providers have to design the network to minimize the call dropping probability for customer care.

We observe that a call is dropped if there is no available channel in the targeted cell during a handoff, that is, a call is dropped when a handoff failure occurs during a call life. Figure 1 can also be used to illustrate the timing diagram for a forced terminated call. As before, let p_o and p_{f} denote the call blocking probability and handoff blocking probability, respectively. Let p_{c} denote the probability that a call is completed (without blocking and forced termination). Then the call dropping probability $p_{\mathrm{d}} = 1 - p_o - p_{\mathrm{c}}$ is given by

$$\begin{aligned} p_{\mathrm{d}} &= Pr(\text{the call is not blocked, it is forced} \\ &\quad \text{to terminatedue to an unsuccessful handoff}) \\ &= (1-p_o)\sum_{k=1}^{\infty}(1-p_{\mathrm{f}})^{k-1} p_{\mathrm{f}} Pr(t_{\mathrm{c}} \geq r + t_2 + \cdots + t_k) \\ &= \frac{(1-p_o)p_{\mathrm{f}}}{2\pi j}\sum_{k=1}^{\infty}(1-p_{\mathrm{f}})^{k-1} \\ &\quad \times \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[f^*(s)]^{k-1}}{s} f_{\mathrm{c}}^*(-s)\mathrm{d}s \\ &= \frac{(1-p_o)p_{\mathrm{f}}}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s[1-(1-p_{\mathrm{f}})f^*(s)]} f_{\mathrm{c}}^*(-s)\mathrm{d}s \end{aligned}$$

where we have again used Lemma 2. Thus, applying the Residue Theorem, we obtain

$$\begin{aligned} p_{\mathrm{d}} = 1 - p_o - p_{\mathrm{c}} &= \frac{(1-p_o)p_{\mathrm{f}}}{2\pi j} \\ &\quad \times \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s[1-(1-p_{\mathrm{f}})f^*(s)]} f_{\mathrm{c}}^*(-s)\mathrm{d}s \\ &= p_{\mathrm{f}} E[H] \end{aligned}$$

where σ is a sufficiently small positive number.

We observe that there is a close relationship between the call dropping probability and the handoff rate: $p_{\mathrm{d}} = p_{\mathrm{f}} E[H]$. This is very intuitive, recall that $E[H]$ is the average number of handoffs during the call life, each handoff has a probability p_{f} to fail, hence the call dropping probability is then the summation of handoff failure probability for each handoff! Thus, if we can compute the handoff rate, then we can easily obtain the call dropping probability. Notice that we have $\lambda_h = \lambda E[H]$, therefore, we have the following result: $p_{\mathrm{d}} = p_{\mathrm{f}} E[H] = \frac{\lambda_h}{\lambda} p_{\mathrm{f}}$. Let p_{c} denote the call completion probability [33], which is defined as the probability that an admitted call completes without forced termination. From Reference [33], we can easily obtain that $p_{\mathrm{c}} = 1 - p_o - p_{\mathrm{d}} = 1 - p_o - p_{\mathrm{f}} E[H]$.

3.4. Actual Call Holding Times

The actual call holding time (or actual call connection time) for a mobile user in a wireless cellular network is an important quantity, which can be used to determine the charging rate for the services the user subscribes to. However, in a wireless network or a mobile computing system, calls may be prematurely terminated due to the limitation of finite resource. The expected actual connection times for a complete and an incomplete call are called the *actual call holding times* for a complete call and an incomplete call, respectively.

We first consider the actual call holding time for an incomplete call. Let ξ denote the actual call connection time, let $G_{\mathrm{d}}(t)$ denote the cumulative distribution function for an incomplete call, then we have

$$\begin{aligned} G_{\mathrm{d}}(t) &= Pr(\xi \leq t | \text{the call is dropped}) \\ &= \frac{Pr(\xi \leq t, \text{the call is dropped})}{Pr(\text{the call is dropped})} \\ &= \frac{1}{p_{\mathrm{d}}}\sum_{k=1}^{\infty} Pr(\xi \leq t, \text{the call is dropped at the} \\ &\quad k\text{th handoff}) \\ &= \frac{1}{p_{\mathrm{d}}}\sum_{k=1}^{\infty}(1-p_o)(1-p_{\mathrm{f}})^{k-1}p_{\mathrm{f}} \\ &\quad \times Pr(r + t_2 + \cdots + t_k \leq t, r + t_2 + \cdots + t_k \leq t_{\mathrm{c}}) \\ &= \frac{(1-p_o)p_{\mathrm{f}}}{p_{\mathrm{d}}}\sum_{k=1}^{\infty}(1-p_{\mathrm{f}})^{k-1}\int_0^t f_k(\tau) Pr(t_{\mathrm{c}} \geq \tau)\mathrm{d}\tau \end{aligned} \quad (6)$$

where $f_k(t)$ is the probability density function of random variable $r + t_2 + \cdots + t_k$, and p_{d} is the call dropping probability. Let $f_k^*(s)$ denote the Laplace transform of $f_k(t)$, which is given by $f_k^*(s) = f_r^*(s)[f^*(s)]^{k-1}$. Let $g_{\rm d}(t)$ denote the density function of the actual call holding time for an incomplete call, then from Equation (6) we obtain

$$\begin{aligned} g_{\rm d}(t) &= \frac{(1-p_o)p_{\rm f}}{p_{\rm d}} \sum_{k=1}^{\infty} (1-p_{\rm f})^{k-1} f_k(t) Pr(t_{\rm c} \geq t) \\ &= \frac{(1-p_o)p_{\rm f}}{p_{\rm d}} \sum_{k=1}^{\infty} (1-p_{\rm f})^{k-1} f_k(t) \int_t^{\infty} f_{\rm c}(t_{\rm c}) {\rm d}t_{\rm c} \end{aligned} \tag{7}$$

Let $g_{\rm d}^*(z)$ denote the Laplace transform of $g_{\rm d}(t)$ and $T_{\rm d}$ the expected actual call holding time for an incomplete call. Applying Laplace transformation on both sides of Equation (7) and using a similar procedure to the one used in the proof of Lemma 2, we obtain

$$\begin{aligned} g_{\rm d}^*(z) &= \frac{(1-p_o)p_{\rm f}}{2p_{\rm d}\pi j} \\ &\quad \times \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{1-(1-p_{\rm f})f^*(s)} \cdot \frac{f_{\rm c}^*(-s+z)-1}{s-z} {\rm d}s \end{aligned} \tag{8}$$

where σ is a sufficiently small positive number.

The expected actual call holding time for an incomplete call $T_{\rm d}$ is given by

$$\begin{aligned} T_{\rm d} &= -\frac{(1-p_o)p_{\rm f}}{2p_{\rm d}\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{1-(1-p_{\rm f})f^*(s)} \\ &\quad \times \frac{s f_{\rm c}^{*(1)}(-s) + f_{\rm c}^*(-s) - 1}{s^2} {\rm d}s \end{aligned}$$

In summary, applying the Residue Theorem, we obtain

$$\begin{aligned} g_{\rm d}^*(z) &= \frac{(1-p_o)p_{\rm f}}{2p_{\rm d}\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{1-(1-p_{\rm f})f^*(s)} \\ &\quad \times \frac{f_{\rm c}^*(-s+z)-1}{s-z} {\rm d}s = -\frac{(1-p_o)p_{\rm f}}{p_{\rm d}} \\ &\quad \times \sum_{p\in\sigma_{\rm p}} \operatorname*{Res}_{s=z+p} \frac{f_r^*(s) f_{\rm c}^*(-s+z)}{(s-z)[1-(1-p_{\rm f})f^*(s)]}, \\ T_{\rm d} &= -\frac{(1-p_o)p_{\rm f}}{2p_{\rm d}\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} g_1(s) {\rm d}s \\ &= \frac{(1-p_o)p_{\rm f}}{p_{\rm d}} \sum_{p\in\sigma_{\rm p}} \operatorname*{Res}_{s={\rm p}} g_1(s) \end{aligned}$$

where

$$g_1(s) = \frac{f_r^*(s)}{1-(1-p_{\rm f})f^*(s)} \cdot \frac{s f_{\rm c}^{*(1)}(-s) + f_{\rm c}^*(-s) - 1}{s^2}$$

Next, we study the actual holding time for a complete call. Suppose that a call is completed when the mobile is in cell k'. Let ξ denote the actual call holding time. If $k'=1$, then $0 \leq t_{\rm c} \leq r$ and $\xi = t_{\rm c}$; while if $k' > 1$, then $r + t_2 + \cdots + t_{k'-1} \leq t_{\rm c} \leq r + t_2 + \cdots + t_{k'}$ and $\xi = t_{\rm c}$. Let $k = k'-1$, then we have

$$\begin{aligned} &\text{For } k = 0, 0 \leq t_{\rm c} \leq r \\ &\text{For } k > 0, r + t_2 + \cdots + t_k \\ &\qquad \leq t_{\rm c} \leq r + t_2 + \cdots + t_{k+1} \end{aligned}$$

Let $g_{\rm c}(t)$ denote the probability density function of the actual call holding time. Using an argument similar to the one for the actual call holding time for an incomplete call, we obtain

$$\begin{aligned} g_{\rm c}(t_{\rm c}) &= \left(\frac{1-p_o}{p_{\rm c}}\right) \left[f_{\rm c}(t_{\rm c}) \int_{t_{\rm c}}^{\infty} f_r(t_1) {\rm d}t_1 \right] + \left(\frac{1-p_o}{p_{\rm c}}\right) \\ &\quad \times \left[\sum_{k=1}^{\infty} (1-p_{\rm f})^k f_{\rm c}(t_{\rm c}) \int_0^{t_{\rm c}} f_k(t) \int_{t_{\rm c}-t}^{\infty} f(\tau) {\rm d}\tau {\rm d}t \right] \end{aligned}$$

Applying the similar technique we used for actual call holding time for an incomplete call, we can obtain

$$\begin{aligned} g_{\rm c}^*(z) &= \frac{1-p_o}{2\pi p_{\rm c} j} \int_{\sigma-j\infty}^{\sigma+j\infty} g_2(s) f_{\rm c}^*(-s+z) {\rm d}s \\ &= -\left(\frac{1-p_o}{p_{\rm c}}\right) \left\{ \sum_{p\in\sigma_{\rm p}} \operatorname*{Res}_{s=z+p} g_2(s) f_{\rm c}^*(-s) \right\} \\ T_{\rm c} &= -\frac{1-p_o}{2\pi p_{\rm c} j} \int_{\sigma-j\infty}^{\sigma+j\infty} g_2(s) f_{\rm c}^{*(1)}(-s) {\rm d}s \\ &= \left(\frac{1-p_o}{p_{\rm c}}\right) \left\{ \sum_{p\in\sigma_{\rm p}} \operatorname*{Res}_{s={\rm p}} g_2(s) f_{\rm c}^{*(1)}(-s) \right\} \end{aligned}$$

where σ is a sufficiently small positive number and

$$g_2(s) = \frac{[1-(1-p_{\rm f})f^*(s)] - p_{\rm f} f_r^*(s)}{s[1-(1-p_{\rm f})f^*(s)]}$$

4. Performance Analysis for Mobility Management Schemes

In this section, we demonstrate how our technique can be used to analyze mobility management schemes.

There are many location management schemes in the literature, Reference [1] presents a very comprehensive survey on all aspects of mobility management. In this section, we only concentrate on three schemes: movement-based mobility management (MB), pointer forwarding scheme (PFS), and two-location algorithm (TLA). Many other schemes can be analyzed in a similar fashion.

We use the terminology used in Reference [41]. An operation *move* means that a mobile user moves from one registration area (RA) (also called location area (LA)) to another while an operation *find* is the process to determine the RA a mobile user is currently visiting. The *move* and *find* in second generation location management schemes (such as in IS-41 or GSM MAP) are called *basic move* and *basic find*. In the *basic move* operation, a mobile detects if it is in a new RA. If it is, it will send a registration message to the new VLR, and the VLR will send a message to the HLR. The HLR will send a de-registration message to the old VLR, which will, upon receiving the de-registration message, send the cancellation confirmation message to the HLR. The HLR will also send a cancellation confirmation message to the new VLR. In the *basic find*, call to a mobile $\mathcal{T}$ is detected at a local switch. If the called party is in the same RA, the connection can be setup directly without querying the HLR. Otherwise, the local switch (VLR) queries the HLR for the callee, then HLR will query the callee's VLR. Upon receiving callee's location, the HLR will forward the location information to the caller's local switch, which will then establish the call connection with the callee. Periodic location update is optional for improving the efficiency of IS-41 and GSM MAP [19,20].

Signaling traffic in mobility management schemes incurs from the operations of *move* (may cause location update) and *find* (may need message exchanges and terminal paging). Most performance evaluation for *move* and *find* was carried out under the assumption that some of the time variables are exponentially distributed. The conclusions drawn from such results may not be extended to the cases when such an assumption is not valid, and the adaptive schemes for choosing certain parameters (such as the threshold for the number of pointers or the threshold in movement-based mobility management scheme) may not be appropriate accordingly. There are not much works handling the nonexponential situations. Recently, we have developed a new approach which handles the nonexponential situations [23,24]. In this section, we demonstrate this approach and present some analytical results for signaling cost for mobility management schemes under more relaxed general assumption. To do so, we need to model some of the time variables appropriately and compute the probability distribution of the number of RA boundary crossings or cell boundary crossings. For example, in the mobility-based mobility management, we need to find the average number of location updates during the inter-service time (the time between two consecutive served call connections for a typical mobile terminal), hence need to determine the distribution of cell boundary crossings. In other schemes, we need to find the probability distribution of RA crossings to determine the cost for location update and call delivery. For the simplicity of presentation, we use *an area* to denote either a RA or a cell so that we can use the results to determine the probability of the number of RA boundary crossings and the probability of the number of cell boundary crossings as long as we use the corresponding residence time.

Assume that the incoming calls to a mobile, say, $\mathcal{T}$, form a Poisson process. The time the mobile stays in an *area* is called the *area residence time* (*ART*). The time the mobile stays in a RA is called the *RA residence time* (*RRT*). The time the mobile stays in a cell is called the *cell residence time* (*CRT*). Thus, ART can be either RRT or CRT for later applications. We assume that the ART is generally distributed with a nonlattice distribution (i.e., the probability distribution does not have discrete component). The time between the end of a call served and the start of the following call served by the mobile is called *the inter-service time*. It is possible that a call arrives while the previously served call is still in progress [59]. In this case, the mobile $\mathcal{T}$ cannot accept the new call (the caller senses busy tone in this case and may hang up). Thus, the inter-arrival (inter-call) times for the calls terminated at the mobile $\mathcal{T}$ are different from the inter-service times. This phenomenon is called the *busy line* effect. We emphasize here that for mobility management, it is the inter-service times, not the inter-arrival times of calls terminating at the mobile, that affects the location update cost, because when a mobile is engaging a call connection, the wireless network knows where the mobile $\mathcal{T}$ is, hence the mobile does not need to carry out any location update during the call connection. Although the incoming calls form a Poisson process (i.e., the interarrival times are exponentially distributed), the inter-service times may not be exponentially distributed. Inter-service time bears the similarity to the inter-departure time of a queueing system with blocking, which has been shown to be nonexponential in general. In this paper, we assume that the service time is negligibly shorter than the inter-service time (i.e., ignoring the service time) and derive the probability $\alpha(K)$ that a mobile moves across K *areas* between two served calls arriving to the mobile $\mathcal{T}$. In this section, we assume that the inter-service times are generally distributed and derive an analytic expression for $\alpha(K)$.

Let $t_1, t_2, \ldots$ denote the area residence times and let r_1 denote the residual area residence time (i.e., the time interval between the time instant the mobile $\mathcal{T}$ registers to the network and the time instant the mobile exits the first *area*). Let t_c denote the inter-service time between two consecutive served calls to a mobile $\mathcal{T}$. Figure 2 shows the time diagram for K area boundary crossings. Suppose that the mobile is in an *area* R_1 when the previous call arrives and accepted by $\mathcal{T}$. It then moves K *areas* during the inter-service time, and $\mathcal{T}$ resides in the jth *area* for a period t_j $(1 \leq j \leq K+1)$. In this paper, we consider a homogeneous wireless mobile network, that is, all *areas* (either registration areas or cells) in the network are statistically identical, hence ARTs $t_1, t_2, \ldots$ are independent and iid with a general probability density function $f(t)$. Let t_c be generally distributed with probability density function $f_c(t)$, and let $f_r(t)$ be the probability density function of r_1. Let $f^*(s), f_c^*(s)$, and $f_r^*(s)$ denote the *Laplace-Stieltjes* (*L-S*) *transforms* (or simply *Laplace transforms*) of $f(t)$, $f_c(t)$, and $f_r(t)$, respectively. Let $E[t_c] = 1/\lambda_c$, and $E[t_i] = 1/\lambda_m$. It is obvious that the probability $\alpha(K)$ is given by

$$\alpha(0) = Pr[t_c \leq r_1], K = 0$$

$$\alpha(K) = Pr[r_1 + t_2 + \cdots + t_K < t_c \leq r_1 + t_2 + \cdots + t_{K+1}]$$

Thus, from Lemma 3, we can directly obtain

$$\alpha(0) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1 - f_r^*(s)}{s} f_c^*(-s)\mathrm{d}s$$

$$\alpha(K) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1 - f^*(s)][f^*(s)]^{K-1}}{s} f_c^*(-s)\mathrm{d}s$$

where $K > 0$ and σ is a sufficiently small positive number. Without any confusion, we will treat this result and Lemma 3 as the same result.

4.1. IS-41 Scheme

As the baseline study, we first briefly go over the IS-41 scheme (or GSM MAP), the de facto standards for wireless cellular systems. The wireless networks standards IS-41 [19] and GSM MAP [20] use two-level strategies for mobility management in that a two-tier system consisting of home location register (HLR) and visitor location register (VLR) databases is deployed (see Figure 3). Although there are some modifications on the mobility management for the third generation wireless systems (such as the introduction of *gateway location register* or *GLR*), the fundamental operations in mobility management remains more or less the same. Hence, we will use the IS-41 standard as the baseline study here. In two-tier mobility management architecture shown in Figure 3, the HLR stores the user profiles of its registered mobiles, which contain the user profile information such as the mobile's identification number, the type of services subscribed, the QoS requirements, and the current location information. The VLR stores the

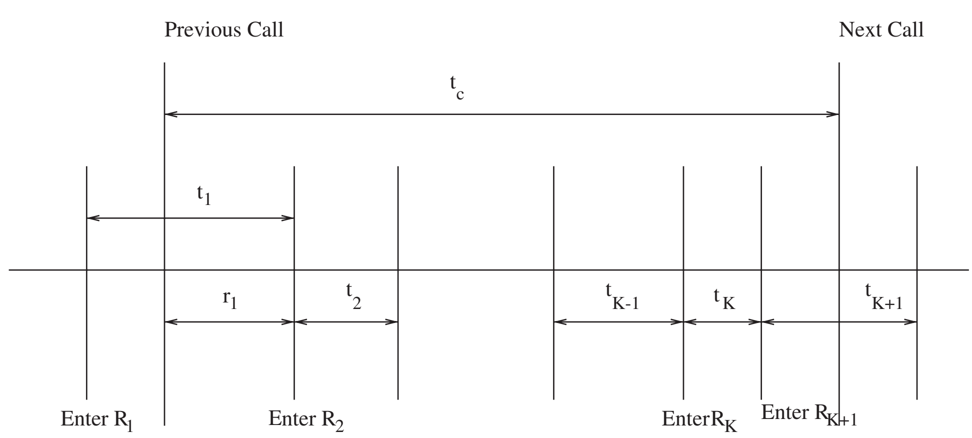


Fig. 2. The time diagram for K area boundary crossings [28].

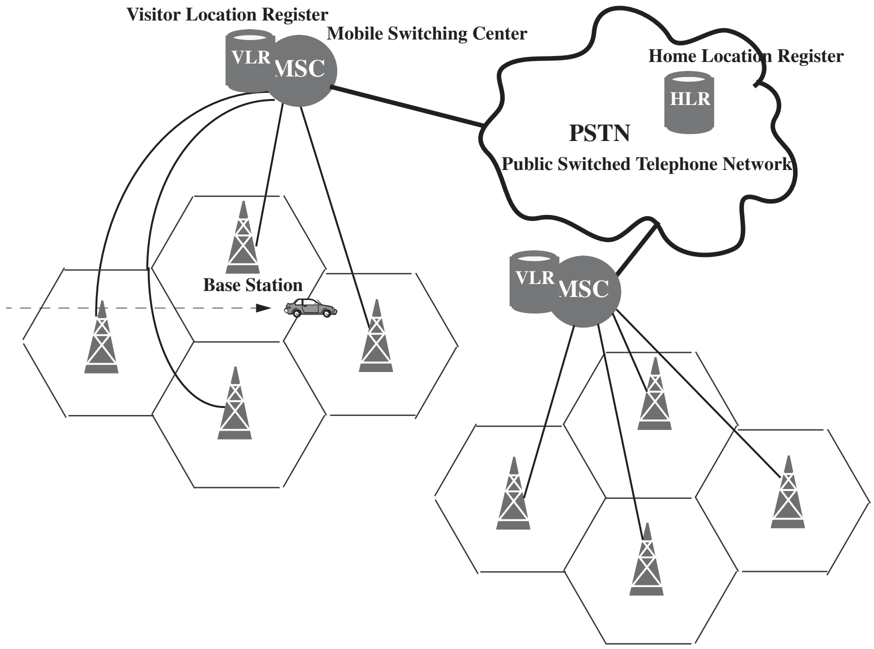


Fig. 3. Two-tier mobility management architecture [60].

replications of user profiles and the temporary identification number for those mobiles which are currently visiting the associated RA. There are two basic operations in mobility management: *registration* and *location tracking*. The registration (also called *location update* in many situations) is the process that a mobile informs the system of its current location. The location tracking (also called *call delivery*) is the process that the system locates its mobile in order to deliver a call service to the mobile. When a user subscribes a service to a wireless network, a record of the mobile user's profile will be created in the system's HLR. Whenever and wherever the mobile user travels in the system's coverage, the mobile's location will be reported to the HLR (registration) according to some strategies. Then the location in HLR will be used to locate (*find*) the mobile. When the mobile visits a new area (called *RA*), a temporary record will be created in the VLR of the visited system, and the VLR will then send a registration message to the HLR. All the signaling messages are exchanged over the overlaying signaling network using signaling system 7 (SS7) standard.

To evaluate this scheme, we use the same notation as before. t_c denotes the inter-service time with average $1/\lambda_c$ and t_i denotes the RA residence time with average $1/\lambda_m$. Let M and F denote the total cost for *basic moves* during the inter-service time and the total cost for *basic find*, respectively (i.e., the costs incurred in the IS-41 scheme). Since all location management schemes will go through the *move* and *find* whenever a terminating call to a mobile $\mathcal{T}$ arrives, the inter-service time forms the fundamental regenerative period for cost analysis, thus we only need to consider the signaling traffic incurred during this period. For the IS-41 scheme, whenever the mobile crosses a RA boundary, a registration is triggered. We assume that the unit cost for a basic registration (i.e., *basic move*) is m. M will be equal to the product of m and the average number of registrations incurred during the inter-service time that is given by $(1/\lambda_c)/(1/\lambda_m)$ from Little's Law, thus $M = m/\rho$ where $\rho = \lambda_c/\lambda_m$.

The *basic find* operation consists of two parts. The first part includes the interactions between the originating switch and the HLR while the second part includes the interactions between the HLR, the VLR, the MSC (terminating switch), and the mobile. Thus, $F \geq m$. The *basic find* will be the paging cost in one RA. Thus, we obtain the total cost for IS-41 during the inter-service time is

$$\mathcal{C}_{IS-41} = M + F = \frac{m}{\rho} + F \tag{9}$$

4.2. Movement-Based Mobility Management (MB)

There are many ways to improve the performance of the mobility management detailed in the IS-41. To minimize the signaling traffic due to the location update while keeping the location information fresh, we need to determine when we would need location update. In the current literature, three location update schemes were proposed and studied [2,10]: *distance-based location update*, *movement-based location update*, and *time-based location update*. In distance-based locate update scheme, location update will be performed when a mobile terminal moves d cells away from the cell in which the previous location update was performed, where d is a distance threshold. In the movement-based location update scheme, a mobile terminal will carry out a location update whenever the mobile terminal completes d movements between cells (whenever the mobile moves from one cell to another, we count it as one move), where d is the movement threshold. In the time-based location update scheme, the mobile terminal will update its location every d time units, where d is the time threshold. It has been shown [10] that the distance-based location update scheme gives the best result in terms of signaling traffic, however, it may not be practical because a mobile terminal has to know its own position information in the network topology. The time-based location update scheme is the simplest to implement, however, unnecessary signaling traffic may result (imagine a terminal stationary for a long period may not need to do any update before it moves). The movement-based location update scheme seems to be the best choice in terms of signaling traffic and implementation complexity. We now analyze the movement-based mobility management scheme here.

4.2.1. Average number of location updates

Let d be the threshold for the movement-based location update scheme, that is, a mobile will make a location update every d crossings of cell boundary when the mobile does not engage in service. Obviously, the average number of location updates during the inter-service time will determine the locate update cost. So, we first want to find this quantity. We assume that $f(t)$ denotes the probability density for the cell residence time (CRT) in this section. All other notations will remain the same as before. Then, the average number of location update during an inter-service time interval under movement-based location update scheme can be expressed as [3]

$$N_u(d) = \sum_{i=1}^{\infty} i \sum_{k=id}^{(i+1)d-1} \alpha(k) \tag{10}$$

where $\alpha(k)$ in this subsection denotes the probability that a mobile crosses k cells during the inter-service time. In what follows in this subsection, we present the computation for $N_u(d)$. Let

$$S(n) = \sum_{k=1}^{n-1} \alpha(k)$$

then, from Lemma 4, we obtain

$$S(n) = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-(f^*(s))^{n-1}]}{s} f_c^*(-s)\mathrm{d}s \tag{11}$$

Moreover, we have

$$\begin{aligned}\sum_{k=1}^{N} S(kd) &= \sum_{k=1}^{N} \frac{1}{2\pi j} \\ &\times \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-(f^*(s))^{kd-1}]}{s} f_c^*(-s)\mathrm{d}s \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)}{s} \\ &\times \left\{ N - \frac{(f^*(s))^{d-1}[1-(f^*(s))^{Nd}]}{1-(f^*(s))^d} \right\} f_c^*(-s)\mathrm{d}s\end{aligned} \tag{12}$$

Thus, from Equations (10), (11) and (12), after some mathematical manipulation (details can be found in Reference [23]), we obtain

$$\begin{aligned}N_u(d) &= \sum_{i=1}^{\infty} i[S((i+1)d) - S(id)] \\ &= \lim_{N\to\infty} \left\{ \sum_{i=1}^{N} i[S((i+1)d) - S(id)] \right\} \\ &= \lim_{N\to\infty} \left\{ NS((N+1)d) - \sum_{i=1}^{N} S(id) \right\} \\ &= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1-(f^*(s))^d]} f_c^*(-s)\mathrm{d}s \\ &= -\sum_{p\in\sigma_c} \operatorname*{Res}_{s=p} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1-(f^*(s))^d]} f_c^*(-s)\end{aligned} \tag{13}$$

where $\sigma > 0$ is a sufficiently small positive number and σ_c is the set of poles of $f_c^*(-s)$.

4.2.2. *Tradeoff cost analysis: a case study*

In this subsection, we present some results for the tradeoff analysis under movement-based location update scheme and some paging scheme with the linear cost functional for illustration purpose.

4.2.2.1. Location update cost. Let m denote the unit cost for location update (i.e., the cost for a *basic move* as in IS-41), then total location update cost is given by

$$\mathcal{C}_u(d) = -m \sum_{p \in \sigma_c} \operatorname*{Res}_{s=p} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1-(f^*(s))^d]} f_c^*(-s) \quad (14)$$

4.2.2.2. Paging cost. We consider the paging strategy used in Reference [52] for our case study. Considering the hexagonal layout for the wireless network, we assume that all cells are statistically identical. According to the movement-based location update scheme, a mobile terminal moves at most d cells away from the previous position where it performs the last location update. Thus, a mobile terminal will surely be located in a cell which is less than d cell away from the previously reported position. If we page in the circular area with d cells as radius and with the previously reported position as the center, then we can definitely find the mobile terminal. Thus, this paging scheme is the most conservative among all paging. If we let P denote the unit cost for each paging in a cell, the maximum paging cost for this paging scheme is given by Reference [52]

$$\mathcal{C}_p(d) = P(1 + 3d(d-1)) \quad (15)$$

4.2.2.3. Total cost. The unit costs (cost factors) m and P can be chosen to reflect the significance of the signaling (which may be significantly different from each other because they use different network resources). Given m, P, and the movement threshold d, the total cost for location update and paging will be given by

$$\begin{aligned}\mathcal{C}_{MB}(d) &= \mathcal{C}_u(d) + \mathcal{C}_p(d) \\ &= -m \sum_{p \in \sigma_c} \operatorname*{Res}_{s=p} \frac{f_r^*(s)(f^*(s))^{d-1}}{s[1-(f^*(s))^d]} f_c^*(-s) \\ &\quad + P(1 + 3d(d-1))\end{aligned} \quad (16)$$

We observe that as long as we find the probability distributions of cell residence time and inter-service time, we can find the total cost using Equation (16). Many numerical results [23] have shown that the cost function $\mathcal{C}_{MB}(d)$ is a convex function of d. If this is true, then we can find the unique optimal threshold d^* to minimize the cost.

4.3. Pointer Forwarding Scheme (PFS)

The PFS modifies the *it move* and *find* used in the IS-41 in the following fashion [41]. When a mobile $\mathcal{T}$ moves from one RA to another, it will inform its local switch (and VLR) at the new RA, which will then determine whether to invoke the *basic move* or the *forwarding move*. In the *forwarding move*, the new VLR exchanges messages with the old VLR to setup pointer from the old VLR to the new VLR, but does not involve the HLR. A subsequent call to the mobile $\mathcal{T}$ from some other switches will invoke the *forwarding find* procedure to locate the mobile: queries the mobile's HLR as in the *basic find*, and obtains a 'potentially outdated' pointer to the old VLR, which will then direct the *find* to the new VLR using the pointer to locate the mobile $\mathcal{T}$. To ensure that the time taken by the *forwarding find* is within the tolerable time limit, the length of the chain of the forwarding pointers must be limited. This can be done by setting up the threshold for chain length to be a number, say, K, that is whenever the mobile $\mathcal{T}$ crosses K RA boundaries, it will register itself through the *basic* move (i.e., basic registration with HLR). In this way, the signaling traffic between the mobile and HLR can be curbed potentially.

To evaluate the performance of this scheme, we need to quantify the signaling traffic incurring in this scheme. Let M' and F' denote the corresponding costs for the pointer forwarding scheme during the inter-service time, in which every K moves will trigger a new registration, where K is the maximum pointer chain length. Here, a move means the crossing from one RA to another. Let S denote the cost of setting up a forwarding pointer between VLRs during a pointer forwarding *move* and let T denote the cost of traversing a forwarding pointer between VLRs during a pointer forwarding *find*. We first derive M' and F'.

Suppose that a mobile $\mathcal{T}$ crosses i RA boundaries during the inter-service time, then there are $i - \lfloor i/K \rfloor$ pointer creations (every K moves require $K-1$ pointer creations) and the HLR is updated $\lfloor i/K \rfloor$ times (with pointer forwarding because the mobile $\mathcal{T}$ registers every Kth move). Here, we use $\lfloor x \rfloor$ to denote the floor function, that is, the largest integer not exceeding x. Let $f(t)$ and $f^*(s)$ denote the probability density function and its corresponding L-S transform of the RA residence time (RRT). Let $\alpha(k)$ denote the probability that the mobile crosses k RAs during the inter-service time. Then, we obtain

$$\begin{aligned} M' &= \sum_{i=0}^{\infty}\left[\left(i-\left\lfloor\frac{i}{K}\right\rfloor\right)S+\left\lfloor\frac{i}{K}\right\rfloor m\right]\alpha(i) \\ &= S\sum_{i=0}^{\infty} i\alpha(i)+(m-S)\sum_{i=0}^{\infty}\left\lfloor\frac{i}{K}\right\rfloor\alpha(i) \\ &= S\sum_{i=0}^{\infty} i\alpha(i)+(m-S)\sum_{r=0}^{\infty} r\left(\sum_{i=rK}^{(r+1)K-1}\alpha(i)\right) \end{aligned} \tag{17}$$

Let

$$\Sigma(K)=\sum_{r=1}^{\infty} r\left(\sum_{i=rK}^{(r+1)K-1}\alpha(i)\right)$$

$$S(n)=\sum_{k=1}^{n-1}\alpha(k)$$

then, from Lemma 4, we obtain

$$S(n)=\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_r^*(s)[1-(f^*(s))^{n-1}]}{s}f_c^*(-s)\mathrm{d}s$$

$$\sum_{i=1}^{N} s(iK)=\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_r^*(s)}{s} \times\left\{N-\frac{(f^*(s))^{K-1}[1-(f^*(s))^{NK}]}{1-(f^*(s))^K}\right\}f_c^*(-s)\mathrm{d}s$$

and

$$\begin{aligned} \Sigma(K) &= \sum_{r=1}^{\infty} r[S((r+1)K)-S(rK)] \\ &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_r^*(s)(f^*(s))^{K-1}}{s[1-(f^*(s))^K]}f_c^*(-s)\mathrm{d}s \end{aligned} \tag{18}$$

Noticing that $\Sigma(1)=\sum_{i=0}^{\infty} i\alpha(i)$, from Equation (17), we obtain

$$\begin{aligned} M' &= S\Sigma(1)+(m-S)\Sigma(K) \\ &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\left[\frac{Sf_r^*(s)}{s[1-f^*(s)]} +\frac{(m-S)f_r^*(s)(f^*(s))^{K-1}}{s[1-(f^*(s))^K]}\right]f_c^*(-s)\mathrm{d}s \end{aligned}$$

Next, we derive F'. After the last *basic move* operation (if any), the mobile $\mathcal{T}$ crosses $n=i-K\lfloor i/K\rfloor$ RA boundaries. Let $\Theta(n)$ denote the number of pointers to be tracked in order to find the mobile $\mathcal{T}$ in the pointer forwarding *find* operation. If the mobile visits a RA more than once (i.e., a 'loop' exists among n moves), then $\Theta(n)$ may not need to trace n pointers, thus, $\Theta(n)\le n$. From this argument and applying Lemma 3, we obtain

$$\begin{aligned} F' &= \sum_{i=0}^{\infty} T\Theta\left(i-K\left\lfloor\frac{i}{K}\right\rfloor\right)\alpha(i)+F \\ &= T\sum_{r=0}^{\infty}\sum_{k=0}^{K-1}\Theta(k)\alpha(rK+k)+F \\ &= T\sum_{k=0}^{K-1}\Theta(k)\left(\sum_{r=0}^{\infty}\alpha(rK+k)\right)+F \\ &= T\sum_{k=0}^{K-1}\Theta(k)\frac{1}{2\pi j} \\ &\times\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_r^*(s)[1-f^*(s)](f^*(s))^{k-1}}{s[1-(f^*(s))^K]}f_c^*(-s)\mathrm{d}s+F \\ &= \frac{T}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_r^*(s)[1-f^*(s)]}{s[1-(f^*(s))^K]} \\ &\times\left(\sum_{k=0}^{K-1}\Theta(k)(f^*(s))^{k-1}\right)f_c^*(-s)\mathrm{d}s+F \end{aligned}$$

Applying the Residue Theorem [57], we finally arrive at

$$\begin{aligned} M' &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} h_1(s)f_c^*(-s)\mathrm{d}s \\ &= -\sum_{p\in\sigma_c}\operatorname*{Res}_{s=p} h_1(s)f_c^*(-s) \\ F' &= F+\frac{T}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} h_2(s)f_c^*(-s)\mathrm{d}s \\ &= F-\sum_{p\in\sigma_c}\operatorname*{Res}_{s=p} h_2(s)f_c^*(-s) \end{aligned}$$

where σ_c denotes the set of poles of $f_c^*(-s)$ and

$$h_1(s)=\left[\frac{Sf_r^*(s)}{s[1-f^*(s)]}+\frac{(m-S)f_r^*(s)(f^*(s))^{K-1}}{s[1-(f^*(s))^K]}\right]$$

$$h_2(s)=\frac{f_r^*(s)[1-f^*(s)]}{s[1-(f^*(s))^K]}\left(\sum_{k=0}^{K-1}\Theta(k)(f^*(s))^{k-1}\right)$$

The worst case for F' would be when all pointers are traced, that is, when $\Theta(n) = n$. In this case, we have the following result:

$$F' = F + \frac{T}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_r^*(s)[1-f^*(s)]}{s[1-(f^*(s))^K]} \times \left(\sum_{k=0}^{K-1} k(f^*(s))^{k-1}\right) f_c^*(-s)\mathrm{d}s = F - T\sum_{p\in\sigma_c} \operatorname*{Res}_{s=p} h_3(s) f_c^*(-s)$$

where

$$h_3(s) = \frac{f_r^*(s)[1-K(f^*(s))^{K-1} + (K-1)(f^*(s))^K}{s[1-(f^*(s))^K][1-f^*(s)]}$$

Thus, the total cost for PFS during the inter-service time can be computed as follows:

$$\mathcal{C}_{\mathrm{PFS}} = M' + F' = F + \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} g(s) f_c^*(-s)\mathrm{d}s = F - \sum_{p\in\sigma_c} \operatorname*{Res}_{s=p} g(s) f_c^*(-s) \tag{19}$$

where

$$g(s) = \frac{Sf_r^*(s)}{s[1-f^*(s)]} + \frac{(m-S)f_r^*(s)(f^*(s))^{K-1}}{s[1-(f^*(s))^K]} + \frac{Tf_r^*(s)[1-f^*(s)]}{s[1-(f^*(s))^K]}\left(\sum_{k=0}^{K-1}\Theta(k)\right)$$

4.4. Two Location Algorithm (TLA)

In the TLA [59], a mobile $\mathcal{T}$ has a small built-in memory to store the addresses for the two most recently visited RAs. The record of $\mathcal{T}$ in the HLR also has an extra field to store the two corresponding locations. The first memory location stores the most recently visited RA. The TLA guarantees that the mobile $\mathcal{T}$ is in one of the two locations. When the mobile $\mathcal{T}$ joins the network, it registers with the network and stores the location in its memory and updates the HLR with its location. When the mobile $\mathcal{T}$ moves to another registration area, it checks whether the RA is in the memory or not. If the new RA is not in the memory, the most recently visited (MRV) RA in the memory is moved to the second memory entry while the new RA is stored in the MRV position in the memory of $\mathcal{T}$. At the same time, a registration operation is performed to make the same modification in the HLR record. If the address for the new RA is already in the memory, swaps the two locations in the memory of $\mathcal{T}$ and no registration is needed and no action is taken in HLR record. Thus, in TLA, no registration is performed when a mobile moves back and forth in two locations. The consequence is that the location entries in the mobile and HLR may not be exactly the same: the MRV RA in HLR may not be the MRV RA in reality!

When a call arrives for the mobile $\mathcal{T}$, the two addresses are used to find the actual location of the mobile $\mathcal{T}$. The order of the addresses used to locate $\mathcal{T}$ will affect the performance of the algorithm. If $\mathcal{T}$ is located in the first try (i.e., a *location hit*), then the *find* cost is the same as the one in IS-41 scheme. Otherwise, the second try (due to the *location miss* in the first try) will find $\mathcal{T}$, which incurs additional cost. Thus, the probability that the HLR has a location miss for a call setup, say, w, is important to capture the signaling traffic. In Reference [59], this probability is derived under the assumption that the inter-service time is exponential. In Reference [24], we obtain more general analytical results under the assumption that the inter-service time is generally distributed and carry out the cost analysis for the TLA, which we will elaborate next.

From the argument in Reference [59], $(1-\omega)$ is the probability that the HLR has the correct view of the latest visited RA when a call arrives (i.e., a location hit occurs and the *find* cost for TLA is the same as that for IS-41). A location hit occurs either when the mobile has not moved since last served call arrival, or when the last location update is followed by an even number of moves during the inter-service time, or when there are an even number of moves with no location update during the inter-service time. Here, in this subsection, a move means a boundary crossing from one RA to another. Let w_1 denote the probability that there is no move during the inter-service time, let w_2 denote the probability that the last served call is followed by an even number of moves without location update during the inter-service time, and let w_3 denote the probability that there are an even number of moves without location update during the inter-service time. Then, we have $1-w = w_1 + w_2 + w_3$, that is, $w = 1 - w_1 - w_2 - w_3$. Let $f(t)$ denote the probability density function for the RRT and other related notations remain the same as before. Let $\alpha(k)$ denote again the probability that a mobile user crosses k RA boundaries during the inter-service time. For w_1, we have

$$w_1 = Pr(r_1 \leq t_c) = \alpha(0) \tag{20}$$

Let θ denote the probability of a move without location update. Let $w_2(K)$ denote the probability that the last registration followed by an even number of moves without location update during the inter-service time given that there are K moves (RA crossings) during the inter-service time. We can easily obtain that [59]

$$\begin{aligned} w_2(K) &= (1-\theta)\sum_{i=0}^{\lfloor (K-1)/2 \rfloor} \theta^{2i} \\ &= \frac{1-\theta^{2\lfloor (K-1)/2 \rfloor + 2}}{1+\theta}, K > 0 \end{aligned}$$

where $\lfloor x \rfloor$ indicates the floor function, that is, the largest integer not exceeding x. Noticing that $w_2(2i+2) = w_2(2i+1)$ for $i \geq 0$, we have

$$\begin{aligned} w_2 &= \sum_{K=1}^{\infty} w_2(K)\alpha(K) \\ &= \sum_{i=0}^{\infty} w_2(2i+1)[\alpha(2i+1) + \alpha(2i+2)] \\ &= \frac{1}{1+\theta}\sum_{i=1}^{\infty}(1-\theta^{2i})[\alpha(2i-1) + \alpha(2i)] \end{aligned} \tag{21}$$

The probability w_3 can be computed as follows:

$$w_3 = \sum_{i=1}^{\infty} \theta^{2i}\alpha(2i)$$

Thus, applying Lemma 3 and after some mathematical manipulations similar to the techniques we used for PFS (the details can be found in Reference [24], we obtain

$$\begin{aligned} w &= 1 - w_1 - w_2 - w_3 = 1 - \alpha(0) - \frac{1}{1+\theta} \\ &\quad \times (1-\alpha(0)) + \frac{1}{1+\theta}\sum_{i=1}^{\infty}\theta^{2i} - \frac{\theta}{1+\theta}\sum_{i=1}^{\infty}\theta^{2i}\alpha(2i) \\ &= \frac{\theta}{1+\theta} - \frac{\theta}{(1+\theta)2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{1-f_r^*(s)}{s}f_c^*(-s)\mathrm{d}s \\ &\quad + \frac{\theta^2}{(1+\theta)2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_r^*(s)[1-f^*(s)]}{s[1-\theta^2 f^{*2}(s)]}f_c^*(-s)\mathrm{d}s \\ &\quad - \frac{\theta^3}{(1+\theta)2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_r^*(s)[1-f^*(s)]f^*(s)}{s[1-\theta^2 f^{*2}(s)]}f_c^*(-s)\mathrm{d}s \\ &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{\theta f_r^*(s)}{s[1+\theta f^*(s)]}f_c^*(-s)\mathrm{d}s \end{aligned} \tag{22}$$

Now, we are ready to carry out the cost analysis for TLA. We still assume that the cost for *basic move* is m and the cost for the *basic find* is F. The *basic find* consists of two parts: the first part is the message exchange from a mobile to the HLR, which is more or less the cost for the *basic move*; the second part is the message forwarding from the HLR to the callee plus the possible terminal paging. Hence, usually we have $F \geq m$ and the cost for the second part in the *basic find* is $F - m$. Suppose that the mobile moves across K RAs during the inter-service time. The conditional probability $Pr[I = i|K]$ that i *location update* operations are performed among the K moves has a Bernoulli distribution:

$$Pr[I = i \,|\, K] = \binom{K}{i}\theta^{K-i}(1-\theta)^i$$

where θ is the probability that when a mobile $\mathcal{T}$ moves, the new RA address is in the mobile's memory. Then the average number of location update during the inter-service time for TLA is given by

$$\begin{aligned} n_{\mathrm{TLA}} &= \sum_{K=0}^{\infty}\sum_{i=0}^{K} iPr[I = i \,|\, K]\alpha(K) \\ &= \sum_{K=0}^{\infty}\left(\sum_{i=0}^{K} i\binom{K}{i}\theta^{K-i}\theta^i\right)\alpha(K) \\ &= \sum_{K=0}^{\infty} K(1-\theta)\alpha(K) = (1-\theta)\Sigma(1) \\ &= \frac{1-\theta}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_r^*(s)}{s[1-f^*(s)]}f_c^*(-s)\mathrm{d}s \\ &= -(1-\theta)\sum_{p\in\sigma_c}\operatorname*{Res}_{s=p}\frac{f_r^*(s)}{s[1-f^*(s)]}f_c^*(-s) \end{aligned} \tag{23}$$

where we have used Equation (18). Thus, the total cost for registration during inter-service time is $c_1 = mn_{\mathrm{TLA}}$.

For the operations of the second part in the *basic find* for TLA, if we have a location hit (i.e., the location entries in the memories of both HLR and the mobile are identical), the *find* cost will be the same as in *basic find*; if there is a location miss, extra cost from HLR to the VLR (second part of the *basic find* operation) will incur, thus the total cost for the *find operation* in TLA will be

$$c_2 = (1-\omega)F + \omega[F + (F-m)] = F + (F-m)\omega$$

Thus, the total signaling cost during the inter-service time for TLA is given by

$$\begin{aligned} \mathcal{C}_{\text{TLA}} &= c_1 + c_2 = \delta n_{TLA} + F + (F - m)\omega \\ &= F + \frac{1}{2\pi j}\int_{\sigma - j\infty}^{\sigma + j\infty} q(s) f_{\text{c}}^*(-s)\text{d}s \\ &= F - \sum_{p \in \sigma_{\text{c}}} \operatorname*{Res}_{s=\text{p}} q(s) f_{\text{c}}^*(-s) \end{aligned} \quad (24)$$

where σ_{c} is the set of poles of $f_{\text{c}}^*(-s)$ and

$$q(s) = \left[\frac{(1-\theta)m}{s[1 - f^*(s)]} + \frac{(F-m)\theta}{s[1 + \theta f^*(s)]}\right] f_r^*(s)$$

5. Conclusions

In this paper, we discuss a newly developed analytical method for the performance evaluation of wireless cellular networks. Under very general assumption on distributions of time variables, we can derive analytical formulae for many performance metrics for call connection performance and mobility management schemes. We hope this survey will enrich the literature of performance evaluation and provide some analytical tools for the design of future wireless networks.

Acknowledgement

This work was supported in part by National Science Foundation Faculty Early Career Development Award under grant ANI-0093241 and the Office of Naval Research Young Investigator Award under grant N000140210464.

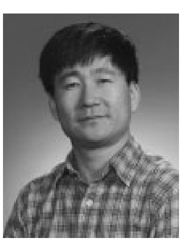
References

1. Akyildiz IF, McNair J, Ho JSM, Uzunalioglu H, Wang W. Mobility management in next-generation wireless systems. In *Proceedings of the IEEE*, August 1999; **87**(8); pp. 1347–1384.
2. Akyildiz IF, Ho JSM. Dynamic mobile user location update for wireless PCS networks. *ACM Wireless Networks* 1995; **1**: 187–196.
3. Akyildiz IF, Ho JSM, Lin Y-B. Movement-based location update and selective paging for PCS networks. *IEEE/ACM Transactions on Networking* 1996; **4**(4): 629–638.
4. Alfa AS, Li W. PCS networks with correlated arrival process and retrial phenomenon. *IEEE Transactions on Wireless Communications* 2002; **1**(4): 630–637.
5. Alfa AS, Li W. A homogeneous PCS network with Markov call arrival process and phase type cell residence time. *ACM Wireless Networks* 2002; **8**(6): 597–605.
6. Asmussen S. *Applied Probability and Queues*. John Wiley & Sons: New York, 1987.
7. Baker GA Jr., Graves-Morris P. *Padé Approximants* (2nd edn). Cambridge University Press, UK 1996.
8. Barcelo F, Bueno S. Idle and inter-arrival time statistics in public access mobile radio (PAMR) systems. *Proceedings of IEEE Globecom'97*, Phoenix, AZ, November 1997.
9. Barcelo F, Jordan J. Channel holding time distribution in cellular telephony. *The 9th International Conference on Wireless Communications (Wireless'97)*, Vol. 1, Alberta, Canada, 9–11 July, 1997; pp. 125–134.
10. Bar-Noy A, Kessler I, Sidi M. Mobile users: to update or not update? *ACM Wireless Networks* 1994; **1**(2): 175–186.
11. Bolotin VA. Modeling call holding time distributions for CCS network design and performance analysis. *IEEE Journal on Selected Areas in Communications* 1994; **12**(3): 433–438.
12. Bria A, Gessler F, Queseth O, Stridh R, Unbehaun M, Wu J, Zander J. 4th-generation wireless infrastructures: scenarios and research challenges. *IEEE Personal Communications* 2001; **8**: 25–31.
13. Camp T, Boleng J, Davis V. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing* 2002; **2**: 483–502.
14. Chang C, Chang CJ, Lo KR. Analysis of a hierarchical cellular system with reneging and dropping for waiting new calls and handoff calls. *IEEE Transactions on Vehicular Technology* 1999; **48**(4): 1080–1091.
15. Chlebus E, Ludwin W. Is handoff traffic really Poissonian? *IEEE ICUPC'95*, Tokyo, Japan, 6–10 November, 1995; pp. 348–353.
16. Cox DC. Wireless personal communications: what is it? *IEEE Personal Communications Magzine* 1995; **2**:20–35.
17. Del Re E, Fantacci R, Giambene G. Handover queueing strategies with dynamic and fixed channel allocation techniques in low earth orbit mobile satellite systems. *IEEE Transactions on Communications* 1999; **47**(1): 89–102.
18. Del Re E, Fantacci R, Giambene G. Efficient dynamic channel allocation techniques with handover queueing for mobile satellite networks. *IEEE Journal on Selected Areas in Communications* 1995; **13**(2): 397–405.
19. EIA/TIA. Cellular radio-telecommunications intersystem operations, *EIA/TIA Technical Report IS-41 Revision B*, 1991.
20. ETSI, *Digital cellular telecommunications system (phase 2+): mobile application part (MAP) specification (GSM 09.02 version 7.51 Release)*, 1998.
21. Everitt DE. Traffic engineering of the radio interface for cellular mobile networks. *Proceedings of the IEEE* 1994; **82**(9); pp. 1371–1382.
22. Fang Y. Modeling and performance analysis for wireless mobile networks: a new analytical approach. Accepted for publication in *IEEE/ACM Transactions on Networking*, in press.
23. Fang Y. Movement-based location management and tradeoff analysis for wireless mobile networks. *IEEE Transactions on Computers* (Special Issue on Wireless Internet) 2003; **52**(6): 791–803.
24. Fang Y. General modeling and performance analysis for location management in wireless mobile networks. *IEEE Transactions on Computers* (Special Issue on Data Management Systems and Mobile Computing) 2002; **51**(10): 1169–1181.
25. Fang Y. Hyper-Erlang distribution model and its applications in wireless mobile networks. *ACM Wireless Networks* 2001; **7**(3): 211–219.
26. Chen X, Liu W, Fang Y. Resource allocation and call admission control in mobile wireless networks. In *Algorithms and Protocols for Wireless and Mobile Networks*, Boukerche A (ed.). CRC Press, Florida, USA; 2005.

27. Fang Y, Zhang Y. Call admission control schemes and performance analysis in wireless mobile networks. *IEEE Transactions on Vehicular Technology* 2002; **51**(2): 371–384.
28. Fang Y, Chlamtac I, Lin YB. Portable movement modeling for PCS networks *IEEE Transactions on Vehicular Technology* 2000; **49**(4): 1356–1363.
29. Fang Y, Chlamtac I. Teletraffic analysis and mobility modeling for PCS networks. *IEEE Transactions on Communications* 1999; **47**(7): 1062–1072.
30. Fang Y, Chlamtac I. A new mobility model and its application in the channel holding time characterization in PCS networks. In Proceeding of INFOCOM'99. New York, New York, March 1999.
31. Fang Y, Chlamtac I, Lin YB. Channel occupancy times and hand-off rate for mobile computing and PCS networks. *IEEE Transactions on Computers* 1998; **47**(6): 679–692.
32. Fang Y, Chlamtac I, Lin YB. Modeling PCS networks under general call holding times and cell residence time distributions. *IEEE Transactions on Networking* 1997; **5**(6): 893–906.
33. Fang Y, Chlamtac I, Lin YB. Call performance for a PCS network. *IEEE Journal on Selected Areas in Communications* 1997; **15**(8): 1568–1581.
34. Frodigh M, Parkvall S, Roobol C, Johansson P, Larsson P. Future-generation wireless networks. *IEEE Personal Communications* 2001; 10–17.
35. Grillo D, Skoog RA, Chia S, Leung KK. Teletraffic engineering for mobile personal communications in ITU-T work: the need to match practice and theory. *IEEE Personal Communications* 1998; **5**: 38–58.
36. Guerin RA. Queueing-blocking system with two arrival streams and guard channels. *IEEE Transactions on Communications* 1988; **36**(2): 153–163.
37. Guerin RA. Channel occupancy time distribution in a cellular radio system. *IEEE Transactions on Vehicular Technology* 1987; **35**(3): 89–99.
38. Hong D, Rappaport SS. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Transactions on Vehicular Technology* 1986; **35**(3): 77–92.
39. Hou J, Fang Y. Mobility-based call admission control schemes for wireless mobile networks. *Wiley International Journal of Wireless Systems and Mobile Computing* 2001; **1**(3): 269–282.
40. Jabbari B. Teletraffic aspects of evolving and next-generation wireless communication networks. *IEEE Communications Magzine* 1996; 4–9.
41. Jain R, Lin YB. An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS. *ACM Wireless Networks* 1995; **1**(2): 197–210.
42. Jedrzycki C, Leung VCM. Probability distributions of channel holding time in cellular telephony systems. In *Proceedings IEEE Vehicular Technology Conference*, May 1996, Atlanta; pp. 247–251.
43. Jordan J, Barcelo F. Statistical modelling of channel occupancy in trunked PAMR systems. *The 15th International Teletraffic Conference (ITC'15)*, Ramaswami V, Wirth PE (eds), Elsevier Science B.V., 1997; pp. 1169–1178.
44. Jordan J, Barcelo F. Statistical modeling of transmission holding time in PAMR systems. In *Proceedings of IEEE Globecom'97*, Phoenix, AZ, November 1997.
45. Katzela I, Naghshineh M. Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey. *IEEE Personal Communications* 1996; **3**(3): 10–31.
46. Kleinrock L. *Queueing Systems: Theory*, Vol. I. John Wiley & Sons: New York, 1975.
47. Kobayashi H. *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology.* Addison-Wesley Publishing Company, 1978.
48. Kulavaratharasah MD, Aghvami AH. Teletraffic performance evaluation of microcellular personal communication networks (PCN's) with prioritized handoff procedures. *IEEE Transactions on Vehicular Technology* 1999; **48**(1): 137–152.
49. Lai WK, Jin YJ, Chen HW, Pan CY. Channel assignment for initial and handoff calls to improve the call completion probability. *IEEE Transactions on Vehicular Technology* 2003; **52**(4): 876–890.
50. Latouche G, Ramaswami V. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM Publishing: Philadelphia, 1999.
51. Lau VKN, Maric SV. Mobility of queued call requests of a new call-queueing technique for cellular systems. *IEEE Transactions on Vehicular Technology* 1998; **47**(2): 480–488.
52. Li J, Kameda H, Li K. Optimal dynamic mobility management for PCS networks. *IEEE/ACM Transactions on Networking* 2000; **8**(3): 319–327.
53. Li W, Alfa AS. Channel reservation for handoff calls in a PCS network. *IEEE Transactions on Vehicular Technology* 2000; **49**(1): 95–104.
54. Li W, Alfa AS. A PCS network with correlated arrival process and splitted-rating channels. *IEEE Journal on Selected Areas in Communications* 1999; **17**(7): 1318–1325.
55. Li W, Chao X. Modelling and performance evaluation of a cellular mobile network. *IEEE/ACM Transactions on Networking* 2004; **12**(1): 131–145.
56. Li W, Fang Y, Henry RR. Actual call connection time characterization for wireless mobile networks under a general channel allocation scheme. *IEEE Transactions on Wireless Communications* 2002; **1**(4): 682–691.
57. LePage WR. *Complex Variables and the Laplace Transform for Engineers*. Dover Publications, Inc.: New York, 1980.
58. Lin P. Channel allocation for GPRS with buffering mechanisms. *ACM Wireless Networks* 2003; **9**: 431–441.
59. Lin YB. Reducing location update cost in a PCS network. *IEEE/ACM Transactions on Networking* 1997; **5**(1): 25–33.
60. Lin YB, Chlamtac I. *Wireless and Mobile Network Architectures*. John Wiley and Sons: New York, 2001.
61. Lin YB, Mohan S, Noerpel A. Queueing priority channel assignment strategies for handoff and initial access for a PCS network. *IEEE Transactions on Vehicular Technology* 1994; **43**(3): 704–712.
62. Lin YB, Noerpel A, Harasty D. The sub-rating channel assignment strategy for PCS Hand-offs. *IEEE Transactions on Vehicular Technology* 1996; **45**(1): 122–130.
63. Marsan MA, Ginella G, Maglione R, Meo M. Performance analysis of hierarchical cellular networks with generally distributed call holding times and dwell times. *IEEE Transactions on Wireless Communications* 2004; **3**(1): 248–257.
64. Nanda S. Teletraffic models for urban and suburban microcells: cell sizes and handoff rates. *IEEE Transactions on Vehicular Technology* 1993; **42**(4): 673–682.
65. Noerpel AR, Lin YB, Sherry H. PACS: Personal access communications system—a tutorial. *IEEE Personal Communications* 1996; **3**(3): 32–43.
66. Orlik P, Rappaport SS. A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions. *IEEE Journal on Selected Areas in Communications* 1998; **16**(5): 788–803.
67. Rajaratnam M, Takawira F. Nonclassical traffic modeling and performance analysis of cellular mobile networks with and without channel reservation. *IEEE Transactions on Vehicular Technology* 2000; **49**(3): 817–834.
68. Ramjee R, Towsley D, Nagarajan R. On optimal call admission control in cellular networks. *Wireless Networks* 1997; **3**: 29–41.
69. Rappaport SS. Blocking, hand-off and traffic performance for cellular communication systems with mixed platforms. *IEE Proceedings–I* 1993; **140**(5): 389–401.

70. Rappaport TS, Annamalai A, Buehrer RM, Tranter WH. Wireless communications: past events and a future perspective. *IEEE Communications Magazine* 2002; **40**: 148–161.
71. Shafi M, Hashimoto A, Umehira M, Ogose S, Murase T. Wireless communications in the twenty-first century: a perspective. *Proceedings of the IEEE* 1997; **85**(10): 1622–1638.
72. Tekinay S, Jabbari B. A measurement-based prioritization scheme for handovers in mobile cellular networks. *IEEE Journal Selected Areas in Communications* **10**(8): 1343–1350.
73. Tekinay S, Jabbari B. Handover and channel assignment in mobile cellular networks. *IEEE Communications Magazine* 1991; New York, USA, 42–46.
74. Tijms HC. *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, 1994.
75. Xie H, Goodman DJ. Mobility models and biases sampling problem. *Proceedings of The 2nd IEEE International Conference on Universal Personal Communications (ICUPC)*, Vol. 2, 1993; pp. 803–807.
76. Xie H, Kuek S. Priority handoff analysis. *Proceedings of IEEE VTC*, 1993; pp. 855–858.
77. Yeo K, Jun C-H. Modeling and analysis of hierarchical cellular networks with general distributions of call and cell residence times. *IEEE Transactions Vehicular Technology* 2002; **51**(6): 1361–1374.
78. Yoon CH, Un CK. Performance of personal portable radio telephone systems with and without guard channels. *IEEE Journal Selected Areas in Communications* 1993; **11**(6): 911–917.
79. Yum T-SP, Yeung KL. Blocking and handoff performance analysis of directed retry in cellular mobile systems. *IEEE Transactions Vehicular Technology* 1995; **44**(3): 645–650.
80. Zhang Y. *Wireless Network Performance Evaluation: Link Unreliability and Parameter Sensitivity*, Ph.D Dissertation, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 2004.
81. Zhang Y, Soong B-H. The effect of unreliable wireless channel on the call performance in mobile network, accepted for publication in *IEEE Transactions on Wireless Communications* 2005, **4**(2): 653–661.
82. Zonoozi MM, Dassanayake P. User mobility modeling and characterization of mobility patterns. *IEEE Journal on Selected Areas in Communications* 1997; **15**(7): 1239–1252.

Authors' Biography



Yuguang Fang received his Ph.D. in Systems and Control Engineering from Case Western Reserve University in January 1994, and his Ph.D. in Electrical Engineering from Boston University in May 1997. From June 1997 to July 1998, he was a visiting assistant professor in Department of Electrical Engineering at the University of Texas at Dallas. From July 1998 to May 2000, he was an assistant professor in the Department of Electrical and Computer Engineering at New Jersey Institute of Technology. In May 2000, he joined the Department of Electrical and Computer Engineering at University of Florida where he got the early promotion to associate professor with tenure in August 2003 and has been a Professor since August 2005. He has published over 150 papers in refereed professional journals and conferences. He received the National Science Foundation Faculty Early Career Award in 2001 and the Office of Naval Research Young Investigator Award in 2002. He is currently serving as an Editor for many journals including *IEEE Transactions on Communications*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Mobile Computing*, and *ACM Wireless Networks*. He is also actively participating in conference organization such as the Program Vice-Chair for IEEE INFOCOM'2005, Program Co-Chair for the Global Internet, and Next Generation Networks Symposium in IEEE Globecom'2004 and the Program Vice Chair for 2000 IEEE Wireless Communications and Networking Conference (WCNC'2000).