A Mobility Management Strategy for GPRS

Yi-Bing Lin, *Fellow, IEEE*, and Shun-Ren Yang

***Abstract*—In general packet radio service (GPRS), a mobile station (MS) is tracked at the cell level during packet transmission, and is tracked at the routing-area (RA) level when no packet is delivered. A READY timer (RT) mechanism was proposed in 3GPP 23.060 to determine when to switch from cell tracking to RA tracking. In this mechanism, a threshold interval T is defined. If no packet is delivered within T, the MS is tracked at the RA level. When a packet arrives, the MS is tracked at the cell level again. However, the RT mechanism has a major fallacy in that the RTs in both the MS and the serving GPRS support node may lose synchronization. This paper considers another mechanism called READY counter (RC) to resolve this problem. In this approach, a threshold K is used. Like the RT approach, the MS is tracked at the cell level during packet transmission. If no packets are delivered after the MS has made K cell crossings, the MS is tracked at the RA level. We also devise an adaptive algorithm called dynamic RC (DRC). This algorithm dynamically adjusts the K value to reduce the location update and paging costs. We propose analytic and simulation models to investigate RC, RT, and DRC. Our study indicates that RC may outperform RT. We also show that DRC nicely captures the traffic–mobility patterns and always adjusts the K threshold close to the optimal values.**

***Index Terms*—General packet radio service (GPRS), mobile network, mobility management (MM), wireless data.**

NOMENCLATURE

α The probability that an ON-period is followed by an OFF-period in the same session.

$A(n)$ The total number of states for an n-layer RA random walk.

$B(n)$ The number of boundary edges in an n-layer RA.

C_T The expected total cost for location update and terminal paging during t_p.

$C_T(K)$ The net cost in an idle period with threshold K.

C_u The expected location update cost during t_p.

C_v The expected terminal paging cost during t_p.

$f_m(t_{m,j})$ The density function for the $t_{m,j}$ distribution.

$f_p(t_p)$ The density function for the t_p distribution.

$f_p^*(s)$ The Laplace Transform for the t_p distribution.

K The RC threshold.

$1/\lambda_m$ The expected value for the $t_{m,j}$ distribution.

$1/\lambda_p$ The expected value for the t_p distribution.

$1/\lambda_{p1}$ The expected value for the OFF-periods t_{p1} distribution.

$1/\lambda_{p2}$ The expected value for the intersession idle periods t_{p2} distribution.

N_c The number of cell crossings during t_p.

$N_r(j)$ The number of RA crossings occurring between the $j+1$th cell crossing and the N_cth cell crossing during t_p.

N_u The number of location updates (cell updates plus RA updates) during t_p.

$P(N_m, k)$ The probability that after N_m cell movements, an MS crosses k RA boundaries provided that the MS is initially in an arbitrary cell of an RA.

$P^*(N_m^*, k^*)$ The probability that after N_m^* cell movements, an MS crosses k^* RA boundaries provided that the MS is initially at a boundary cell of an RA.

$S(n)$ The number of cells in an n-layer RA.

T The RT threshold.

$t_{m,j}$ The cell residence time of an MS at cell C_j.

t_p The time interval between the end of a packet transmission and the beginning of the next packet transmission.

U The cost for a cell/RA update.

V The cost for paging in a cell.

V_m The variance for the $t_{m,j}$ distribution.

$\theta_1(k)$ The probability that an MS will leave an RA at the kth step provided that the MS is initially in an arbitrary cell of the RA.

$\theta_2(k)$ The probability that after an MS enters an RA, it moves out of the RA at the kth step.

I. INTRODUCTION

GENERAL PACKET radio service (GPRS) provides packet-switched data services for existing mobile telecommunication networks such as global system for mobile communications (GSM) and digital advanced mobile phone service [10]. GPRS core network has also evolved into 3G network (i.e., universal mobile telecommunications (UMTS) [11]). Most GSM-based mobile operators are deploying GPRS for wireless Internet services. The network architecture of GSM/GPRS is shown in Fig. 1. In this figure, the dashed lines represent signaling links, and the solid lines represent data and signaling links. The core network consists of two service domains: a *circuit-switched* (CS) service domain (i.e., PSTN/ISDN) and a *packet-switched* (PS) service domain (i.e., IP). GPRS is evolved from GSM by introducing two new core network nodes: *serving GPRS support node* (SGSN) and *gateway GPRS support node*. Existing GSM nodes including *base station subsystem* (BSS), *visitor location register* (VLR),

Manuscript received September 11, 2001; revised May 24, 2002; accepted July 3, 2002. The editor coordinating the review of this paper and approving it for publication is Y. Fang. The work of Y.-B. Lin was supported in part by MOE Program of Excellence Research under Contract 92-E-FA04–4, FarEastone, the Lee and MTI Center for Networking Research, NCTU. The work of S.-R. Yang was supported by the MediaTek Fellowship.

Y.-B. Lin was with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu 30050, Taiwan, R.O.C. He is now with the Institute of Information Science, Academia Sinica, Taiwan, R.O.C. (e-mail: liny@csie.nctu.edu.tw).

S.-R. Yang is with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu 30050, Taiwan, R.O.C. (e-mail: sjyoun@csie.nctu.edu.tw).

Digital Object Identifier 10.1109/TWC.2003.819034

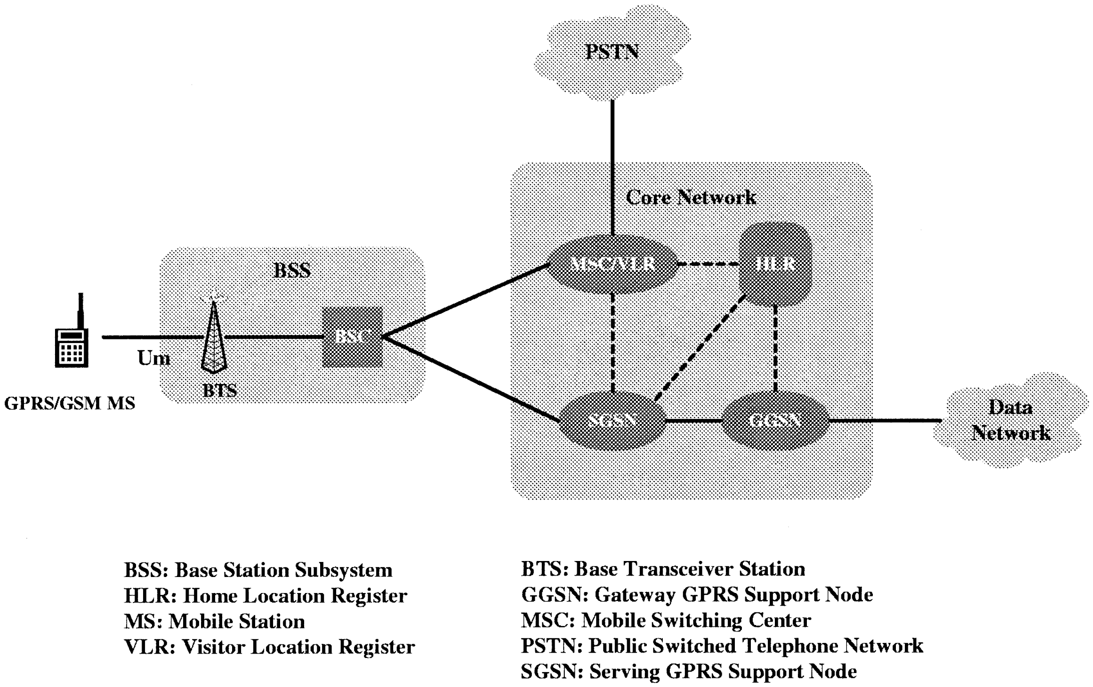

PSTN
Core Network
MSC/VLR
HLR
BSS
BSC
Um
BTS
GPRS/GSM MS
SGSN
GGSN
Data Network
BSS: Base Station Subsystem
HLR: Home Location Register
MS: Mobile Station
VLR: Visitor Location Register
BTS: Base Transceiver Station
GGSN: Gateway GPRS Support Node
MSC: Mobile Switching Center
PSTN: Public Switched Telephone Network
SGSN: Serving GPRS Support Node

Fig. 1. Network architecture of GSM/GPRS.

and *home location register* (HLR) are upgraded. GPRS BSS consists of *base transceiver stations* (BTSs) and *base station controller* (BSC) where the BSC is connected to the SGSN through frame relay link. The BTS communicates with the mobile station (MS) through the radio interface *Um* based on the time-division multiple-access technology.

The cells (i.e., radio coverages of BTSs) in a GPRS service area are partitioned into several groups. To deliver services to an MS, the cells in the group covering the MS will page the MS to establish the radio link. Location change of an MS is detected as follows. The cells broadcast their cell identities. The MS periodically listens to the broadcast cell identity and compares it with the cell identity stored in the MS's buffer. If the comparison indicates that the location has been changed, then the MS sends the location update message to the network.

In the CS domain, cells are partitioned into *location areas* (LAs). The LA of an MS is tracked by the VLR. In the PS domain, the cells are partitioned into *routing areas* (RAs). An RA is typically a subset of an LA. The RA of an MS is tracked by the SGSN. The SGSN also tracks the cell of an MS when packets are delivered between the MS and the SGSN.

In GPRS, the mobility management (MM) activities for an MS are characterized by an MM finite state machine exercised in both the SGSN and the MS. There are three states in the machine. In the IDLE state, the MS is not known (i.e., not attached) to GPRS. In the STANDBY state, the MS is attached to GPRS and the MS is tracked by the SGSN at the RA level. In the READY state, the SGSN tracks the MS at the cell level. Packet data units can only be delivered in this state. Descriptions of transitions among the MM states can be found in [11] and are briefly described as follows.

T1) IDLE→READY. This transition is triggered by an MS when the MS performs GPRS attach.

T2) READY→IDLE. This transition is triggered by the MS or the SGSN when the MS is detached from the GPRS network.

T3) STANDBY→READY. This transition occurs when the MS sends a packet data unit to the SGSN, possibly in response to a page from the SGSN.

T4) READY→STANDBY. This transition is triggered by either the SGSN or the MS. In GPRS, a READY timer (RT) is maintained in the MS and the SGSN. If no packet data unit is transmitted before the timer expires, then this MM transition occurs. The length of the RT can only be changed by the SGSN. The MS is informed of the RT value change through messages such as Attach Accept and RA Update Accept. This MM transition may also occur when the SGSN forces to do so or when abnormal condition is detected during radio transmission.

T5) STANDBY→IDLE. This transition is triggered by the SGSN when tracking of MS is lost. This transition may also be triggered by SGSN when the SGSN receives a Cancel Location message from the HLR, which implies that the MS has moved to the service area of another SGSN.

Transition T4 merits further discussion. In the READY state, the MS expects to receive packets in short intervals. Therefore, when the MS moves to a new cell, it should inform the SGSN of the movement immediately. In this way, the SGSN can deliver the next packet to the destination cell without paging the whole RA. On the other hand, if the communication session between the MS and the SGSN completes, the SGSN may not send the next packet (the first packet of the next session) to the MS in a long period. In this case, tracking the MS at the cell level is too expensive. Thus, the MM state should be switched

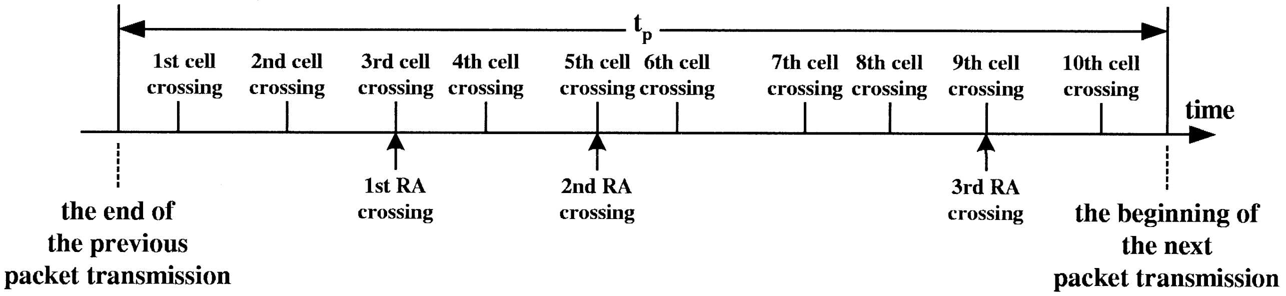


Fig. 2. Cell and RA crossings in an idle period.

to STANDBY and the MS is tracked at the RA level. To conclude, in the READY state, no paging is required (the packets are sent directly to the MS) while the location update cost is high (location update is performed for every cell movement). In the STANDBY state, the paging cost is high (all cells in the RA are paged), while the location update cost is low (location update is performed for every RA movement). The T4 transition can be implemented by two approaches. In the RT approach [1], an RT threshold T is defined. At the end of a packet transmission, the RT timer is set to the T value and is decremented as time elapses. Transition T4 occurs if the MS does not receive the next packet before the RT timer expires. However, the RT approach has a major fallacy that the RT timers in both the MS and the SGSN may lose synchronization (i.e., when the SGSN moves to STANDBY, the MS may be still in READY). To resolve this problem, we consider the *READY counter* (*RC*) approach. In the RC approach, an RC counter counts the number of cell movements in the packet idle period between two packet transmissions to an MS. If the number of movements reaches a threshold K, then the MM state switches from READY to STANDBY. To capture user mobility more accurately, one may dynamically adjust the K value to further reduce the net cost of cell/RA updates and paging. Consider the interval t_p between the end of a packet transmission and the beginning of the next packet transmission. If we know the number of cell crossings in t_p and the distribution of RA crossings among these cell crossings, then we can find the optimal K value such that the net cost is minimized. Let N_c be the number of cell crossings during t_p. Let $N_r(j)$ be the number of RA crossings occurring between the $j+1$th cell crossing and the N_cth cell crossing, where $j < N_c$. By convention, $N_r(K) = 0$ for $K \geq N_c$. Fig. 2 illustrates the cell and RA crossings in an idle period. In this example, $N_c = 10$. If $K = 2$, then $N_r(2) = 3$. If $K = 4$, then $N_r(4) = 2$. Let U be the cost for a cell/RA update and V be the cost for paging in a cell. Let S be the number of cells in an RA. Consider the RC algorithm with threshold K. The net cost $C_T(K)$ in an idle period can be expressed as

$$C_T(K) = \begin{cases} UN_c, & \text{for } K > N_c \\ U[K + N_r(K)] + SV, & \text{for } K \leq N_c. \end{cases} \quad (1)$$

In the following theorem, we show how to find the optimal threshold value K^* for RC in an idle period such that the net cost is minimized.

Theorem 1: Consider an idle period where no packet is delivered. Let N_c be the number of cell crossings in this period. In the RC algorithm, let K^* be the optimal threshold value that minimizes the net cost $C_T^* = C_T(K^*)$ in the idle period. Then, $K^* = 0$ or $K^* = N_c + 1$.

Proof: As previously defined, $N_r(K)$ is the number of RA updates in RC with threshold K. If $N_c \geq K$, then

$$C_T(K) = U[K + N_r(K)] + SV. \quad (2)$$

The number of RA crossings within the first K cell crossings is $N_r(0) - N_r(K)$. It is clear that

$$N_r(0) - N_r(K) \leq K. \quad (3)$$

From (2) and (3), we have

$$C_T(K) \geq U[N_r(0)] + SV = C_T(0), \quad \text{for } K \leq N_c. \quad (4)$$

For $K > N_c$, we have $C_T(K) = C_T(N_c + 1)$. Therefore

$$C_T^* = \min_{0 \leq K \leq \infty} C_T(K) = \min_{0 \leq K \leq N_c + 1} C_T(K). \quad (5)$$

From (4) and (5), we have

$$C_T^* = \min\left[C_T(0), C_T(N_c + 1)\right]. \quad (6)$$

In other words, $K^* = 0$ or $K^* = N_c + 1$.

Based on Theorem 1, we devise an algorithm to select K as follows. Let $t_p(i)$ be the interval between the end of the $i-1$th packet transmission and the beginning of the ith packet transmission. Let $K(i)$ be the optimal K value for $t_p(i)$. The K value can be dynamically adjusted using the following algorithm.

Dynamic RC (DRC) Algorithm

Initialization. Assign an arbitrary value to $K(0)$. Exercise the RC approach with threshold $K(0)$ before the first packet arrives.

When the ith packet transmission is completed. Compute the optimal $K(i)$ that minimizes the net cost for the period $t_p(i)$. Based on Theorem 1, $K(i)$ is either zero or $N_c + 1$ in $t_p(i)$, which can be quickly computed with very low cost. Exercise the RC approach with threshold $\bar{K}$ during $t_p(i+1)$, where

$$\bar{K} = \begin{cases} \left\lceil \sum_{j=i-M+1}^{i} \frac{K(j)}{M} \right\rceil, & \text{for } i \geq M \\ \left\lceil \sum_{j=1}^{i} \frac{K(j)}{i} \right\rceil, & \text{for } i < M \end{cases}. \quad (7)$$

In other words, the threshold $\bar{K}$ between the ith and the $i+1$th packet transmissions

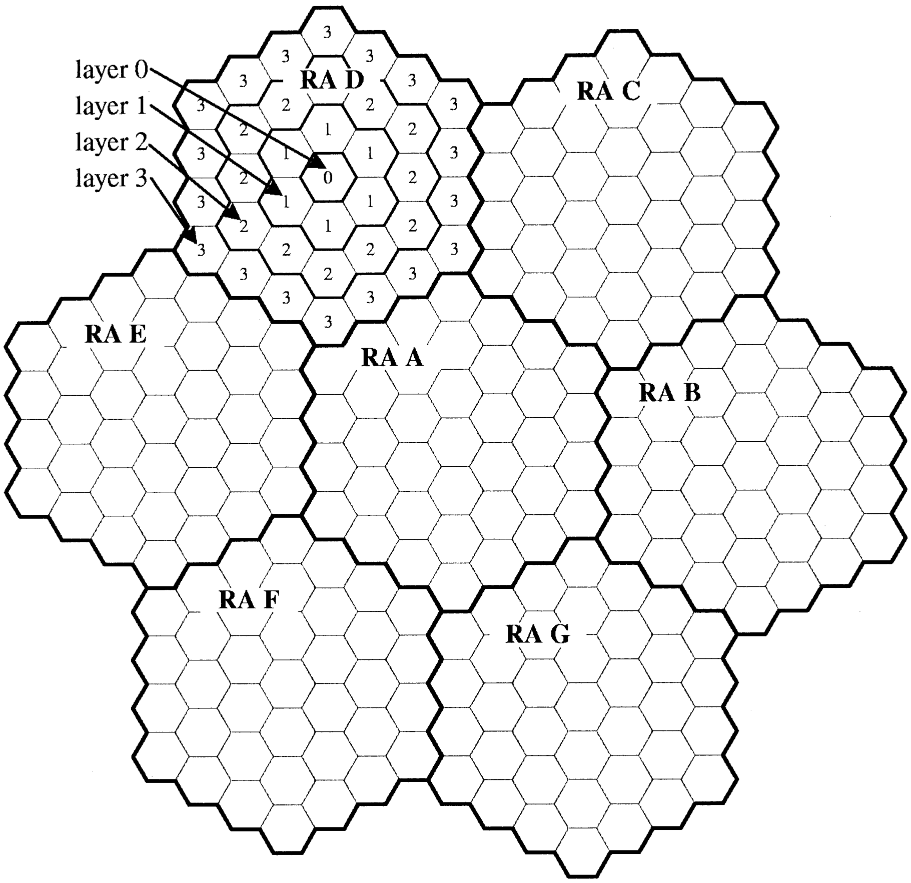


Fig. 3. Cell/RA layout in a GPRS network.

is selected as the average of the previous M optimal K values. Simulation experiments show that $M \geq 5$ is appropriate when using (7) to compute $\bar{K}$. In this paper, we consider $M = 25$.

Since the MS has the complete information on the numbers of cell and RA crossings in t_p, the DRC algorithm is implemented in MS. In DRC, the state transition of SGSN (i.e., from READY to STANDBY) is triggered by the MS. When the MS crosses the $\bar{K}$th cell boundary during t_p, it sends an RA update message to the SGSN instead of the cell update message. Once the SGSN receives the first RA update message from the MS, it switches the MM state from READY to STANDBY. No extra message is introduced to switch the MM states in DRC. Thus, the network signaling cost is the same as that of the static RC mechanism. The only cost incurred by DRC is the computation of $\bar{K}$ in the MS. As shown in (7), the computation can be done in microseconds. This cost is not significant and can be ignored.

This paper proposes analytic and simulation models to study the RT, RC, and DRC approaches. We investigate how the RA size and the location update–paging costs affect the performance of these approaches. Specifically, we show that RC is better than RT, and that DRC can automatically adjust the K value to minimize the net cost. The notation used in this paper is listed in the Nomenclature.

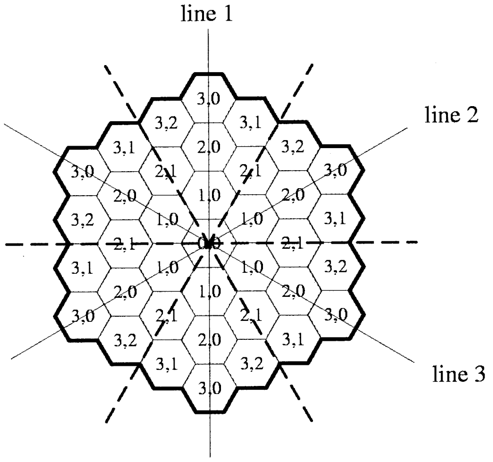


Fig. 4. Type classification for a four-layer RA.

II. Analytic Model for READY Counter Mechanism

This section develops an analytic model to study the GPRS RC mechanism. We first describe a uniform random walk model for user movement. Then, we show how to use this model to investigate the GPRS RC performance. For a specific threshold K, we derive the number of cell updates and RA updates between the end of a packet transmission and the beginning of the next packet transmission. Our model considers a GPRS network with hexagonal cell layout. Such a layout with small cells is reasonable for mobile services in big cities. In this configuration, the cells are grouped into several n-layer RAs. Every RA covers $S(n) = 3n^2 - 3n + 1$ cells, as shown in Fig. 3 (where $n = 4$). The figure plots seven RAs (A, B, C, D, E, F, and G) and the cells within the RAs. The cell at the center of an RA is referred to as the *Layer 0* cell. The cells surrounding layer $x - 1$ cells are referred to as *layer* x cells. There are $6x$ cells in layer x except for Layer 0 which contains exactly one cell. An n-layer RA consists of cells from Layer 0 to layer $n - 1$. Based on this RA/cell structure, we derive the number of cell crossings before a user crosses an RA boundary.

Based on the equal routing probability assumption (i.e., the MS moves to each of the neighboring cells with probability 1/6), we classify the cells in an RA into several cell types [2]. For $x \geq 0$ and $y \geq 0$, a cell type is of the form $\langle x, y\rangle$, where x represents that the cell is in layer x, and y represents the $y + 1$th type in layer x. Cells of the same type are indistinguishable in terms of movement pattern because they are at the symmetrical positions (with respect to lines 1, 2, and 3 in Fig. 4) on the hexagonal RA. According to the type classification algorithm in [2], Fig. 4 illustrates the types of cells for a four-layer RA. We develop a random walk model to compute when an MS crosses the boundary of an n-layer RA. A state of this random walk is of the form (x, y). For $0 \leq x < n$ and $0 \leq y \leq x - 1$, the state (x, y) is transient, which represents that the MS is in one of the cells of type $\langle x, y\rangle$. For $x = n$ and $0 \leq y < n - 1$, the state (n, y) is absorbing, which represents that the MS crosses the boundary of the RA from a cell of type $\langle n - 1, y\rangle$. The total number of states for an n-layer RA random walk is equal to $A(n) = n(n+1)/2$. Let $p_{(x,y),(x',y')}$ be the one-step transition probability from state (x, y) to state (x', y'), i.e., the probability that the MS moves from a $\langle x, y\rangle$ cell to a $\langle x', y'\rangle$ cell in one step. From [2], the transition probability matrix $P = (p_{(x,y),(x',y')})$ of the random walk is given as

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{6} & \frac{1}{3} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & \frac{1}{3} & \frac{1}{6} & \cdots & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}_{A(n)\times A(n)}. \quad (8)$$

Note that the number of states for the random walk can be further reduced if we consider the symmetry along the dashed lines in Fig. 4. The details will not be presented in this paper.

Let $p^{(k)}_{(x,y),(x',y')}$ be the probability that the random walk moves from state (x, y) to state (x', y') with exact k steps. Let $p_{k,(x,y),(n,j)}$ be the probability that an MS initially resides at a $\langle x, y\rangle$ cell, moves into a $\langle n - 1, j\rangle$ cell at the $k - 1$th step and then moves out of the RA at the kth step. Then

$$p_{k,(x,y),(n,j)} = \begin{cases} p_{(x,y),(n,j)}, & \text{for } k = 1 \\ p^{(k)}_{(x,y),(n,j)} - p^{(k-1)}_{(x,y),(n,j)}, & \text{for } k > 1. \end{cases} \quad (9)$$

Equation (9) can be solved using the transition probability matrix (8).

Suppose that an MS is in any cell of an RA with equal probability. In other words, the MS is in cell $\langle 0, 0\rangle$ with probability $1/S(n)$ and is in a cell of type $\langle x, y\rangle$ ($0 < x < n$, $0 \leq y \leq x - 1$) with probability $6/S(n)$, where $S(n) = 3n^2 - 3n + 1$ is the number of cells covered by an n-layer RA. Consider the example where a four-layer RA covers $S(4) = 37$ cells (see Fig. 4). Since there are one cell of type $\langle 0, 0\rangle$ and six cells of type $\langle 2, 1\rangle$ in a four-layer RA, the MS is in cell $\langle 0, 0\rangle$ with probability 1/37 and in a cell of type $\langle 2, 1\rangle$ with probability 6/37. Let $\theta_1(k)$ be the probability that the MS will leave the RA at the kth step. Then

$$\theta_1(k) = \left[\frac{1}{S(n)}\right]\left[\sum_{j=0}^{n-2} p_{k,(0,0),(n,j)}\right] + \left[\frac{6}{S(n)}\right]\left[\sum_{x=1}^{n-1}\sum_{y=0}^{x-1}\sum_{j=0}^{n-2} p_{k,(x,y),(n,j)}\right]. \quad (10)$$

Let $\theta_2(k)$ be the probability that after an MS enters an RA, it moves out of the RA at the kth step. Probability $\theta_2(k)$ is derived as follows. It can be shown [12] that after entering the RA, the MS is in a boundary cell with probability in proportion to the number of boundary edges for that boundary cell. Under the condition that an MS is moving into a boundary cell, the MS is in a boundary cell of type $\langle n - 1, 0\rangle$ with probability $3 \cdot 6/B(n)$ and is in a boundary cell of type $\langle n - 1, y\rangle$ ($1 \leq y \leq n - 2$) with probability $2 \cdot 6/B(n)$, where $B(n) = 6[3 + 2(n - 2)]$ is the number of boundary edges in an n-layer RA. In Fig. 4, $B(n) = 42$ for a four-layer RA. In this example, there are three boundary edges for each of the six $\langle 3, 0\rangle$ cells, and the MS is in a boundary cell of type $\langle 3, 0\rangle$ with probability 18/42. Similarly, there are two boundary edges for each of the six $\langle 3, 1\rangle$ cells, and the MS is in a boundary cell of type $\langle 3, 1\rangle$ with probability 12/42. Based on the above discussion, we have

$$\theta_2(k) = \left[\frac{3\cdot 6}{B(n)}\right]\left[\sum_{j=0}^{n-2} p_{k,(n-1,0),(n,j)}\right] + \left[\frac{2\cdot 6}{B(n)}\right]\left[\sum_{y=1}^{n-2}\sum_{j=0}^{n-2} p_{k,(n-1,y),(n,j)}\right]. \quad (11)$$

Suppose that an MS is in an arbitrary cell of an RA. Let $P(N_m, k)$ be the probability that after N_m cell movements, the MS crosses k RA boundaries. Similarly, consider an MS that is initially at a boundary cell of an RA. Let $P^*(N^*_m, k^*)$ be the probability that after N^*_m cell movements, the MS crosses k^* RA boundaries. From (10) and (11), we have

$$P(N_m, k) = \begin{cases} 1, & \text{for } k = N_m = 0 \\ \sum_{j=N_m+1}^{\infty} \theta_1(j), & \text{for } k = 0, N_m > 0 \\ \sum_{j=1}^{N_m} \theta_1(j) \\ \quad \times P^*(N_m - j, k - 1), & \text{for } k \geq 1,\ N_m \geq k \\ 0, & \text{for } N_m < k \end{cases} \quad (12)$$

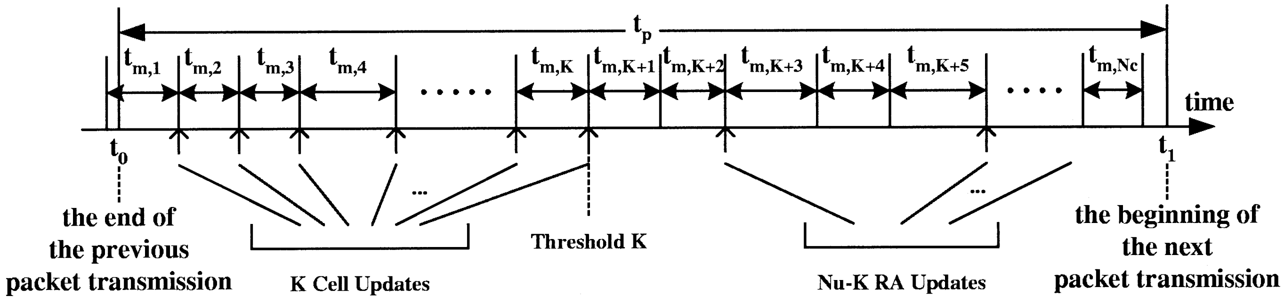


Fig. 5. Timing diagram for cell and RA crossings.

$$
P^*(N_m^*, k^*) = \begin{cases} 1, & \text{for } k^* = N_m^* = 0 \\ \sum\limits_{j=N_m^*+1}^{\infty} \theta_2(j), & \text{for } k^*=0,\ N_m^*>0 \\ \sum\limits_{j=1}^{N_m^*} \theta_2(j) \\ \times P^*(N_m^* - j, k^* - 1), & \text{for } k^* \geq 1,\ N_m^* \geq k^* \\ 0, & \text{for } N_m^* < k^*. \end{cases} \tag{13}
$$

Equations (12) and (13) can be effectively computed using dynamic programming technique [13]. Note that the above derivations are based on the uniform movement assumption. To accommodate nonuniform random walk, we need to modify the state diagram and the transition probability matrix. With (12) and (13), we derive the number of cell/RA updates between the end of a packet transmission and the beginning of the next packet transmission as follows. Fig. 5 shows the timing diagram of the activities for an MS. Suppose that the previous packet transmission of the MS ends at time t_0 and the next packet transmission begins at time t_1. Let $t_p = t_1 - t_0$, which has a general distribution with density function $f_p(t_p)$, expected value $1/\lambda_p$, and Laplace Transform

$$
f_p^*(s) = \int_{t_p=0}^{\infty} e^{-st_p} f_p(t_p) dt_p.
$$

For RC with a specific threshold K, let N_u be the number of location updates (cell updates plus RA updates) during the period t_p. Based on the random walk model mentioned above, the distribution of N_u can be derived as follows. Suppose that the cell residence time $t_{m,j}$ at cell C_j has an Erlang distribution with mean $1/\lambda_m = m/\lambda$, variance $V_m = m/\lambda^2$, and density function

$$
f_m(t) = \frac{\lambda e^{-\lambda t}(\lambda t)^{m-1}}{(m-1)!}, \quad \text{for } t \geq 0 \tag{14}
$$

where $m = 1, 2, 3, \ldots$ We select Erlang distribution because this distribution can be easily extended into hyper-Erlang distribution. Hyper-Erlang distribution has been proven as a good approximation to many other distributions as well as measured data [5], [9].

The probability mass function of the number of cell crossings N_c within t_p is shown in (15), at the bottom of the next page. Within the time interval t_p, the MS performs cell updates for the first K cell crossings and then performs RA updates whenever it crosses RA boundaries. From (12), (13), and (15), the probability mass function for N_u is

$$
\Pr[N_u = j] = \begin{cases} \Pr[N_c = j], & \text{for } j < K \\ \sum\limits_{k=j}^{\infty} \Pr[N_c = k] \\ \times P(k - K, j - K), & \text{for } j \geq K. \end{cases} \tag{16}
$$

Based on (15) and (16), we derive the expected total cost C_T for location update and paging during t_p. Assume that the cost for performing a location update is U and the cost for paging at one cell is V. Let C_u be the expected location update cost during t_p. From (16), we have

$$
C_u = U \sum_{j=0}^{\infty} j \Pr[N_u = j]. \tag{17}
$$

If the number of cell crossings $N_c < K$, then no cell needs to page the MS. Otherwise, all cells of the RA should page the MS. Let C_v be the expected terminal paging cost during t_p. From (15), we have

$$
C_v = VS(n) \sum_{j=K}^{\infty} \Pr[N_c = j]. \tag{18}
$$

From (17) and (18), the expected total cost C_T for location update and terminal paging during t_p is

$$
C_T = C_u + C_v. \tag{19}
$$

We have developed a discrete simulation model to validate against our analytic analysis. The simulation actually simulates the movement of an MS on the hexagonal plane. The θ_1, θ_2, $P(N_m, k)$, $P^*(N_m^*, k^*)$, and C_T values produced by the analytic and simulation models show that both models are consistent. For example, Fig. 6 plots the C_T curves for analytic and simulation results based on the three-layer RA configuration. The t_p intervals are exponentially distributed, and the cell residence times have Erlang distribution given in (14). Other parameters such as λ_{p1} and λ_{p2} will be explained in Section III. In this figure, the dashed curves represent the analytic model, and the solid curves represent the simulation experiments. The results indicate that the discrepancy between analytic analysis and simulation is within 1% in most cases. The comparisons for various input parameters and t_p, t_m distributions show similar results (i.e., both models are consistent) and will not be elaborated in this paper.

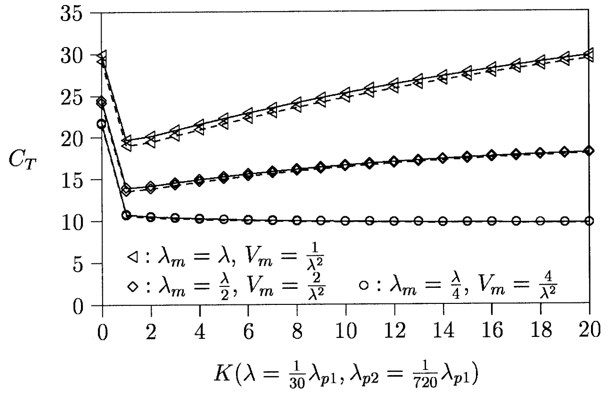


Fig. 6. Comparison between the analytic and simulation results.

III. Numerical Results

This section investigates the performance of the GPRS RC mechanism. Then, based on the simulation experiments, we compare RC with RT and show that DRC can capture an appropriate K value that reduces the net signaling cost for RC. In this study, we combine the ETSI packet data model [4] with the ON–OFF source model (also known as a packet train model) [7]. As shown in Fig. 7, we assume that the packet data traffic consists of communication sessions, where the intersession idle period has an exponential distribution with mean $1/\lambda_{p2}$. For general intersession idle periods, the variances of the idle period distributions have similar effects as that of the variances for cell residence times (to be elaborated later in Fig. 8), and the details will not be presented in this paper. Within a communication session, packet traffic is modeled by the ON–OFF source model. In an ON-period, a burst of data packets are sent. In an OFF-period, no packets are sent. Following the ETSI packet data model, the number of OFF-periods in a session has a geometric distribution with mean $\alpha/1 - \alpha$, where $0 \leq \alpha < 1$. In other words, an ON-period is followed by an OFF-period (in the same session) with probability α and is followed by the intersession idle period (for the next session) with probability $1 - \alpha$. The OFF-periods t_{p1} are drawn from a Pareto distribution [8] with mean $1/\lambda_{p1}$ and infinite variance, which has been found to match very well with the actual data traffic measurements [14]. A Pareto distribution has two parameters β and l, where β describes the "heaviness" of the tail of the distribution. The probability density function is

$$f_p(t_{p1}) = \left(\frac{\beta}{l}\right)\left(\frac{l}{t_{p1}}\right)^{\beta+1}$$

and the expected value is

$$E[t_{p1}] = \left(\frac{\beta}{\beta - 1}\right) l.$$

If β is between one and two, the variance for the distribution becomes infinity. The typical parameter values obtained in [14] are $E[t_{p1}] = 10.5$ s and $\beta = 1.2$ for OFF-periods. Our study follows the above $E[t_{p1}]$ and β values. We also assume that the cell residence time is Gamma distributed with mean $1/\lambda_m$ and variance V_m. The Gamma distribution with shape parameter η and scale parameter λ (i.e., mean $1/\lambda_m = \eta/\lambda$ and variance $V_m = \eta/\lambda^2$) has the following density function:

$$f_m(t) = \frac{\lambda e^{-\lambda t}(\lambda t)^{\eta - 1}}{\Gamma(\eta)}, \quad \text{for } t \geq 0$$

where $\Gamma(\eta) = \int_{z=0}^{\infty} z^{\eta-1} e^{-z} dz$ is the Gamma function. It has been shown that the distribution of any positive random variable can be approximated by a mixture of Gamma distributions [9,

$$\begin{aligned}
&\Pr[N_c = k] \\
&= \int_{t_p=0}^{\infty} \frac{e^{-\lambda t_p}}{m} \left\{ \sum_{j=km}^{km+m-1} \left[\frac{(km+m-j)(\lambda t_p)^j}{j!}\right] - \sum_{j=km-m}^{km-1} \left[\frac{(j-km+m)(\lambda t_p)^j}{j!}\right] \right\} f_p(t_p) dt_p \\
&= \frac{1}{m} \left\{ \sum_{j=km}^{km+m-1} \left[\frac{(km+m-j)\,\lambda^j}{j!}\right] \left[\int_{t_p=0}^{\infty} t_p^j f_p(t_p) e^{-\lambda t_p} dt_p\right] - \sum_{j=km-m}^{km-1} \left[\frac{(j-km+m)\,\lambda^j}{j!}\right] \left[\int_{t_p=0}^{\infty} t_p^j f_p(t_p) e^{-\lambda t_p} dt_p\right] \right\} \\
&= \frac{1}{m} \left\{ \sum_{j=km}^{km+m-1} \left[\frac{(km+m-j)\,(-\lambda)^j}{j!}\right] \left[\frac{d^j f_p^*(s)}{ds^j}\right]\Bigg|_{s=\lambda} - \sum_{j=km-m}^{km-1} \left[\frac{(j-km+m)\,(-\lambda)^j}{j!}\right] \left[\frac{d^j f_p^*(s)}{ds^j}\right]\Bigg|_{s=\lambda} \right\},\ k = 1, 2, \ldots \\
&\Pr[N_c = 0] \\
&= \int_{t_p=0}^{\infty} \frac{e^{-\lambda t_p}}{m} \sum_{j=0}^{m-1} \left[\frac{(m-j)(\lambda t_p)^j}{j!}\right] f_p(t_p) dt_p \\
&= \frac{1}{m} \left\{ \sum_{j=0}^{m-1} \left[\frac{(m-j)\,\lambda^j}{j!}\right] \left[\int_{t_p=0}^{\infty} t_p^j f_p(t_p) e^{-\lambda t_p} dt_p\right] \right\} \\
&= \frac{1}{m} \left\{ \sum_{j=0}^{m-1} \left[\frac{(m-j)(-\lambda)^j}{j!}\right] \left[\frac{d^j f_p^*(s)}{ds^j}\right]\Bigg|_{s=\lambda} \right\}
\end{aligned} \tag{15}$$

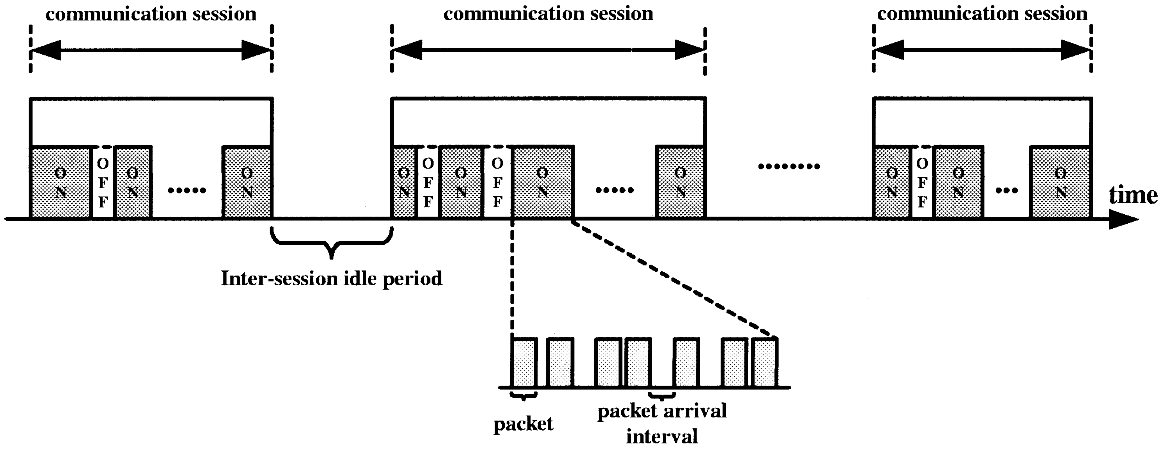
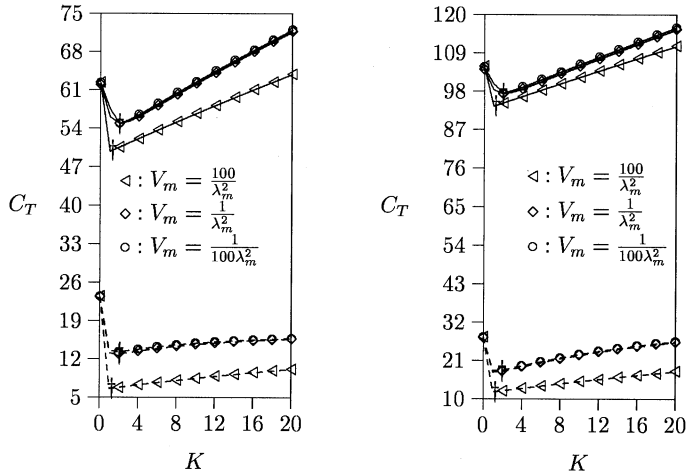


Fig. 7. Packet data traffic.



(a) $\lambda_{p2} = \frac{1}{100}\lambda_{p1}$

(b) $\lambda_{p2} = \frac{1}{200}\lambda_{p1}$

Fig. 8. Effects of V_m on C_T (solid: $\lambda_m = \lambda_{p1}$; dashed: $\lambda_m = (1/10)\lambda_{p1}$; $U/V = 4$, $\alpha = 0.6$).

Lemma 3.9]. The Gamma distribution was used to model MS movement in many studies [3], [5], [6] and is used in this paper to investigate the impact of variance for cell residence times. (The impact of V_m is shown in Fig. 8. In Figs. 9–12, $V_m = 1/\lambda_m^2$ for cell residence times.) Note that the Erlang distribution used in Fig. 6 is a special case of the Gamma distribution, where $V_m = 1/(m\lambda_m^2)$. The impacts of several input parameters are discussed as follows.

A. Effects of U/V Ratio

Fig. 9 plots C_T as functions of U/V, where U is the cost of a cell/RA update, and V is the cost for paging at a cell. The figure indicates that for a large U/V (e.g., $U/V > 20$), the C_T cost for a small RA layout is higher than that for a large RA layout. It is apparent that for the same number of cell crossings within a period, small RA layout has more RA crossings. If a location update operation is expensive (i.e., U/V is large), then a small RA layout will result in large C_T. On the other hand, the RA paging cost increases as the RA size increases. Therefore, when U/V is small (i.e., the paging cost dominates), C_T is an increasing function of the RA size. In the following discussions, we assume that $U/V = 4$. Other U/V ratios show similar results and will not be presented in this paper.

B. Effects of RA Size

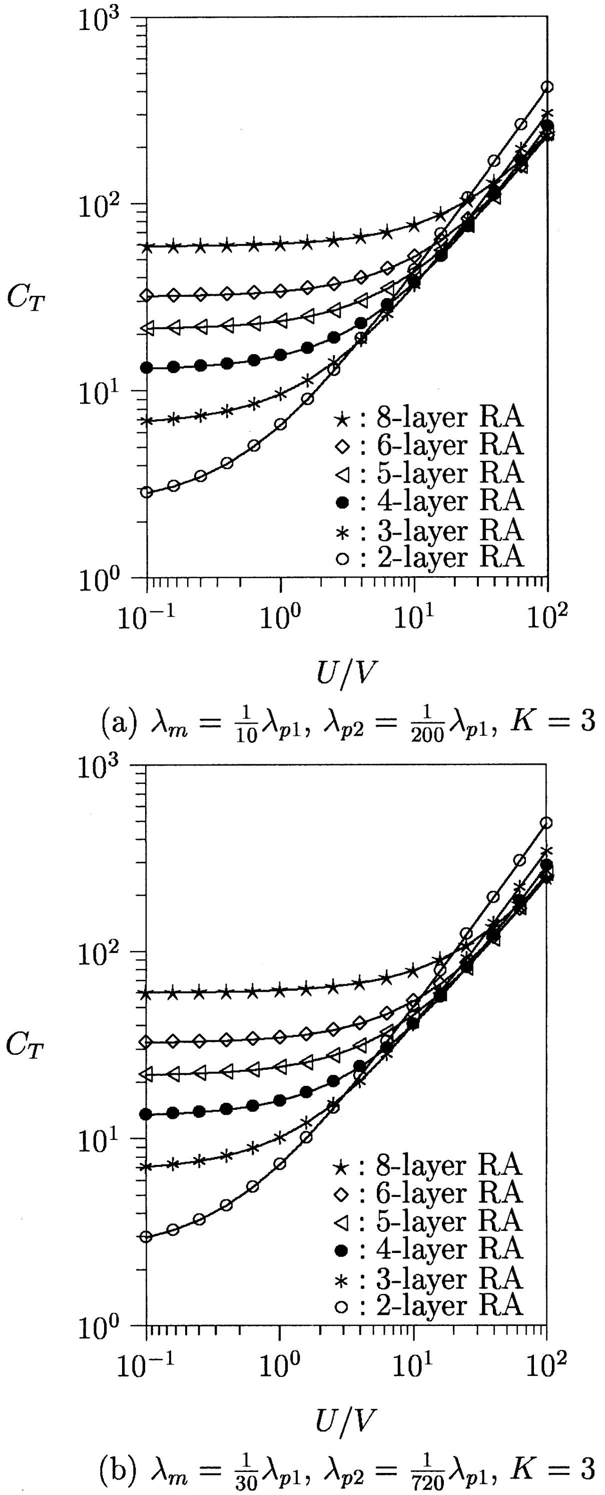
Fig. 10(a) shows that the five-layer RA layout has the lowest C_T for mobility rate $\lambda_m = \lambda_{p1}$ (where $\lambda_{p2} = (1/200)\lambda_{p1}$, $K = 3$, and $\alpha = 0.6$). Both the two-layer and the three-layer RA layouts have smaller C_T when λ_m is less than $(1/10)\lambda_{p1}$. Fig. 10(b) shows that the four-layer RA layout has the lowest C_T when $\lambda_{p2} = (1/720)\lambda_{p1}$ (where $\lambda_m = (1/10)\lambda_{p1}$, $K = 3$ and $\alpha = 0.6$). Both the two-layer and the three-layer RA layouts have lower C_T when λ_{p2} is larger than $(1/200)\lambda_{p1}$. Fig. 10(c) and (d) shows that both the two-layer and the three-layer RA layouts have smaller C_T for various K and α values, where $\lambda_m = (1/10)\lambda_{p1}$ and $\lambda_{p2} = (1/200)\lambda_{p1}$. For the packet traffic and mobility patterns considered in Fig. 10, the three-layer RA layout is likely to produce lower network operation cost (i.e., C_T) in most cases. The above discussion shows that the C_T curves in Figs. 9 and 10 provide guidelines for mobile service operators to deploy optimal cell layout. In the remainder of this paper, we assume three-layer RA configuration.

C. Effects of K

Fig. 11 shows how K affects C_T. As K increases, the location update cost increases while the paging cost decreases. Therefore, paging and location update are two conflict factors which result in concave C_T curves, and optimal K values exist. This observation justifies the motivation to develop a mechanism that dynamically selects appropriate K values when the traffic–mobility patterns change.

D. Effects of λ_m

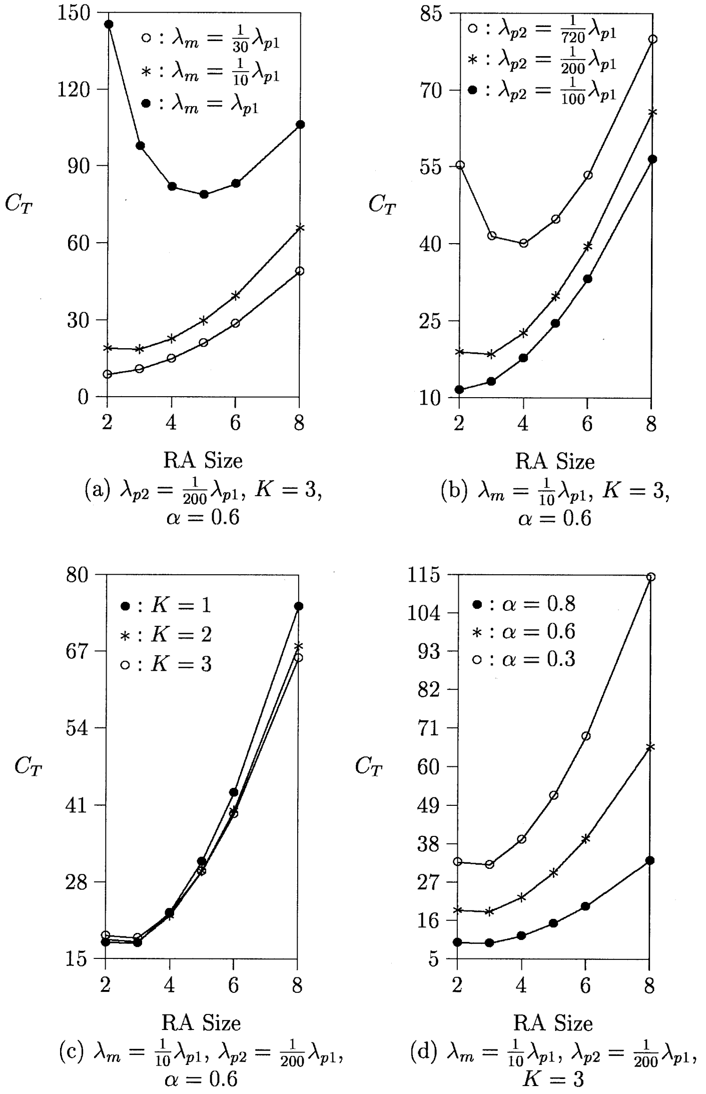
As λ_m increases, more cell crossings are expected, and C_T increases (see the $\diamond$ and the $\circ$ curves in Fig. 11). Note that when K is large and λ_m is small, increasing K only has insignificant effect on C_T. This phenomenon is due to the fact that $C_T(K) = C_T(N_c+1)$ for $K > N_c$. When λ_m is small and K is very large, it is likely that $K > N_c$, and increasing K only insignificantly increases the C_T value.



(a) $\lambda_m = \frac{1}{10}\lambda_{p1}$, $\lambda_{p2} = \frac{1}{200}\lambda_{p1}$, $K = 3$

(b) $\lambda_m = \frac{1}{30}\lambda_{p1}$, $\lambda_{p2} = \frac{1}{720}\lambda_{p1}$, $K = 3$

Fig. 9. Effects of U/V on C_T ($V_m = 1/\lambda_m^2$, $\alpha = 0.6$).



(a) $\lambda_{p2} = \frac{1}{200}\lambda_{p1}$, $K = 3$, $\alpha = 0.6$

(b) $\lambda_m = \frac{1}{10}\lambda_{p1}$, $K = 3$, $\alpha = 0.6$

(c) $\lambda_m = \frac{1}{10}\lambda_{p1}$, $\lambda_{p2} = \frac{1}{200}\lambda_{p1}$, $\alpha = 0.6$

(d) $\lambda_m = \frac{1}{10}\lambda_{p1}$, $\lambda_{p2} = \frac{1}{200}\lambda_{p1}$, $K = 3$

Fig. 10. Effects of RA size on C_T ($U/V = 4$, $V_m = 1/\lambda_m^2$).

E. Effects of λ_{p2}

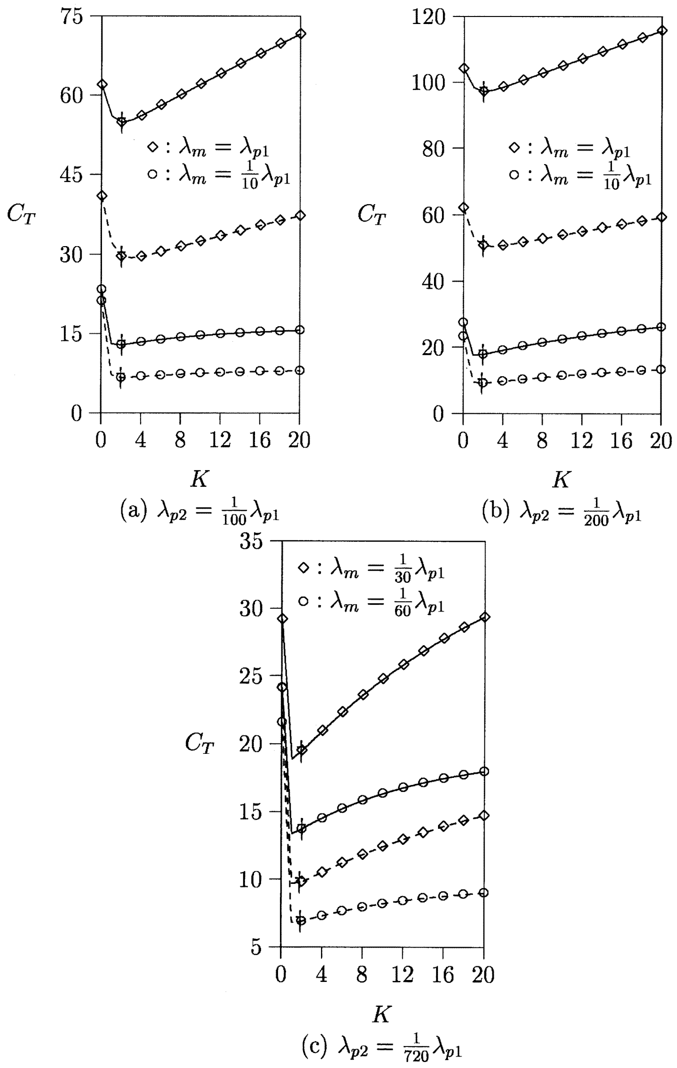
Fig. 11(a)–(c) plots the C_T curves for $\lambda_{p2} = (1/100)\lambda_{p1}$, $(1/200)\lambda_{p1}$, and $(1/720)\lambda_{p1}$, respectively. The figures show that C_T increases as λ_{p2} decreases. A small λ_{p2} implies large intersession idle periods, and more cell movements will occur during this period. Thus, a large C_T is expected. As in our previous discussion for the interaction between K and λ_m, if K is sufficiently large, C_T is not sensitive to the change of K when λ_{p2} is large (increasing λ_{p2} has similar effect as decreasing λ_m).

F. Effects of α

Fig. 11 shows the effect of α on C_T (the solid curves are for $\alpha = 0.6$ and the dashed curves are for $\alpha = 0.8$). A smaller α implies more intersession idle periods. Since intersession idle periods are longer than the OFF-periods in the sessions, C_T increases as α decreases. The figure also indicates that C_T is more sensitive to the mobility rate λ_m for a small α than a large α.

G. Effects of Variance V_m

Fig. 8 plots the C_T curves where the cell residence times have a Gamma distribution with mean $1/\lambda_m$ and variance V_m. The figure shows that a large V_m (e.g., $V_m = 100/\lambda_m^2$) results in a small C_T. This phenomenon is explained as follows. When V_m increases, more short and long cell residence times are observed. Long cell residence times imply small number of cell crossings in t_p, which result in a small C_T. On the other hand, short cell residence times imply large N_c values (more cell crossings), which increases C_T. However, when $N_c > K$, the number of



(a) $\lambda_{p2} = \frac{1}{100}\lambda_{p1}$

(b) $\lambda_{p2} = \frac{1}{200}\lambda_{p1}$

(c) $\lambda_{p2} = \frac{1}{720}\lambda_{p1}$

Fig. 11. Effects of K, λ_m, λ_{p2}, and α on C_T (solid: $\alpha = 0.6$; dashed: $\alpha = 0.8$; $U/V = 4$, $V_m = 1/\lambda_m^2$).

RA crossings among $N_c - K$ cell crossings does not increase as fast as N_c does. Therefore, the (negative) effect of short cell residence times are not as significant as the (positive) effect of long cell residence times, and the net effect is that C_T decreases as V_m increases. Fig. 8 also indicates that C_T is not affected by the change of V_m when $V_m \leq 1/\lambda_m^2$.

H. Comparison of RC and RT

Both RC and RT have similar performance in examples shown in Figs. 8, 10, and 11. However, when the MS mobility rate changes from time to time, RC may significantly outperform RT. The reason is given below. The threshold T of RT is a fixed time interval. When λ_m changes, T cannot adapt to the change. On the other hand, the threshold K of RC always captures the Kth cell movement of an MS no matter how λ_m changes. Consider the example in Fig. 12. In this experiment, we mix two types of packet traffic and mobility patterns.

Type I pattern: Mean OFF-period time $E[t'_{p1}] = 1/\lambda'_{p1} = 100$ s, mobility rate $\lambda'_m = 1/1000$ s and $V'_m = 1/\lambda'^2_m$.

Type II pattern: Mean OFF-period time $E[t''_{p1}] = 1/\lambda''_{p1} = 1000$ s, mobility rate $\lambda''_m = 1/10$ s and $V''_m = 1/\lambda''^2_m$.

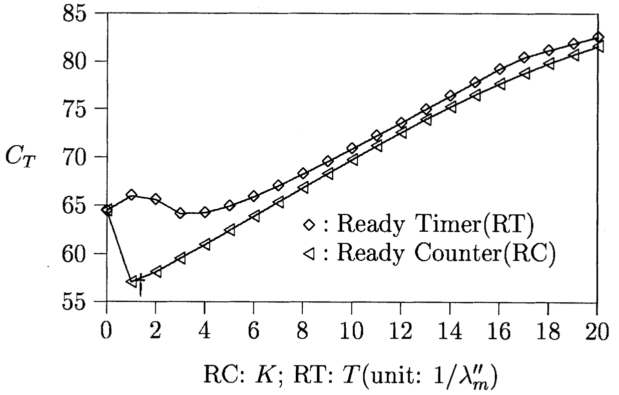


Fig. 12. Comparison between RC and RT ($U/V = 4$).

Consider 1 000 000 OFF-periods. Type I pattern is exercised for the first 500 000 OFF-periods, and Type II pattern is exercised for the remaining 500 000 OFF-periods. Let C'_T be the expected cost of an OFF-period for Type I pattern and C''_T be the expected cost for Type II pattern. In RC, the lowest C'_T is expected when $K = 1$ and the lowest C''_T is expected when $K = 0$. Thus, by choosing $K = 1$, good performance can be expected for both Type I and II patterns. In RT, the best threshold value for C'_T occurs when $T > 100$ s. On the other hand, the best threshold value for C''_T occurs when $T < 10$ s. Fig. 12 plots the C_T curve against K for RC and the C_T curve against T (measured by $1/\lambda''_m$) for RT. The figure indicates that selection of any T value will not satisfy both Type I and II patterns, and the resulting C_T cost for RT is higher than that of RC.

I. Performance of DRC

The C_T curves in Figs. 8, 11, and 12 are concave with respect to K, which indicate that optimal K values exist to minimize C_T. However, the optimal K values are not the same for different traffic–mobility patterns. Therefore, a mechanism that automatically selects the optimal K values in real time is required. The DRC algorithm proposed in Section I serves this purpose. In Figs. 8, 11, and 12, the "†" points represent the (C_T, $E[\overline{K}]$) pairs of DRC. These points indicate that DRC nicely captures the traffic–mobility patterns and always adjusts the RC threshold close to the optimal values.

IV. Conclusion

In GPRS, an MS is tracked at the cell level during packet transmission and is tracked at the RA level when no packet is delivered. An RT mechanism was proposed in 3GPP 23.060 [1] to determine when to switch from cell tracking to RA tracking. In this mechanism, a threshold interval T is defined. If no packet is delivered within T, the MS is tracked at the RA level. When a packet arrives, the MS is tracked at the cell level again. However, the RT mechanism has a major fallacy that the RTs in both the MS and the SGSN may lose synchronization. This paper considered another mechanism called RC to resolve this problem. In this approach, a threshold K is used. Like the RT approach, the MS is tracked at the cell level during packet transmission. If no packets are delivered after the MS has made K cell crossings, the MS is tracked at the RA level. We also devised an adaptive algorithm called DRC. This algorithm dynamically adjusts the K value to reduce the location update and paging costs. We proposed analytic and simulation models to investigate RC, RT, and DRC. Our study indicates the following.

- If the location update cost dominates, then a large RA layout should be chosen. On the other hand, if paging cost dominates, a small RA layout is appropriate.
- We quantitatively showed how the network operation cost (C_T) increases as intersession idle times and user mobility increase.
- The variance V_m of cell residence times affect the network operation cost C_T. As V_m increases, C_T decreases.
- RC and RT have similar performance when the mobility rate λ_m is fixed. RC may significantly outperform RT when λ_m changes from time to time.
- DRC nicely captures the traffic–mobility patterns and always adjusts the K threshold close to the optimal values.

ACKNOWLEDGMENT

The authors would like to thank the three anonymous reviewers. Their comments have significantly improved the quality of this paper.

REFERENCES

[1] *Technical Specification 3G TS 23.060 Version 3.6.0 (2001–01)*, 3GPP, 3rd Generation Partnership Project, Technical Specification Group Services and Systems Aspects, GPRS, Service Description, Stage 2, 2000.

[2] I. F. Akyildiz, Y.-B. Lin, W.-R. Lai, and R.-J. Chen, "A new random walk model for PCS networks," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 1254–1260, July 2000.

[3] I. Chlamtac, Y. Fang, and H. Zeng, "Call blocking analysis for PCS networks under general cell residence time," presented at the IEEE WCNC, New Orleans, LA, Sept. 1999.

[4] "UMTS Terrestrial Radio Access (UTRA), Concept Evaluation, Version 3.0.0," ETSI, UMTS 30.06, 1997.

[5] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling for PCS networks," *IEEE Trans. Commun.*, vol. 47, pp. 1062–1072, July 1999.

[6] Y. Fang, I. Chlamtac, and H.-B. Fei, "Analytical results for optimal choice of location update interval for mobility database failure restoration in PCS networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, pp. 615–624, June 2000.

[7] R. Jain and S. A. Routhier, "Packet trains: Measurements and a new model for computer network traffic," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 986–995, June 1986.

[8] N. L. Johnson, *Continuous Univariate Distributions-1*. New York: Wiley, 1970.

[9] F. P. Kelly, *Reversibility and Stochastic Networks*. New York: Wiley, 1979.

[10] Y.-B. Lin and I. Chlamtac, *Wireless and Mobile Network Architectures*. New York: Wiley, 2001.

[11] Y.-B. Lin, Y.-R. Haung, Y.-K. Chen, and I. Chlamtac, "Mobility management: From GPRS to UMTS," *Wireless Commun. Mobile Computing*, vol. 1, pp. 339–359, 2001.

[12] Y.-B. Lin, W.-R. Lai, and R.-J. Chen, "Performance analysis for dual band PCS networks," *IEEE Trans. Comput.*, vol. 49, pp. 148–159, Feb. 2000.

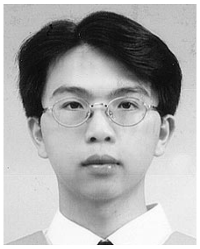
[13] U. Manber, *Introduction to Algorithms – A Creative Approach*. Reading, MA: Addison-Wesley, 1989.

[14] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level," *IEEE/ACM Trans. Networking*, vol. 5, pp. 71–86, Feb. 1997.

Yi-Bing Lin (M'95–SM'95–F'03) received the BSEE degree from the National Cheng Kung University, Tainan, Taiwan, R.O.C., in 1983, and the Ph.D. degree in computer science from the University of Washington, Seattle, in 1990.

From 1990 to 1995, he was with the Applied Research Area at Bell Communications Research (Bellcore), Morristown, NJ. In 1995, he became a Professor in the Department of Computer Science and Information Engineering (CSIE), National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C. In 1996, he was appointed Deputy Director of the Microelectronics and Information Systems Research Center, NCTU. From 1997 to 1999, he was Chairman of CSIE, NCTU. He is an Adjunct Research Fellow of Academia Sinica. His current research interests include design and analysis of personal communications services network, mobile computing, distributed simulation, and performance modeling.

Dr. Lin is an Associate Editor of *IEEE Network*, an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Associate Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, an Associate Editor of IEEE COMMUNICATIONS SURVEY AND TUTORIALS, an Editor of *IEEE Personal Communications Magazine*, an Editor of *Computer Networks*, an Area Editor of *ACM Mobile Computing and Communication Review*, a Columnist of *ACM Simulation Digest*, an Editor of *International Journal of Communications Systems*, an Editor of *ACM/Baltzer Wireless Networks*, an Editor of *Computer Simulation Modeling and Analysis*, an Editor of *Journal of Information Science and Engineering*, Program Chair for the 8th Workshop on Distributed and Parallel Simulation, General Chair for the 9th Workshop on Distributed and Parallel Simulation, Program Chair for the 2nd International Mobile Computing Conference, Guest Editor for the *ACM/Baltzer MONET Special Issue on Personal Communications*, a Guest Editor for IEEE TRANSACTIONS ON COMPUTERS SPECIAL ISSUE ON MOBILE COMPUTING, a Guest Editor for IEEE TRANSACTIONS ON COMPUTERS SPECIAL ISSUE ON WIRELESS INTERNET, and a Guest Editor for *IEEE Communications Magazine Special Issue on Active, Programmable, and Mobile Code Networking*. He is the coauthor (with Imrich Chlamtac) of the book Wireless and Mobile Network Architecture (New York: Wiley, 2001). He received the 1998 and 2000 Outstanding Research Awards from the National Science Council, R.O.C., and the 1998 Outstanding Youth Electrical Engineer Award from CIEE, R.O.C.



Shun-Ren Yang received the B.S.C.S.I.E. and M.S.C.S.I.E. degrees from the National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C., in 1998 and 1999, respectively. He is currently working toward the Ph.D. degree at NCTU.

His current research interests include design and analysis of a personal communications services network, computer telephony integration, mobile computing, and performance modeling.