Reducing Authentication Signaling Traffic in Third-Generation Mobile Network

Yi-Bing Lin, *Fellow, IEEE,* and Yuan-Kai Chen

***Abstract*—In universal mobile telecommunication system (UMTS), authentication functions are utilized to identify and authenticate a mobile station (MS) and validate the service request type to ensure that the user is authorized to use the particular network services. The authenticating parties are the authentication center (AuC) in the home network and the MS. In UMTS, the serving general packet radio service support node (SGSN) accesses the AuC to obtain the authentication data, and delegates the AuC to perform mutual authentication with the MS. Since the cost for accessing AuC is expensive, the SGSN may obtain an array of authentication vectors (AVs) at a time so that the number of accesses can be reduced. On the other hand, if the size K of the AV array is large, the AV array transmission from the AuC to the SGSN may be expensive. Thus, it is desirable to select an appropriate K value to minimize the authentication network signaling cost. We propose an analytic model to investigate the impact of K on the network signaling traffic, which is validated by simulation experiments. Then, we propose an automatic K-selection mechanism that dynamically selects the size of the AV array to reduce the network signaling cost. Our study indicates that the automatic K-selection mechanism effectively identifies appropriate size of the authentication vector array.**

***Index Terms*—Authentication, authentication center (AuC), general packet radio service (GPRS), home location register (HLR), mobility management, serving GPRS support node (SGSN), universal mobile telecommunication system (UMTS).**

NOMENCLATURE

ADR	Authentication data request and response.
AuC	Authentication center.
AUTN	Authentication token.
AV	Authentication vector.
CDR	Call detail record.
CK	Cipher key.
GGSN	Gateway GPRS support node.
GPRS	General packet radio service.
HLR	Home location register.
IK	Integrity key.
IMSI	International mobile subscriber identity.
MS	Mobile station.
OMC	Operation and maintenance center.
RA	Routing area.
RAND	Random number.
SGSN	Serving GPRS support node.
UAR	User authentication request and response.
UMTS	Universal mobile telecommunication system.
USIM	User service identity module.
UTRAN	UMTs terrestrial radio access network.
XRES	Expected response.

Manuscript received May 16, 2001; revised September 21, 2001; accepted October 15, 2001. The editor coordinating the review of this paper and approving it for publication is Y. M. Fang. This work was supported in part by MOE Program of Excellence Research under Contract 89-E-FA04-4, TAHOE Network, Ericsson, InterVideo, FarEastone, National Science Council under Contract NSC 90-2213-E-009-156, and by the Lee and MTI Center for Networking Research, NCTU.

The authors are with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan, R.O.C. (e-mail: liny@csie.nctu.edu.tw).

Digital Object Identifier 10.1109/TWC.2003.811171

I. INTRODUCTION

UNIVERSAL mobile telecommunication system (UMTS) is a third-generation (3G) mobile service technology evolved from general packet radio service (GPRS), which supports multimedia services to the mobile users [12], [15]–[17]. The UMTS architecture is illustrated in Fig. 1. In this architecture, the packet data services of an mobile station (MS) are provided by the serving GPRS support node (SGSN) connecting to the UMTS terrestrial radio access network (UTRAN) that covers the MS. For the discussion purpose, we refer to the area covered by an SGSN as the *SGSN service area*. The SGSN connects the MS to the external data network through the gateway GPRS support node (GGSN). Furthermore, the SGSN communicates with the home location register (HLR) and the authentication center (AuC) to receive subscriber data and authentication information of an MS. The HLR maintains the current location (SGSN number) of an MS. The AuC is used in the security data management for the authentication of subscribers. The AuC may be colocated with the HLR. More details of UMTS and GPRS can be found in [4] and [13].

An SGSN service area is partitioned into several routing areas (RAs). When the MS moves from one RA to another, a location update is performed, which informs the SGSN of the MSs current location. Note that a crossing of two RAs within an SGSN area requires an *intra-SGSN* location update, while a crossing of two RAs of different SGSN areas requires an *inter-SGSN* location update. Details of location update and mobility management modeling can be found in [5] and [6].

In UMTS, authentication function identifies and authenticates an MS, and validates the service request type to ensure that the user is authorized to use the particular network services. Specifically, authentication is performed for every location update (either inter-SGSN or intra-SGSN), call origination, and (possibly) call termination. UMTS authentication supports mutual authentication, i.e., authentication of the MS by the network and authentication of the network by the MS. The procedure also establishes a new UMTS cipher key (CK) and integrity key (IK) agreement between the SGSN and the MS. In

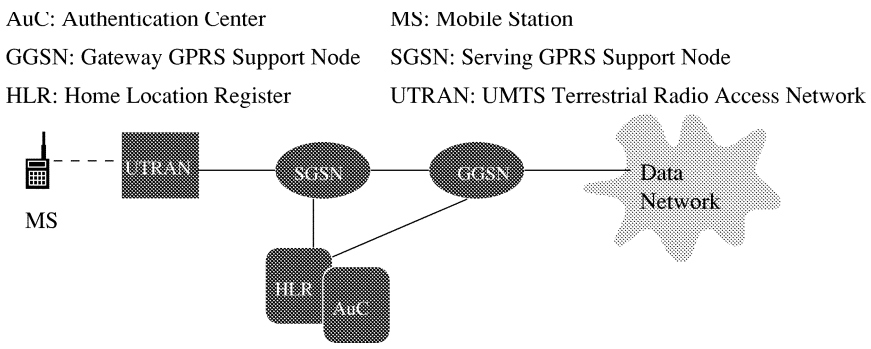

Fig. 1. Simplified UMTS architecture.

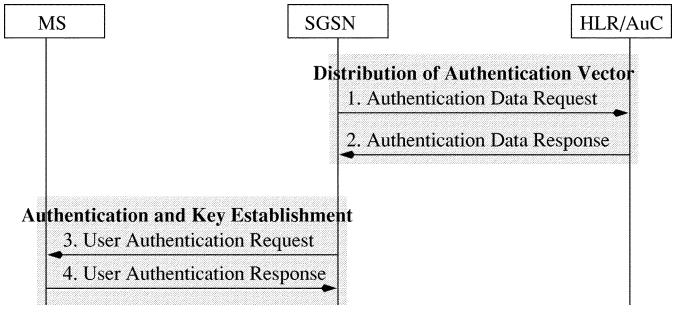

Fig. 2. UMTS authentication procedure.

UMTS authentication, the authenticating parties are HLR/AuC in the home network and the user service identity module (USIM) in the user's MS. Two major authentication procedures are described in this paper.

Distribution of Authentication Vector: This procedure distributes authentication vectors (AVs) from the HLR/AuC to the SGSN. An AV is temporary authentication data, which enables an SGSN to engage in UMTS authentication and key agreement with a particular user. We assume the following.

- The HLR/AuC trusts that the SGSN will handle authentication information securely.
- The communication path between the SGSN and HLR/AuC are adequately secure.
- The user (MS/USIM) trusts the HLR/AuC.

Authentication and Key Establishment: GSM only provides one-way authentication. In UMTS, mutual authentication is achieved by sharing a secret key between the USIM and the HLR/AuC [3]. This procedure follows a challenge/response protocol identical to the GSM subscriber authentication and key establishment protocol [13] combined with a sequence-number-based one-pass protocol for network authentication derived from ISO/IEC 9798-4 [7].

Signaling flows of above two procedures are described in the following steps (see Fig. 2).

Step 1) When an MS moves into a new SGSN area, the SGSN does not have previously stored authentication information (i.e., AVs). The SGSN invokes the distribution of authentication vector procedure by sending the `Authentication Data Request` message to the HLR/AuC. This message includes the international mobile subscriber identity (IMSI) that uniquely identifies the MS.

Step 2) Upon receipt of a request from the SGSN, the IMSI is used to identify the HLR/AuC record of the MS, and the HLR/AuC sends an ordered array of K AVs (generated based on the MS record) to the SGSN through the `Authentication Data Response` message. An AV consists of a random number RAND, an expected response XRES, a cipher key CK, an integrity key IK, and an authentication token AUTN. Each AV is good for one authentication and key agreement between the SGSN and the USIM.

The HLR/AuC may have precomputed the required AVs and retrieve these AVs from the HLR database or may compute them on demand.

The next two steps authenticate the user and establish a new pair of cipher and integrity keys between the SGSN and the USIM.

Step 3) When the SGSN initiates an authentication and key agreement, it selects the next unused authentication vector from the ordered AV array and sends the parameters RAND and AUTN (from the selected authentication vector) to the USIM through the `User Authentication Request` message.

Step 4) The USIM checks whether AUTN can be accepted and, if so, produces a response RES which is sent back to the SGSN through the `User Authentication Response` message. The SGSN compares the received RES with XRES. If they match, then the authentication and key agreement exchange is successfully completed. Note that in this mutual authentication procedure, AUTN is used by the USIM to authenticate the network, and RES/XRES is used by the network to authenticate the USIM.

In this step, the USIM also computes CK and IK using the received AUTN. During data delivery, CK and IK are utilized to perform ciphering and integrity functions in the MS. On the other hand, the SGSN retrieves CK and IK from the AV and passes them to the UTRAN for data ciphering and integrity. Details of CK and IK generation at the MS side are given in [3, Secs. 6.5, 6.6].

Note that the message names in the above descriptions are based on [3]. In [1], the actual SS7 messages sent in Steps 1 and 2 are `MAP_SEND_AUTHENTICATION_INFO` and

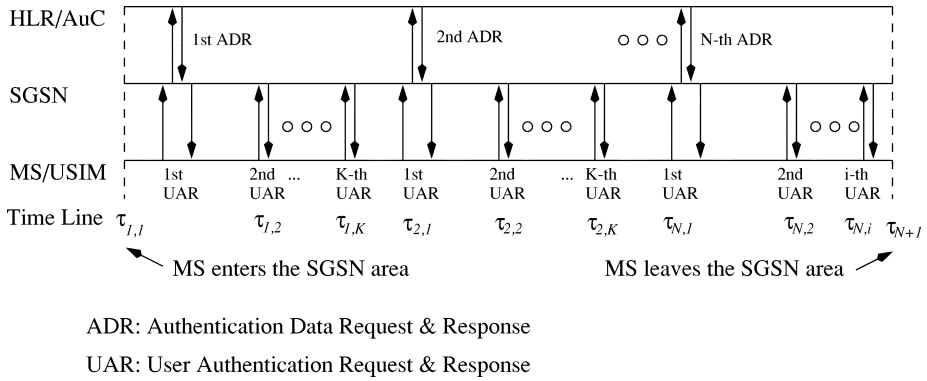

Fig. 3. Timing diagram.

`MAP_SEND_AUTHENTICATION_INFO_ack`. In [2], the messages exchanged in Steps 3 and 4 are `Authentication Request` and `Authentication Response`.

Third-generation partnership project (3GPP) TS 33.102 [3] describes a procedure to distribute authentication data from a previously visited SGSN to the newly visited SGSN. This procedure assumes that the links between SGSNs are adequately secure. In reality, it may not be true (especially when SGSNs are owned by different service providers or even different countries). Thus, we assume that when the MS moves from the old SGSN area to a new SGSN area, the authentication data stored in the old SGSN are not sent to the new SGSN. Instead, the new SGSN obtains a new AV array from the HLR/AuC through an `Authentication Request and Response` (ADR) message pair exchange (that is, Steps 1 and 2 in Fig. 2).

Consider the timing diagram in Fig. 3. Suppose that an MS enters a new SGSN area at time $\tau_{1,1}$. The MS sends a registration message to the SGSN. Since the SGSN does not have authentication information, the distribution of authentication vector procedure is executed through an ADR, where the number of AVs obtained from the HLR/AuC is K. After the SGSN has obtained the AV array from the HLR/AuC, mutual authentication is performed between the SGSN and the MS/USIM through a `User Authentication Request and Response` (UAR) message pair exchange (that is, Steps 3 and 4 in Fig. 2) by using the first AV. After $\tau_{1,1}$, the second *authentication event* (a call request or an inter-SGSN RA update) occurs at time $\tau_{1,2}$. The MS/USIM initiates the second UAR, and the SGSN uses the second AV in the array for mutual authentication. At time $\tau_{1,K}$, the last AV in the array is used for the UAR of the Kth authentication event. After $\tau_{1,K}$, the next authentication event occurs at $\tau_{2,1}$. The SGSN realizes that no AV is available, which issues the second ADR to obtain the next AV array from the HLR/AuC and then performs a UAR. For the next incoming authentication events, ADRs and UARs are executed accordingly as described above. At time τ_{N+1}, the MS leaves the SGSN area. The last authentication event before τ_{N+1} occurs at $\tau_{N,i}$ (where $1 \le i \le K$), which utilizes the ith AV in the AV array. Thus, during the period $\tau_{N+1} - \tau_{1,1}$, $(N-1)K+i$ UARs and N ADRs are performed, and at time τ_{N+1}, $K-i$ AVs are left unused in the SGSN.

We note that the ADR operation is expensive (especially when SGSN and HLR/AuC are located at different countries). Therefore, one may increase the AV array size K to reduce the number of ADRs performed when an MS is in an SGSN area. On the other hand, with a large K, the AVs may occupy too much network bandwidth for every transmission from the AuC to the SGSN. Thus, it is desirable to select an appropriate K value to minimize the authentication network signaling cost. In the next section, we propose an analytic model to investigate the impact of K on the network signaling traffic. Based on the analysis, we then propose a mechanism to select appropriate size of the authentication vector array. The notation used in this paper is listed in the Nomenclature.

II. Analytic Modeling With Fixed K

Let N be the total number of ADRs performed when the MS resides in an SGSN area. For each ADR, the number of AVs obtained from the HLR/AuC is K. Suppose that the aggregate incoming/outgoing call and registration arrivals form a Poisson process with rate λ. As we mentioned earlier, for every incoming/outgoing call and registration, a UAR is performed. For a specific period τ, let $\Theta(n, K, \tau)$ be the probability that there are n ADRs to the HLR/AuC. Note that n ADRs are performed if there are $(n-1)K+k$ $(1 \le k \le K)$ UARs in the period τ. According to the probability function of the Poisson distribution, we have

$$\Theta(n, K, \tau) = \sum_{k=1}^{K} \left\{ \frac{(\lambda\tau)^{(n-1)K+k}}{[(n-1)K+k]!} \right\} e^{-\lambda\tau}. \tag{1}$$

Let t be the period that an MS resides in an SGSN service area. That is, $t = \tau_{N+1} - \tau_{1,1}$ in Fig. 3. Suppose that t has a general distribution with the density function $f(t)$, the mean $1/\mu$, and the Laplace transform $f^*(s) = \int_{t=0}^{\infty} f(t)e^{-st}\,dt$. Let $P(n, K)$ be the probability that there are n ADRs during the MSs residence in the SGSN area. Then

$$\begin{aligned} P(n, K) &= \int_{t=0}^{\infty} \Theta(n, K, t) f(t)\,dt \\ &= \sum_{k=1}^{K} \int_{t=0}^{\infty} \left\{ \frac{(\lambda t)^{(n-1)K+k}}{[(n-1)K+k]!} \right\} e^{-\lambda t} f(t)\,dt \\ &= \sum_{k=1}^{K} \left\{ \frac{\lambda^{(n-1)K+k}}{[(n-1)K+k]!} \right\} \int_{t=0}^{\infty} t^{(n-1)K+k} \\ &\quad \times f(t) e^{-\lambda t}\,dt \end{aligned} \tag{2}$$

$$= \sum_{k=1}^{K} \left\{ \frac{\lambda^{(n-1)K+k}}{[(n-1)K+k]!} \right\} (-1)^{(n-1)K+k} \times \left[\frac{d^{(n-1)K+k} f^*(s)}{ds^{(n-1)K+k}} \right] \Bigg|_{s=\lambda} \quad (3)$$

where (3) is derived from (2) using Rule P.1.1.9 in [18]. Let $E[N]$ be the expected number of ADRs when the MS resides in an SGSN service area. Then

$$E[N] = \sum_{n=1}^{\infty} nP(n, K). \quad (4)$$

We derive $P(n, K)$ and $E[N]$ based on three SGSN residence time distributions as follows.

Gamma Distribution: If $f(t)$ is a gamma density function with mean $1/\mu$ and variance v, then

$$f^*(s) = (1 + \mu v s)^{-(1/\mu^2 v)}$$

and

$$\frac{d^l f^*(s)}{ds^l} = (-\mu v)^l \left[\prod_{j=0}^{l-1} \left(\frac{1}{\mu^2 v} + j \right) \right] (1 + \mu v s)^{-(1/\mu^2 v + l)}. \quad (5)$$

We are particularly interested in the gamma distribution. It has been shown that the distribution of any positive random variable can be approximated by a mixture of gamma distributions (see [8, Lemma 3.9]). One may also measure the SGSN residence times of an MS in a real PCS network and the measured data can be approximated by a gamma distribution as the input to our analytic model. It suffices to use the gamma distributions with different shape and scale parameters to represent different SGSN residence time distributions [14].

From (5), (3) is rewritten as

$$P(n, K) = \sum_{k=1}^{K} \left\{ \frac{(\lambda \mu v)^{(n-1)K+k}}{[(n-1)K+k]!} \right\} \times \left\{ \left[\prod_{j=0}^{(n-1)K+k-1} \left(\frac{1}{\mu^2 v} + j \right) \right] \times (1 + \mu v \lambda)^{-[1/\mu^2 v + (n-1)K + k]} \right\}. \quad (6)$$

Hyper-Erlang Distribution: Another distribution that has very general approximation capability to the probability distribution of any nonnegative random variable is the hyper-Erlang distribution [6], [8] with the mean

$$E[t] = \frac{1}{\mu} = \sum_{i=1}^{I} \left(\frac{\beta_i m_i}{\mu_i} \right)$$

where m_i is a positive integer,

$$\beta_i \geq 0, \text{ and } \sum_{i=1}^{I} \beta_i = 1.$$

The Laplace transform of the hyper-Erlang distribution is

$$f^*(s) = \sum_{i=1}^{I} \beta_i \left(1 + \frac{s}{m_i \mu_i} \right)^{m_i}. \quad (7)$$

From (7), (3) is rewritten as

$$P(n, K) = \sum_{i=1}^{I} \beta_i \left\{ \left[\frac{(m_i - 1)! \mu_i^{m_i} (m_i \lambda)^{(n-1)K}}{(m_i \lambda + \mu_i)^{(n-1)K + m_i}} \right] \times \left[\sum_{k=1}^{K} \binom{(n-1)K + k + m_i - 1}{m_i - 1} \left(\frac{m_i \lambda}{m_i \lambda + \mu_i} \right)^k \right] \right\}. \quad (8)$$

The hyper-Erlang distribution is appropriate to approximate t that is either Coxian or SOHYP distributed or more generally phase-type distributed [9], [10].

Exponential Distribution: Let $\mu^2 v = 1$ in (6) or let $I = 1$, $m_1 = 1$ in (8). In this case, t is exponentially distributed, and (3) is rewritten as

$$P(n, K) = \left(\frac{\lambda}{\lambda + \mu} \right)^{(n-1)K} \left[1 - \left(\frac{\lambda}{\lambda + \mu} \right)^K \right].$$

Also, (4) is rewritten as

$$E[N] = \frac{1}{1 - \left(\frac{\lambda}{\lambda + \mu} \right)^K}. \quad (9)$$

The exponential distribution may not capture the details of the real SGSN residence times for a mobile network. However, this distribution is good for mean value analysis, which does capture the trend for the performance of a system [11].

Let $C(K)$ be the total message transmission cost for ADRs when an MS resides in an SGSN area. Then

$$C(K) = E[N] \times (K + 2\alpha) \quad (10)$$

where α represents the cost for an SS7 message overhead normalized by the cost of an AV transmission and the processing time for generating the AV through the authentication procedure. In the right-hand side of (10), this overhead is considered for the `Request` and `Response` message pair exchanged in an ADR. From (9) and (10), the total ADR transmission cost for exponential SGSN residence times is expressed as

$$C(K) = \frac{K + 2\alpha}{1 - \left(\frac{\lambda}{\lambda + \mu} \right)^K}. \quad (11)$$

III. Performance of Fixed-K Mechanism

This sections shows how the expected number $E[N]$ and cost $C(K)$ of ADR are affected by the AV array size K. We say that the AV array size is selected by the *fixed-K mechanism* if the selected K value is fixed throughout an MS's lifetime. Our analytic model in the previous section can be used to compute $E[N]$ and $C(K)$ for the fixed-K mechanism. The analytic results are compared with the simulation experiments as shown in Figs. 4 and 6. In both figures, the dashed curves represent the analytic results, and the symbols $\diamond$, $\bullet$, $\circ$, and $*$ represent the simulation results. These figures indicate that the analytic results are consistent with the simulation experiments. Fig. 4 plots $E[N]$ against K with various UAR (authentication event) arrival rates λ. The SGSN residence times are assumed to be exponentially

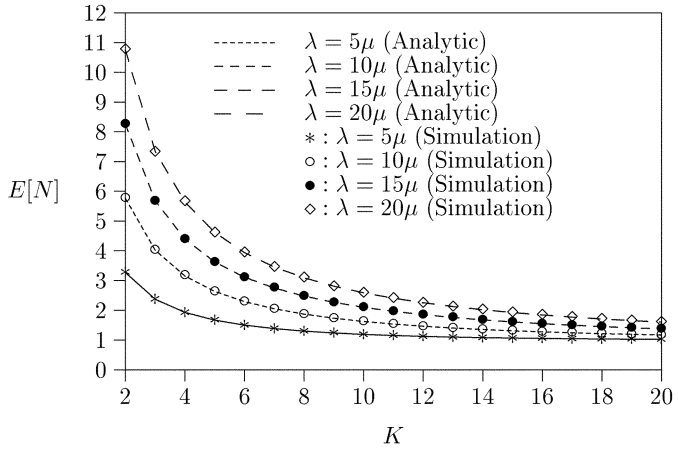

Fig. 4. Effect of the UAR arrival rate λ (exponential SGSN residence times with mean $1/\mu$).

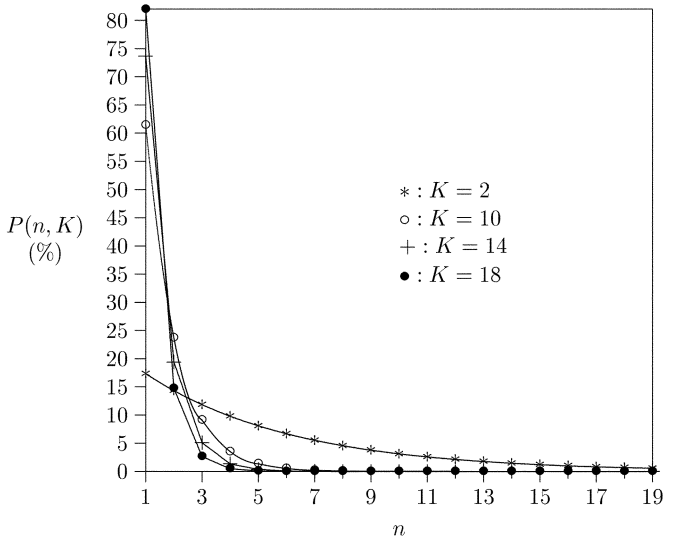

Fig. 5. $P(n, K)$ probability distribution (exponential residence times; $\lambda = 10\mu$).

distributed with mean $1/\mu$. It is obvious that $E[N]$ is a decreasing function of K. The figure indicates that for $5\mu < \lambda < 20\mu$ and $K > 10$, $E[N]$ is only insignificantly reduced by increasing K. Fig. 5 plots the $P(n, K)$ probability distribution, where $\lambda = 10\mu$. The figure shows that the shapes of $P(n, K)$ distribution are similar for $K > 10$, which are very different from the shape when $K = 2$. This observation is consistent with our statement for Fig. 4; i.e., the $E[N]$ values are roughly the same for large K values, and increasing K does not improve the $E[N]$ performance.

Fig. 6 shows the effect of the variance v for the gamma SGSN residence time distribution. The mean SGSN residence time is $1/\mu$, and the UAR arrival rate is $\lambda = 10\mu$. The figure shows that as v increases, $E[N]$ increases. This phenomenon is explained as follows. As v increases, more short SGSN residence times and more long SGSN residence times are observed. For short residence times, it is likely that $\tau_{N+1} < \tau_{1,2}$ in Fig. 3 and, thus, $N = 1$ is expected. Fig. 7 indicates that $P(1, K)$ significantly increases as v increases (especially when $v < 1/\mu^2$).

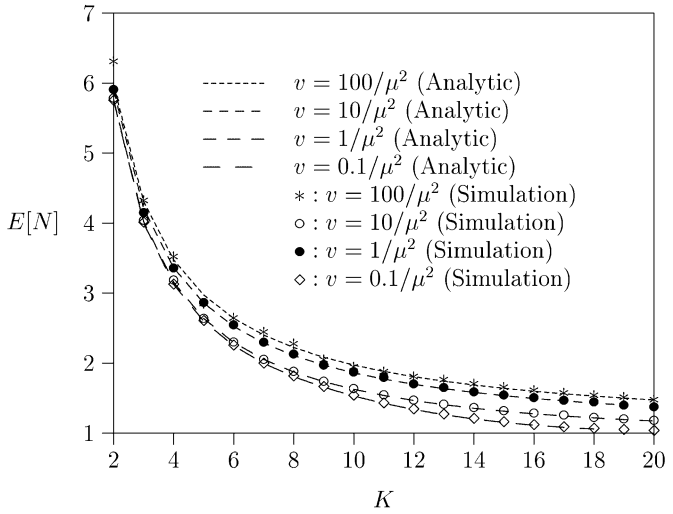

Fig. 6. Effect of the variance v for gamma SGSN residence times ($\lambda = 10\mu$).

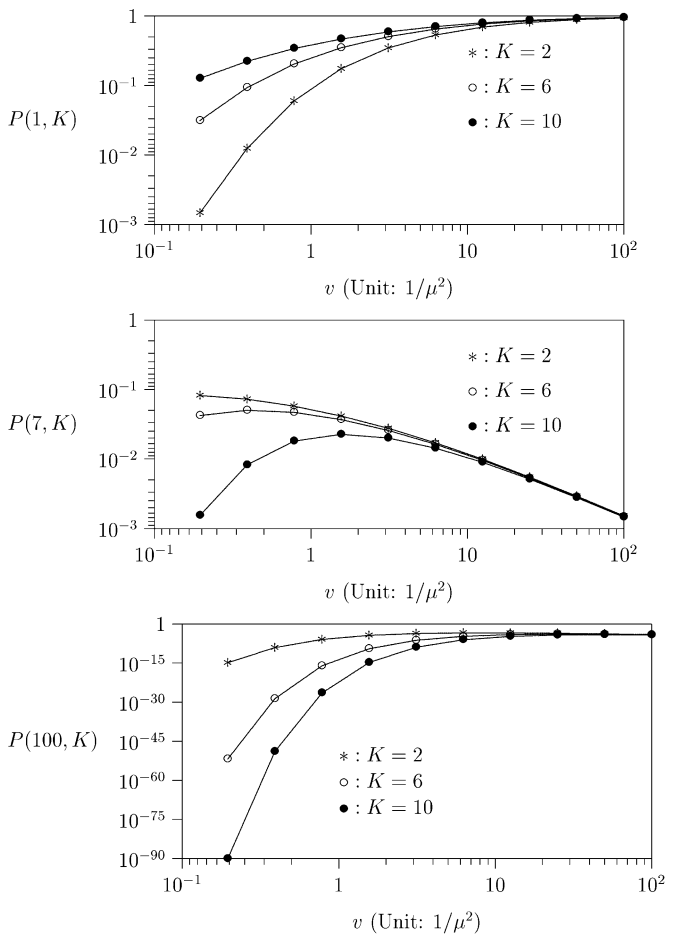

Fig. 7. Effect of the variance v for gamma SGSN residence time distribution ($\lambda = 20\mu$).

In this case, it implies that after the SGSN has obtained the AV array, only one AV is used for the first UAR. The $K-1$ AVs are wasted. Similarly, as v increases, more long residence times are observed, and a larger ADR number is expected [see $P(100, K)$ in Fig. 7]. Thus, $E[N]$ is an increasing function of v. The above discussion leads to the conclusion that for MSs with irregular

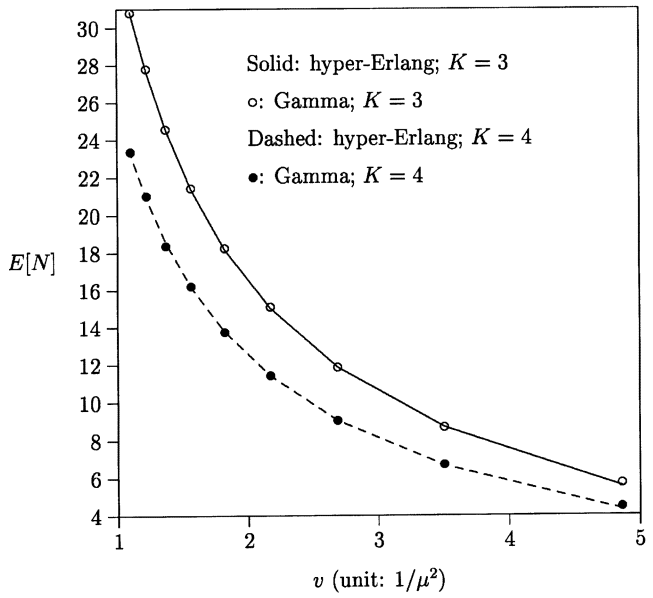

Fig. 8. Effect of the SGSN residence time distribution.

moving patterns, more network signaling traffic for authentication will be observed. As a final remark for Fig. 7, more residence times with lengths close to its mean value are observed for smaller variance v [see $P(7, K)$ in Fig. 7].

Fig. 8 plots $E[N]$ when the SGSN residence time distributions are gamma and hyper-Erlang. In our example, the hyper-Erlang distribution has parameters $I = 2$, $m_1 = m_2 = 1$, $\mu_2 = \gamma\mu_1$, $\beta_1 + \beta_2 = 1$, and $\lambda = \theta\mu_1$. From the above parameters, we have

$$\mu = \left[\frac{\gamma}{1+(\gamma-1)\beta_1}\right]\mu_1, \quad v = \left\{\frac{1+(\gamma^2-1)\beta_1}{[1+(\gamma-1)\beta_1]^2\mu^2}\right\}, \quad \lambda = \theta\left[\frac{1+(\gamma-1)\beta_1}{\gamma}\right]\mu. \tag{12}$$

In Fig. 8, we set $0.1 \le \beta_1 \le 0.9$, $\theta = 100$ and $\gamma = 20$. The μ, v and λ values for the case of gamma distribution are assigned based on (12). Note that if $\gamma > 1$, then as β_1 decreases, v increases and λ decreases. Thus, the $E[N]$ curves are decreasing (because λ decreases as v increases), which are different from what we observed in Fig. 6, where λ is fixed. Fig. 8 indicates that $E[N]$ curves are similar for gamma and hyper-Erlang distributions. In the remainder of this paper, we only consider the gamma SGSN residence time distribution for our performance study. The results for hyper-Erlang residence times are similar and will not be presented.

By assuming $\alpha = 1$ in (10), Fig. 9 plots $C(K)$ against K with various UAR arrival rates λ. The SGSN residence times are exponentially distributed in Fig. 9(a), and are gamma distributions with various variance values in Fig. 9(b). Since $E[N]$ is a decreasing function of K (see Figs. 4 and 6), the product $KE[N]$ [and thus $C(K)$] results in concave curves shown in Fig. 9. That is, as K increases, $C(K)$ decreases and then increases. Let K^* be the optimal value that minimizes the cost C. For exponential SGSN residence time distribution, K^* can be obtained by differentiating (11). That is K^* can be approximated by

$$K^* = \lceil x \rceil, \qquad \text{where } \left(\frac{\lambda+\mu}{\lambda}\right)^x = (x+2)\ln\left(\frac{\lambda}{\lambda+\mu}\right)+1.$$

Fig. 9(a) shows that K^* increases as λ increases. This result is consistent with our intuition that if there are more UARs, then more authentication vectors should be delivered from the AuC to the SGSN at a time. Fig. 9(b) shows that K^* is only insignificantly affected by the variance of the SGSN residence time distribution. Specifically, $5 \le K^* \le 6$ for $0.1/\mu^2 \le v \le 100/\mu^2$.

IV. Automatic K-Selection Mechanism

In 3GPP TS 29.002 [1], $K = 5$ is recommended for fixed-K mechanism. However, our study in the previous section indicates that the optimal value K^* is affected by the traffic of UARs. This section proposes a per-user *automatic K-selection mechanism* that dynamically selects the K value based on the SGSN residence time distribution and the UAR traffic pattern of an MS. The automatic K-selection mechanism can be implemented in MS or in AuC. If the mechanism is exercised at the MS, then every time the MS moves into a new SGSN service area, it forwards the new K value to the AuC through the registration procedure. If the mechanism is exercised by the AuC, then the number of the UARs must be provided by the old SGSN when the MS moves to the new SGSN or when the MS detaches from the network. The old SGSN can either supplies the number of the UARs (the number of calls and intra-SGSN registrations when the MS resides in the old SGSN) to the HLR/AuC when the `Cancel Location` message pair is exchanged between the old SGSN and the HLR (see [4, Sec. 6.9] or [13, Sec. 2, Ch. 9]), or the SGSN can generate *call detail records* (CDRs) for calls and intra-SGSN location updates (see [13, Sec. 6, Ch. 18]). The CDRs are sent to either a billing gateway or an operation and maintenance center (OMC), and the OMC forwards the related information to the HLR/AuC.

Let $K(j)$ be the K value selected for the jth *iteration* (i.e., when the MS resides in the jth SGSN area). The automatic K-selection mechanism is described as follows.

Initialization: When the user subscribes to the UMTS service, an initial K value is assigned (e.g., $K(1) = 5$ as 3GPP suggested). Then, every time the MS enters an SGSN area, the following two steps are executed.

Measurement Step: During the jth SGSN residence period, the number M of UARs are counted.

Decision Step: When the MS leaves the jth SGSN area, we determine if the K value should be adjusted to $K_1 = K(j) - 1$ (decrement $K(j)$ by one), $K_2 = K(j)$ (no change), $K_3 = K(j)+1$ (increment $K(j)$ by one). We use (10) as the heuristic to compute three costs as follows:

$$c_i = \left\lceil \frac{M}{K_i} \right\rceil \times (K_i + 2\alpha), \qquad \text{for } i = 1, 2, 3. \tag{13}$$

Based on (13), $K(j+1)$ is selected as

$$K(j+1) = K_l \qquad \text{where } 1 \le l \le 3, \text{ and } c_l = \min_{1 \le i \le 3} c_i. \tag{14}$$

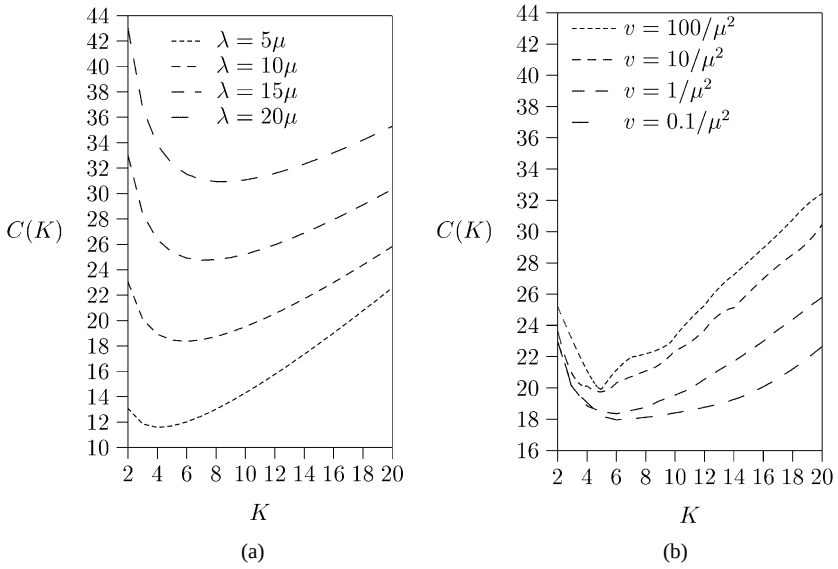

Fig. 9. Cost of ADRs (the mean of the SGSN residence times is $1/\mu$). (a) Exponential residence times. (b) Gamma residence times ($\lambda = 10\mu$).

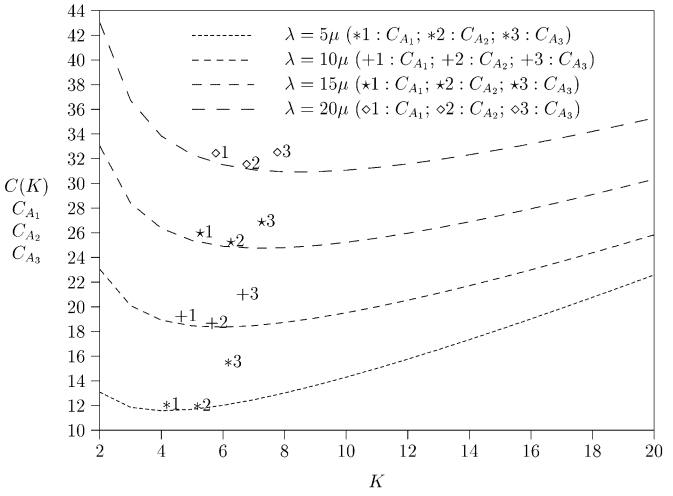

Fig. 10. Performance of the K-selection mechanism: Exponential SGSN residence times with mean $1/\mu$.

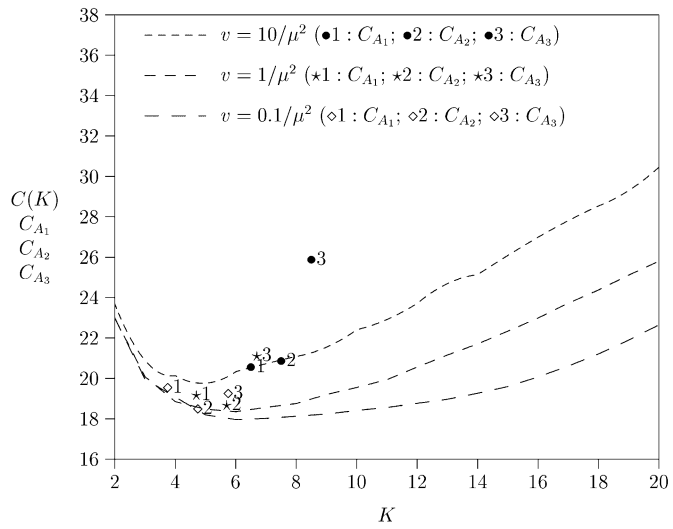

Fig. 11. Performance of the K-selection mechanism: Gamma SGSN residence times ($\lambda = 10\mu$).

From the above discussion, it seems reasonable that when the MS enters the $j+1$st SGSN area, the AV array size for ADRs is assigned the value $K(j+1)$ computed in (14). However, from our experiments (to be elaborated later), this is not the best choice. Instead, the best performance is achieved when a smaller K value is selected. Define three AV array size assignments for the $j+1$st iteration as follows.

A_1: AV array size is $\max(K(j+1)-2, 1)$.
A_2: AV array size is $\max(K(j+1)-1, 1)$.
A_3: AV array size is $K(j+1)$.

As we will see later, A_2 will yield the best performance in most cases. Also note that choosing any AV array sizes larger than $K(j+1)$ always yields performance worse than that for A_3 and these scenarios will not be presented in this paper.

Figs. 10–12 show the performance of A_1, A_2, and A_3, and compares them with the fixed-K mechanism. We assume that $\alpha = 1$ in (13). Note that the actual α value depends on the implementation of the AV data structure in a mobile network. However, the conclusions presented in this paper holds for any

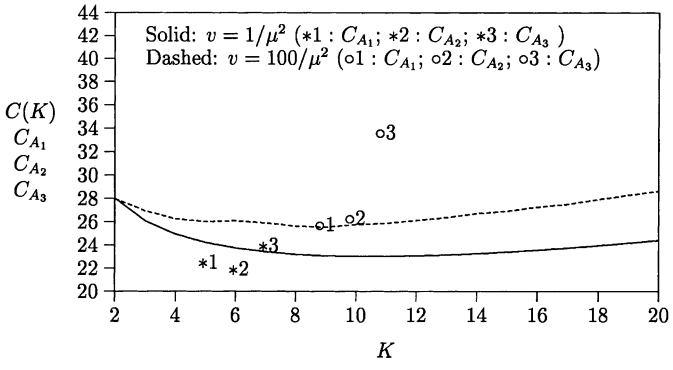

Fig. 12. Performance of the K-selection mechanism: Mixed UAR arrival patterns.

α values. Let C_{A_1}, C_{A_2}, and C_{A_3} be the ADR transmission cost for A_1, A_2, and A_3, respectively. As defined before, let $C(K)$ be the cost of fixed-K mechanism with the array size K. Three

scenarios with various SGSN residence time distributions and UAR traffic patterns are considered.

Exponential SGSN Residence Time Distribution: Fig. 10 plots C_{A_1}, C_{A_2}, C_{A_3}, and $C(K)$ for $2 \leq K \leq 20$. The SGSN residence times have exponential distribution with mean $1/\mu$. The UAR arrival rates are $\lambda = 5\mu$, 10μ, 15μ, and 20μ, respectively. The figure indicates that, in most cases, the costs for the automatic K-selection mechanisms are close to each other and

$$C_{A_3} > C_{A_1} > C_{A2} \simeq C(K^*) \tag{15}$$

where K^* is the optimal K value that minimizes the cost C in the fixed-K mechanism.

Gamma SGSN Residence Time Distribution: Fig. 11 plots the cost curves, where the UAR arrival rate is $\lambda = 20\mu$, and the SGSN residence times have gamma distribution with variances $v = 10/\mu^2$, $1/\mu^2$, and $0.1/\mu^2$, respectively. Inequality (15) also holds for most cases in this scenario. Although the cost for A_2 is slightly greater than that for A_1 when v is large, the difference is insignificant and can be ignored. However, A_3 becomes less accurate in terms of capturing the K^* value. The cost of A_2 is very close to $C(K^*)$ as in the case for exponential SGSN residence times.

Mixed UAR Arrival Patterns: Fig. 12 plots the cost curves where the UAR arrivals have mixed patterns. In this figure, the solid curve represents the case where the SGSN residence times have an exponential distribution with mean $1/\mu$. The dashed curve represents the case where the residence time distribution is gamma with mean $1/\mu$ and variance $v = 100/\mu^2$. In both curves, the UAR arrivals are Poisson with rate $\lambda = 5\mu$ for the first 10 000 UARs, then the rate changes to $\lambda = 10\mu$ for the second 10 000 UARs, $\lambda = 15\mu$ for the third 10 000 UARs, and $\lambda = 20\mu$ for the last 10 000 UARs.

The figure indicates that with the mixed UAR traffic patterns, A_1 and A_2 may outperform the fixed-K mechanism for all K values (see the solid curve).

The above analysis indicates that the performance of the automatic K-selection mechanism (particularly A_2) is close to (and may be better than) the performance of the fixed-K mechanism with the optimal K value.

V. Conclusion

In the mobile network, authentication is exercised for every location update and every call event. In UMTS, mutual authentication is performed between the MS and the SGSN. In order to carry out authentication, the SGSN obtains authentication data from the HLR/AuC. Since the cost for accessing AuC is expensive, the SGSN may obtain an array of AVs at a time so that the number of accesses can be reduced. On the other hand, if the size K of the AV array is large, every AV array transmission from the AuC to the SGSN may be expensive. Thus, it is desirable to select an appropriate K value to minimize the authentication network signaling cost. We proposed an analytic model to investigate the fixed-K mechanism (the selected K value is fixed throughout an MS's lifetime). The analytic results were validated against the simulation experiments. We observed the following results.

- Increasing K will decrease the expected number $E[N]$ of accesses from the SGSN to the HLR/AuC. However, when K is large, increasing K only insignificantly reduces $E[N]$.
- When the variance of the SGSN residence times increases, $E[N]$ increases.
- Let $C(K)$ be the cost for the SGSN to access the authentication data of HLR/AuC [see (10)]. $C(K)$ is a concave curve, and there exists an optimal K value that minimizes $C(K)$.
- The optimal K value increases as the authentication event (UAR) rate λ increases.

Study on the fixed-K mechanism suggests that K should be adjusted based on the authentication traffic so that the cost $C(K)$ is reduced. We proposed a per-user automatic K-selection mechanism that dynamically selects the K value based on the SGSN residence time distribution and the UAR traffic pattern of an MS. Our study showed that the automatic K-selection mechanism effectively identifies appropriate K value to reduce the network signaling cost. Specifically, our study indicated that the performance of the automatic K-selection mechanism (particularly A_2) is close to (and may be better than) the performance of the fixed-K mechanism with the optimal K value.

Acknowledgment

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

References

[1] 3GPP, 3rd Generation Partnership Project; Technical Specification Core Network; Mobile Application Part (MAP) Specification (Release 1999), 2000, Technical Specification 3G TS 29.002 V3.7.0 (2000-12).

[2] 3GPP, 3rd Generation Partnership Project; Technical Specification Group Core Network; Mobile Radio Interface Layer 3 Specification; Core Network Protocols—Stage 3 for Release 1999, 2000, 3G TS 24.008 version 3.6.0 (2000-12).

[3] 3GPP, 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; 3G Security; Security Architecture, 2000, Technical Specification 3G TS 33.102 V3.7.0 (2000-12).

[4] 3GPP, 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; General Packet Radio Service (GPRS); Service Descripton; Stage 2, 2000, Technical Specification 3G TS 23.060 version 3.6.0 (2001-01).

[5] I. F. Akyildiz, J. McNair, J. S. M. Ho, H. Uzunalioglu, and W. Wang, "Mobility management in next generation wireless systems," *Proc. IEEE*, vol. 87, pp. 1347–1384, Aug. 1999.

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

[7] ISO/IEC, "Information technology-security techniques—Entity authentication—Part 4: Mechanisms using a cryptographic check function," ISO/IEC, Tech. Rep. ISO/IEC 9798-4, 1999.

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

[9] L. Kleinrock, *Queueing Systems: Volume I—Theory*. New York: Wiley, 1976.

[10] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia, PA: SIAM, 1999.

[11] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik, *Quantitative System Performance*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[12] B. Li *et al.*, "QoS-enabled voice support in the next-generation Internet: Issues, existing approaches and challenges," *IEEE Commun. Mag.*, vol. 38, no. 4, pp. 54–61, Apr. 2000.

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

[14] 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.
[15] UMTS Forum. (2000) Enabling UMTS/third generation services and applications. Tech. Rep. 11, UMTS. [Online]. Available: www.umts-forum.org
[16] UMTS Forum. (2000) Shaping the mobile multimedia future—An extended vision from the UMTS forum. Tech. Rep. 10, UMTS. [Online]. Available: www.umts-forum.org
[17] UMTS Forum. (2000) The UMTS third generation market—Structuring the service revenues opportunities. Tech. Rep. 9, UMTS. [Online]. Available: www.umts-forum.org
[18] E. J. Watson, *Laplace Transforms and Applications*. Cambridge, MA: Birkhauserk, 1981.

Yi-Bing Lin (M'96–SM'96–F'03) received the B.S.E.E. degree from National Cheng Kung University, Taiwan, 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, Bell Communications Research (Bellcore), Morristown, NJ. In 1995, he was appointed Professor of computer science and information engineering (CSIE), National Chiao Tung University, Hsinchu, Taiwan, R.O.C. (NCTU). In 1996, he was appointed Deputy Director of Microelectronics and Information Systems Research Center, NCTU. From 1997 to 1999, he was elected as Chairman of CSIE, NCTU, Hsinchu, Taiwan, R.O.C. He is an Editor for *Computer Networks*, an Area Editor of ACM *Mobile Computing and Communication Review*, a columnist of ACM Simulation Digest, an Editor of the *International Journal of Communications Systems*, an Editor of ACM/Baltzer *Wireless Networks*, an Editor of *Computer Simulation Modeling and Analysis*, an Editor of the *Journal of Information Science and Engineering*, and Guest Editor for the ACM/Baltzer MONET Special Issue on Personal Communications. He is the coauthor of *Wireless and Mobile Network Architecture* (New York: Wiley, 2001). 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 the IEEE NETWORKING, an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Editor of the IEEE *Personal Communications Magazine*, a Guest Editor for IEEE TRANSACTIONS ON COMPUTERS: Special Issue on Mobile Computing, and a Guest Editor for IEEE *Communications Magazine*: Special Issue on Active, Programmable, and Mobile Code Networking. He is Program Chair for the 8th Workshop on Distributed and Parallel Simulation, General Chair for the 9th Workshop on Distributed and Parallel Simulation, and Program Chair for the 2nd International Mobile Computing Conference. He received the 1998 and 2000 Outstanding Research Awards from National Science Council, R.O.C., and the 1998 Outstanding Youth Electrical Engineer Award from CIEE, R.O.C. He is an Adjunct Research Fellow of Academia Sinica.

Yuan-Kai Chen received the B.S.C.S.I.E. and M.S.C.S.I.E. degrees from National Chiao Tung University, Hsinchu, Taiwan, R.O.C., in 1989 and 1991, respectively. He is currently working toward the Ph.D. degree in the Department of Computer Science and Information Engineering, National Chiao Tung University.

In 1991, he joined the Telecommunication Laboratories, Chunghwa Telecom Company, Ltd., Taiwan, R.O.C. His current research interests include design and analysis of personal communications services network, mobile computing, and UMTS.