Reducing Credit Re-authorization Cost in UMTS Online Charging System

Sok-Ian Sou, Yi-Bing Lin, *Fellow, IEEE*, and Jeu-Yih Jeng

***Abstract*—During an online charging General Packet Radio Service (GPRS) session, a number of mid-session events, such as changes of Quality of Service (QoS), could dynamically affect the rating of the in-progress service. When such events occur, the GPRS support node needs to re-authorize the granted credit units with the Online Charging System (OCS). This paper proposes a threshold-based scheme that utilizes a threshold parameter δ to reduce the signaling traffic for the credit re-authorization procedure. By selecting an appropriate δ value, the signaling overhead in the OCS can be significantly reduced while the inaccuracy of the credit information insignificantly increases. The mobile operator can choose appropriate parameter values in the threshold-based scheme based on our study.**

***Index Terms*—Credit reservation, streaming services, online charging system (OCS), UMTS.**

I. Introduction

UNIVERSAL Mobile Telecommunications System (UMTS) supports real-time IP multimedia services through IP Multimedia Subsystem (IMS) and General Packet Radio Service (GPRS) transport network which consists of GPRS Support Nodes (GSNs) such as Serving GSNs (SGSNs) and Gateway GSNs (GGSNs) [7]. The IP-based GPRS/IMS services specify critical charging requirements that impose flexible mobile billing schemes (e.g., time-based, volume-based, content-based) [3], [4], [11]. Therefore, 3GPP Release 6 introduces the convergent charging solution that elaborates on an IP-based Online Charging System (OCS) [1], [2].

The Diameter Credit Control protocol is used for communications between the GGSN and the OCS [2]. To start an online GPRS session, the GGSN first sends a Credit Control Request (CCR) to the OCS. The OCS determines the rating and then allocates the granted credit units (based on the charge for time units or data volume units) to the GGSN by sending the Credit Control Answer (CCA). During the GPRS session, a number of mid-session events, such as change of *Quality of Service* (QoS) or change of SGSN, could dynamically affect the rating of the in-progress service. When such events occur, the GGSN needs to re-authorize the granted credit units . If the change of QoS occurs frequently, the signaling load incurred by the re-authorization procedure will increase heavily. This paper proposes a cost reduction method for credit re-authorization procedure and conducts both simulations and analytic models to evaluate the performance of the proposed method.

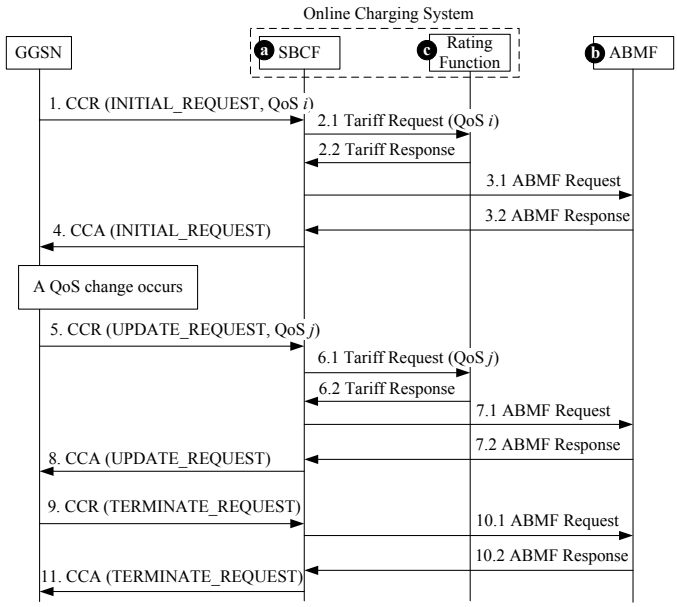


Fig. 1. Message flow for credit re-authorization procedure.

II. Online Charging System for GPRS Sessions

In the OCS, the Session Based Charging Function (SBCF; Fig. 1 (a)) is responsible for online charging of network bearer and user sessions [1]. Through the Rc reference point, the SBCF interacts with the Account Balance Management Function (ABMF; Fig. 1 (b)) to query and update the user's account. The ABMF keeps the subscriber's account data and controls the account balance. At the time of writing, the message exchanges in the Rc reference point are not defined. In this paper, we use ABMF Request and ABMF Response to represent the message exchanges between SBCF and ABMF. The Rating Function (Fig. 1 (c)) handles a wide variety of rateable instances such as data volume and session connection time. By exchanging the Diameter Tariff Request/Response and the Price Request/Response message pairs, the SBCF

Manuscript received April 16, 2007; accepted July 5, 2007. The associate editor coordinating the review of this paper and approving it for publication was S. Shen. This work was sponsored in part by NSC Excellence project NSC 95-2752-E-009-005-PAE, NSC 95-2218-E-009-201-MY3, NSC 95-2221-E-009-024, NSC 95-2219-E-009-010, NSC 95-2219-E-009-019, Intel, Chung Hwa Telecom, IIS/Academia Sinica, ITRI/NCTU Joint Research Center and MoE ATU.

S.-I. Sou is with the Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan, R.O.C. (e-mail: sisou@mail.ncku.edu.tw). The work of S.-I. Sou was supported by the Media Tek Fellowship.

Y.-B. Lin is with the Department of Computer Science, National Chiao Tung University, Hsinchu 30010, Taiwan, R.O.C. (e-mail: liny@csie.nctu.edu.tw). Y.-B. Lin is also with the Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan.

J.-Y. Jeng is with the Information Technology Laboratory of Telecommunication Laboratories, Chunghwa Telcom Co., Ltd., R.O.C. (e-mail: jyjeng@cht.com.tw).

Digital Object Identifier 10.1109/TWC.2008.070403.

interacts with the Rating Function to determine the price and tariff of the requested service.

A. Credit Re-authorization Procedure

Consider a scenario where a mobile user is viewing a streaming video through the GPRS network. The streaming services are charged according to the amount of time units and the related QoS provisioned (i.e., the bandwidth allocated to the bearer session). Due to user mobility between different UMTS coverage ares, and depending on the work load of the radio network, the QoS of the streaming session may change from time to time. In terms of bandwidth allocated, assume that there are N QoS classes for the GPRS bearer session. For $1 \leq i \leq N$, let α_i be the number of credit units charged for every time unit in a class i session. In each credit reservation, the OCS grants $\alpha_i \tau_g$ credit units (which can last for τ_g time units) to the GGSN for the streaming session. Note that the bandwidth allocated to a class i session is typically proportioned to the charge (i.e., α_i). Whenever the QoS of the GPRS bearer changes, the credit re-authorization procedure illustrated in Fig. 1 is executed with the following steps.

Step 1. The GGSN sends the INITIAL_REQUEST `CCR` message to the OCS. This message indicates the QoS parameter (e.g., QoS class i) of the session.

Step 2. When the OCS receives the `CCR` message, the SBCF sends the `Tariff Request` message, including the QoS parameter in the *Service-Information* field, to the Rating Function. The Rating Function replies with the `Tariff Response` message to indicate the applicable tariff α_i of this session [1].

Step 3. Based on the received tariff information, the SBCF grants $\alpha_i \tau_g$ credit units to the session and reserves $\alpha_i \tau_g$ credit units in the subscriber's account by exchanging the `ABMF Request` and `Response` message pair with the ABMF.

Step 4. After the reservation is performed, the OCS acknowledges the GGSN with the `CCA` message including the granted credit units ($\alpha_i \tau_g$) and the trigger event type (i.e., "CHANGE_IN_QOS"). This message indicates that the GGSN should trigger credit re-authorization procedure when the QoS change occurs. Then, the GGSN starts the streaming service.

Step 5. When the serving QoS of the session is changed from class i to class j (because, e.g., the mobile user moves to another base station with different bandwidth capacity), the credit re-authorization procedure is executed. The GGSN suspends service delivery and sends a UPDATE_REQUEST `CCR` message to the OCS. This CCR message includes the *Reporting-Reason* field with value "RATING_CONDITION_CHANGE" and the *Trigger-Type* field with value "CHANGE_IN_QOS". The GGSN also reports that $\alpha_i(\tau_g - \tau_u)$ credit units have been consumed and requests new credit units based on the QoS parameter (i.e., QoS class j).

Step 6. Upon receipt of the `CCR` message from the GGSN, the SBCF calculates the remaining credit units for the bearer session and revaluates the rating by exchanging `Tariff Request` and `Response` messages. The `Tariff Response` message includes the new tariff α_j of the session.

Step 7. Based on the old tariff α_i and new tariff α_j of the bearer session, the SBCF debits $\alpha_i(\tau_g - \tau_u)$ credit units and then reserves extra $\alpha_j \tau_g$ credit units with the ABMF.

Step 8. The OCS acknowledges the GGSN with the `CCA` message to indicate that $\alpha_j \tau_g$ credit units have been reserved (which lasts for τ_g time units). The GGSN resumes the service delivery. Note that Steps 5-8 may be repeated whenever the QoS is changed or when the allocated credit units are depleted.

Steps 9-11. When the video streaming service is complete, the GGSN terminates the session. The GGSN reports the used credit units to the OCS by sending the TERMINATE_REQUEST `CCR` message. The SBCF calculates the consumed credit units and instructs the ABMF to debit the user account. Finally, the OCS acknowledges the GGSN with a `CCA` message.

For the discussion purpose, the above re-authorization procedure (i.e., Steps 6 and 7) is referred to as the "basic" scheme.

B. Threshold-based Scheme for Credit Re-authorization Procedure

In the basic scheme, the ABMF message exchanges (see Step 7 in Fig. 1) can be omitted if the remaining credit units are large enough to accommodate the new reservation. Based on this observation, we propose a "threshold-based" scheme that utilizes a threshold parameter δ to reduce the signaling cost incurred by the re-authorization procedure between the SBCF and the ABMF. In this scheme, the SBCF determines whether to interact with the ABMF or not based on δ as follows:

Step 1. As described in Step 6 of Fig. 1, the SBCF retrieves the old tariff α_i and new tariff α_j for the bearer session from the Rating Function.

Step 2a. If $\alpha_i \tau_u \geq \delta \alpha_j \tau_g$, the SBCF directly allocates $\alpha_i \tau_u$ credit units to the GGSN. The ABMF message exchange (i.e., Step 7 in Fig. 1) is skipped.

Step 2b. Otherwise (i.e., $\alpha_i \tau_u < \delta \alpha_j \tau_g$), the SBCF allocates $\alpha_j \tau_g$ credit units to the GGSN by executing Step 7 in Fig. 1. That is, the SBCF debits $\alpha_i(\tau_g - \tau_u)$ credit units to the ABMF and reserves $\alpha_j \tau_g$ credit units.

In the OCS, a mobile operator (or a user) can check the account balance at anytime. When there is no in-progress session, the OCS accurately reports the account balance of the user. On the other hand, when credit units are reserved for some in-progress sessions, the account balance reported by the OCS may not be up-to-date because some reserved credit units might already be used. When a "balance check" occurs, the OCS reports the account balance including the reserved credit units. This reported value may be larger than the actual balance. Denote C as the inaccuracy of credit information, which is the difference between the balance stored in the OCS (including the reserved credit units) and the actual balance (excluding the credit units already consumed by the GGSN). In other words, C is the amount of credit consumed by the GGSN between when the previous ABMF message exchange occurs and when the balance check occurs. In the threshold-based

scheme, the inaccuracy of credit information C may be larger than the basic scheme because it skips some ABMF message exchanges. Therefore, it is important to select appropriate τ_g and δ values to optimize the performance of the threshold-based scheme in terms of the following output measures:

- M: the expected number of ABMF message exchanges for a GPRS session. The larger the M value, the higher the ABMF message overhead.
- C: the expected inaccuracy of credit information when a balance check occurs during an in-progress session. It is apparent that the smaller the C value, the more accurate the account balance reported by the OCS.

In this paper, the above output measures are subscripted with "B" and "T" (i.e., M_B/C_B, and M_T/C_T) to represent the basic scheme and the threshold-based scheme, respectively.

III. Analytic Modeling

This section describes an analytic model for the basic scheme and the threshold-based scheme. Assume that there are N QoS classes. We make the following assumptions:

- A GPRS session starts with QoS class i ($1 \le i \le N$) with probability $1/N$.
- When a QoS change occurs, an in-progress session either terminates (with probability p_0) or switches to another QoS class with probability $\frac{1-p_0}{N-1}$.
- For $1 \le i \le N$, let α_i be the amount of credit charged for each time unit in a QoS class i session. For each credit reservation (through the ABMF message exchange), the SBCF reserves $\alpha_i \tau_g$ credit units in the ABMF, and then grants these credit units (which lasts for τ_g time units) to the GGSN, where τ_g is an exponential random variable with mean $1/\mu$. Fixed τ_g will be considered in the simulation experiments later.
- Suppose that the QoS changes partition the streaming session into several subsessions, where x_n is the holding time of the n-th subsession. We assume that x_n is independent and identically distributed (i.i.d.) exponential random variable with mean $E[x_n] = 1/\lambda$.

In our previous study [9], the output measures for the basic scheme were derived as

$$M_B = \frac{\mu+\lambda}{p_0\lambda} \quad \text{and} \quad C_B = \left[\frac{1}{N(\mu+\lambda)}\right]\left(\sum_{i=1}^{N}\alpha_i\right) \tag{1}$$

To analytically model the threshold-based scheme, we assume that $\delta = 0$ and $p_0 \to 0$ such that the stationary behavior for the ABMF message exchanges during a subsession can be observed. For $p_0 \to 0$, it is clear that the output measure $M_T \to \infty$. Therefore, we shall derive the expected number m_T of the ABMF message exchanges for a subsession. This new measure will be used to partially validate our simulation model. Then we will use the simulation experiments to investigate M_T in Section IV. To simplify our discussion, the number of QoS classes is restricted to $N = 2$. These two QoS classes represent the low and the high bandwidth classes, respectively. The analytic model can be directly extended for $N > 2$.

A. Derivation for m_T

The number m_T of the ABMF message exchanges in a subsession is determined by two factors: (i) the QoS class of the subsession and (ii) the remaining credits left at the end of the previous subsession. Consider the timing diagram in Fig. 2, where the n-th subsession starts at t_n. Let J_n be the QoS class of the n-th subsession. Since $N = 2$, a GPRS bearer session consists of subsessions with QoS class 1 and class 2 alternatively. Suppose that during subsession n, the granted credit units are depleted, and an ABMF message exchange occurs at $t_{n+1} - \tau_r$ (note that if $t_n < t_{n+1} - \tau_r$ as illustrated in Fig. 2, the QoS class is not changed). At t_{n+1}, the $n+1$-st subsession starts and the QoS class is changed from J_n to J_{n+1}. At t_{n+1}, the SBCF retrieves the new tariff (i.e., α_2) from the Rating Function. Following the threshold-based scheme for $\delta = 0$, the SBCF directly assigns the remaining credit units to the new subsession without interacting with the ABMF. Suppose that the remaining credit units are not depleted in subsession $n+1$. Therefore, no ABMF message exchange occurs during $[t_{n+1}, t_{n+2}]$. At t_{n+2}, the $n+2$-nd subsession starts and the QoS is changed to $J_{n+2} = 1$.

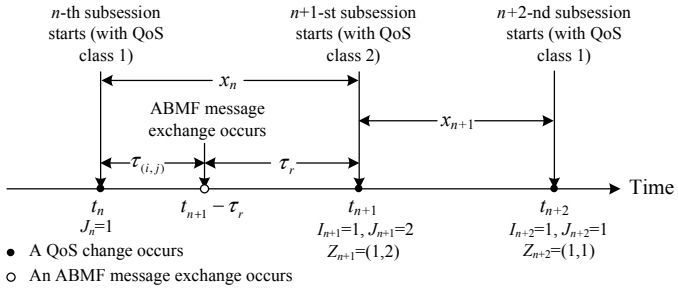


Fig. 2. Timing diagram for the threshold-based scheme.

For every J_n, we define a corresponding random variable I_n that represents the QoS class of the subsession immediately before the last ABMF message exchange occurs. When subsession n starts, the amount of credit units left is affected by I_n. Define $Z_n = (I_n, J_n)$ as a two dimensional random variable. It is clear that the process $\{Z_n, n \ge 1\}$ is a Markov chain.

Let $\pi_{(i,j)} = \lim_{n\to\infty} \Pr[Z_n = (i,j)]$ be the stationary distribution of the Markov chain, where $i, j \in \{1, 2\}$. Since the subsession holding time x_n is exponentially distributed, the QoS change occurrences form a Poisson stream. From Poisson Arrival See Time Average (PASTA) [5], $\pi_{(i,j)}$ also represents the proportion of time that the subsession resides in state (i, j). Let $m_{(i,j)}$ be the expected number of ABMF message exchanges for a subsession in state (i, j). Then m_T can be computed as

$$m_T = \textstyle\sum_{i,j\in\{1,2\}} m_{(i,j)}\pi_{(i,j)} \tag{2}$$

Assume that the Markov chain moves from state $(1, 2)$ to state $(1, 1)$ with transition probability p_a, and moves from state $(2, 1)$ to state $(2, 2)$ with transition probability p_b. Based on Fig. 3, the limiting probabilities are derived as

$$\left.\begin{aligned}
\pi_{(1,1)} &= p_a/\left[1 + p_a + (1-p_a)(1+p_b)/(1-p_b)\right] \\
\pi_{(1,2)} &= 1/\left[1 + p_a + (1-p_a)(1+p_b)/(1-p_b)\right] \\
\pi_{(2,1)} &= 1/\left[1 + p_b + (1-p_b)(1+p_a)/(1-p_a)\right] \\
\pi_{(2,2)} &= p_b/\left[1 + p_b + (1-p_b)(1+p_a)/(1-p_a)\right]
\end{aligned}\right\} \tag{3}$$

p_a and p_b in (3) are derived as follows: Let $\alpha_i\tau_u$ be the credit

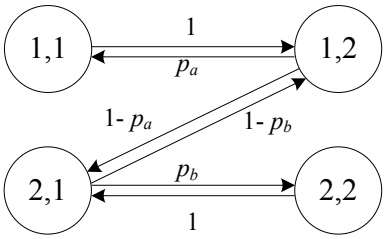


Fig. 3. Probability transition diagram ($N = 2$).

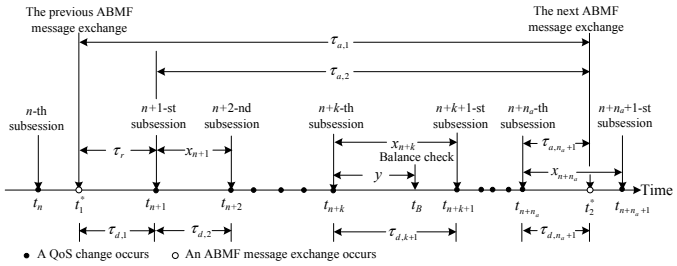


Fig. 4. Timing diagram for deriving C_T.

units left at the end of the previous subsession. As illustrated in Fig. 2, suppose that the current subsession is in state (i, j) and are granted $\alpha_j \tau_{(i,j)}$ credit units, where

$$\tau_{(i,j)} = \alpha_i \tau_u / \alpha_j \tag{4}$$

The subsession holding time $x_n = t_{n+1} - t_n = \tau_{(i,j)} + \tau_r$ has the density function

$$f_x(x_n) = \lambda e^{-\lambda x_n} \tag{5}$$

From (4), p_a and p_b can be expressed as

$$\left.\begin{aligned} p_a &= \Pr[\tau_{(1,2)} > x_n] = \Pr[\alpha_1 \tau_u / \alpha_2 > x_n] \\ p_b &= \Pr[\tau_{(2,1)} > x_n] = \Pr[\alpha_2 \tau_u / \alpha_1 > x_n] \end{aligned}\right\} \tag{6}$$

Since τ_g is exponentially distributed with rate μ, and from the memoryless property [8], τ_u is also exponential distributed with the density function

$$f_u(\tau_u) = \mu e^{-\mu \tau_u} \tag{7}$$

Let $\hat{\alpha} = \alpha_2 / \alpha_1$, then $\tau_{(1,2)} = \tau_u / \hat{\alpha}$ and $\tau_{(2,1)} = \hat{\alpha} \tau_u$. From (5)-(7), the transition probability p_a is derived as

$$p_a = \Pr[\tau_u / \hat{\alpha} > x_n] = \lambda / (\lambda + \hat{\alpha} \mu) \tag{8}$$

Similarly, the transition probability p_b is derived as

$$p_b = \Pr[\hat{\alpha} \tau_u > x_n] = \hat{\alpha} \lambda / (\hat{\alpha} \lambda + \mu) \tag{9}$$

Substituting (8) and (9) into (3), $\pi_{(i,j)}$ is re-written as

$$\left.\begin{aligned} \pi_{(1,1)} &= \tfrac{\lambda}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2 \lambda)} \quad , \quad \pi_{(2,1)} = \tfrac{\hat{\alpha}(\hat{\alpha}\lambda + \mu)}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2 \lambda)} \\ \pi_{(1,2)} &= \tfrac{\lambda + \hat{\alpha}\mu}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2 \lambda)} \quad , \quad \pi_{(2,2)} = \tfrac{\hat{\alpha}^2 \lambda}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2 \lambda)} \end{aligned}\right\} \tag{10}$$

$m_{(i,j)}$ in (2) is derived as follows: For $i = j$, where $\tau_{(i,i)} = \tau_u$ is exponentially distributed with rate μ. When $\alpha_i \tau_{(i,i)}$ credit units are depleted, an ABMF message exchange occurs and the SBCF grants another $\alpha_i \tau_g$ credit units to the subsession. Therefore, the ABMF message arrivals during a subsession period x_n are a Poisson stream with rate μ. By Little's formula [5], $m_{(i,i)}$ is expressed as

$$m_{(i,i)} = \mu E[x_n] = \mu / \lambda \tag{11}$$

$m_{(2,1)}$ is derived as follows: As shown in Fig. 2, assume that the credit units $\alpha_1 \tau_{(2,1)}$ are depleted at time $t_{n+1} - \tau_r$. From the memoryless property [8], τ_r is also exponential distributed with mean $E[\tau_r] = 1/\lambda$. In period τ_r, the ABMF message arrivals are a Poisson stream with rate μ. From the Little's formula, the expected number of ABMF message exchanges is $1 + \mu E[\tau_r]$. Since the probabilities that no ABMF message exchange in subsession n (i.e., $Z_{n+1} = (2, 2)$) or at least one ABMF message exchange occurs during subsession n (i.e., $Z_{n+1} = (1, 2)$) are denoted as p_b and $1 - p_b$, respectively. Therefore, $m_{(2,1)}$ is expressed as

$$m_{(2,1)} = \tfrac{\mu(\mu + \lambda)}{\lambda(\hat{\alpha}\lambda + \mu)} \tag{12}$$

Similarly,

$$m_{(1,2)} = \tfrac{\hat{\alpha}\mu(\mu + \lambda)}{\lambda(\lambda + \hat{\alpha}\mu)} \tag{13}$$

Substituting (10)-(13) into (2) to yield

$$m_T = \tfrac{2\hat{\alpha}\lambda\mu + \mu(\lambda + 2\hat{\alpha}\mu + \hat{\alpha}^2\lambda)}{2\lambda(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} \tag{14}$$

B. Derivation for C_T

This subsection derives the inaccuracy of credit information C_T in the threshold-based scheme. In Fig. 4, an ABMF message exchange occurs at t_1^* in subsession n. Suppose that the granted credit units are consumed at t_2^* in subsession $n + n_a$. Note that if $n_a = 0$, two consecutive ABMF message exchanges occur in subsession n. In $[t_1^*, t_2^*]$, a balance check occurs at t_B in subsession $n + k$, where $0 \le k \le n_a$. Let K be the QoS class of the session at t_1^* and $C_{T,i}$ be the C_T value under the condition that $K = i$. By considering whether $K = 1$ or $K = 2$, we can express C_T consumed in $[t_1^*, t_B]$ as

$$C_T = \textstyle\sum_{i=1}^{2} C_{T,i} \Pr[K = i] \tag{15}$$

$\Pr[K = i]$ can be derived in the following four situations.

Situation I: $Z_{n+k} = (1, 1)$**.** In this case, whether an ABMF message exchange occurs in subsession $n + k$ or not, we have $K = 1$.

Situation II: $Z_{n+k} = (2, 1)$**.** Since the balance check is a random observer of subsession $n+k$, and from the reverse residual life theorem [8], $t_B - t_{n+k}$ is also exponential distributed with rate λ. Therefore, the probability that no ABMF message occurring in period $[t_{n+k}, t_B]$ (i.e., $K = 2$) is p_b. On the other hand, if any ABMF message occurs in period $[t_{n+k}, t_B]$, we have $K = 1$ (with probability $1 - p_b$).

Situation III: $Z_{n+k} = (1, 2)$**.** Similar to Situation II, we have $K = 1$ and $K = 2$ with probability p_a and $1 - p_a$, respectively.

Situation IV: $Z_{n+k} = (2, 2)$**.** Similar to Situation I, we have $K = 2$ with probability 1.

Let $\pi_{(i,j)}$ be the probability that when a random observer arrives (i.e., the balance check occurs) at t_B, and the $n + k$-th subsession is in state (i, j). Based on the above discussion, and from (8)-(10), we have

$$\left.\begin{aligned} \Pr[K = 1] &= \tfrac{2\lambda + \hat{\alpha}\mu}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} \\ \Pr[K = 2] &= \tfrac{2\hat{\alpha}^2\lambda + \hat{\alpha}\mu}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} \end{aligned}\right\} \tag{16}$$

In (15), $C_{T,i}$ is derived as follows: As illustrated in Fig. 4, $\tau_r = t_{n+1} - t_1^*$. From the memoryless property [8], τ_r has the density function

$$f_r(\tau_r) = \lambda e^{-\lambda \tau_r} \tag{17}$$

In Fig. 4, the remaining granted credit units at t_1^* lasts for $\tau_{a,1}$ time units. Similarly, the remaining granted credit units at t_{n+l-1} $(2 \le l \le n_a + 1)$ can last for $\tau_{a,l}$ time units, where

$$\tau_{a,l} = \begin{cases} t_2^* - t_1^* & , l = 1 \\ t_2^* - t_{n+l-1} & , 2 \le l \le n_a + 1 \end{cases}$$

$$\text{Let} \quad \tau_{d,l} = \begin{cases} \min(\tau_r, \tau_{a,1}) & , l = 1 \\ \min(x_{n+l-1}, \tau_{a,l}) & , 2 \le l \le n_a + 1 \end{cases}$$

Assume that a balance check occurs at t_B in subsession $n+k$. Let $y = t_B - \max(t_1^*, t_{n+k})$, then $t_B - t_1^* = \sum_{l=1}^{k} \tau_{d,l} + y$. Denote $P_{i,k}$ as the probability that exactly k QoS changes occur in $[t_1^*, t_B]$ under the condition that $K = i$. Let $\alpha_{i,l} = \alpha_{(i+l \bmod 2)+1}$ for $i \in \{1, 2\}$, $C_{T,i}$ can be expressed as

$$\begin{aligned} C_{T,i} &= \sum_{k=0}^{\infty} P_{i,k} \Big\{ \sum_{l=1}^{k} \alpha_{i,l} E[\tau_{d,l} | K = i, t_{n+l} < t_2^*] \\ &\quad + \alpha_{i,k+1} E[y | K = i, t_{n+k} < t_B < t_{n+k+1}] \Big\} \end{aligned} \tag{18}$$

In (18), $E[\tau_{d,l} | K = i, t_{n+l} < t_2^*]$ is derived as follows: For $K = i$, let $f_{i,l}(\tau_{a,l})$ be the density function of $\tau_{a,l}$. We have

$$\left.\begin{aligned} f_{1,l}(\tau_{a,l}) &= \begin{cases} \mu e^{-\mu \tau_{a,l}} & , l \text{ is odd} \\ \hat{\alpha}\mu e^{-\hat{\alpha}\mu \tau_{a,l}} & , l \text{ is even} \end{cases} \\ f_{2,l}(\tau_{a,l}) &= \begin{cases} \mu e^{-\mu \tau_{a,l}} & , l \text{ is odd} \\ \frac{\mu}{\hat{\alpha}} e^{-\frac{\mu}{\hat{\alpha}} \tau_{a,l}} & , l \text{ is even} \end{cases} \end{aligned}\right\} \tag{19}$$

Eq. (19) is explained as follows: For $K = i$ and when l is odd, subsession $n + l - 1$ is served with QoS class i. Therefore, $f_{i,l}(\tau_{a,l})$ is exponentially distributed with rate μ. When l is even, subsession $n + l - 1$ is served with QoS classes 2 and 1 for $K = 1$ and $K = 2$, respectively. Therefore, $f_{i,l}(\tau_{a,l})$ is exponentially distributed with rates $\hat{\alpha}\mu$ and $\mu/\hat{\alpha}$, respectively. From (17) and (19), $E[\tau_{d,1} | K = i, t_{n+1} < t_2^*]$ is derived as

$$\begin{aligned} & E[\tau_{d,1} | K = i, t_{n+1} < t_2^*] \\ = \; & E[\tau_r | \tau_r < \tau_{a,1}, K = i] \\ = \; & \frac{\int_{\tau_r=0}^{\infty} \tau_r f_r(\tau_r) \int_{\tau_{a,1}=\tau_r}^{\infty} f_{i,1}(\tau_{a,1}) d\tau_{a,1} d\tau_r}{\int_{\tau_r=0}^{\infty} f_r(\tau_r) \int_{\tau_{a,1}=\tau_r}^{\infty} f_{i,1}(\tau_{a,1}) d\tau_{a,1} d\tau_r} \end{aligned} \tag{20}$$

For $l \ge 2$, and from (5) and (19), $E[\tau_{d,l} | K = i, t_{n+l} < t_2^*]$ can be derived as (21). Combining (20) and (21), $E[\tau_{d,l} | K = i, t_{n+l} < t_2^*]$ can be expressed as

$$\begin{aligned} & E[\tau_{d,l} | K = i, t_{n+l} < t_2^*] \\ = \; & \begin{cases} 1/(\lambda + \mu) & , l \text{ is odd} \\ 1/(\lambda + \hat{\alpha}\mu) & , K = 1 \text{ and } l \text{ is even} \\ 1/(\lambda + \mu/\hat{\alpha}) & , K = 2 \text{ and } l \text{ is even} \end{cases} \end{aligned} \tag{22}$$

In (18), $E[y | K = i, t_1^* < t_B < t_{n+1}]$ and $E[y | K = i, t_{n+k} < t_B < t_{n+k+1}]$ are derived as follows: Since the balance check is a random observation point of $\tau_{d,k+1}$, and from reverse residual life theorem for the exponential random variables [8], $E[y | K = i, \max(t_1^*, t_{n+k}) < t_B < t_{n+k+1}] = E[\tau_{d,k+1}]$. If $t_1^* < t_B < t_{n+1}$, we have

$$\begin{aligned} & E[y | K = i, t_1^* < t_B < t_{n+1}] \\ = \; & \int_{\tau_{a,1}=0}^{\infty} \tau_{a,1} f_{i,1}(\tau_{a,1}) \int_{\tau_r=\tau_{a,1}}^{\infty} f_r(\tau_r) d\tau_r d\tau_{a,1} \\ & + \int_{\tau_r=0}^{\infty} \tau_r f_r(\tau_r) \int_{\tau_{a,1}=\tau_r}^{\infty} f_{i,1}(\tau_{a,1}) d\tau_{a,1} d\tau_r \end{aligned} \tag{23}$$

For $t_{n+k} < t_B < t_{n+k+1}$, we have (24). Combining (23) and (24), $E[y | K = i, \max(t_1^*, t_{n+k}) < t_B < t_{n+k+1}]$ can be derived as

$$\begin{aligned} & E[y | K = i, \max(t_1^*, t_{n+k}) < t_B < t_{n+k+1}] \\ = \; & E[\tau_{d,k+1} | K = i] \\ = \; & \begin{cases} 1/(\lambda + \mu) & , k \text{ is even} \\ 1/(\lambda + \hat{\alpha}\mu) & , K = 1 \text{ and } k \text{ is odd} \\ 1/(\lambda + \mu/\hat{\alpha}) & , K = 2 \text{ and } k \text{ is odd} \end{cases} \end{aligned} \tag{25}$$

To derive $P_{i,k}$ in (18), we consider $[t_1^*, t_2^*]$ as a renewal interval and define a continuous-time stochastic process $\Delta(t)$, where

$$\Delta(t) = \begin{cases} 1 & , t_1^* \le t < t_{n+1} \\ l & , t_{n+l-1} \le t < t_{n+l} \text{ and } 2 \le l \le n_a \\ n_a + 1 & , t_{n+n_a} \le t < t_2^* \end{cases}$$

Therefore, $P_{i,k} = \Pr[\Delta(t_B) = k + 1 | K = i]$. In Fig. 4, $t_2^* > t_{n+k}$ and the length of time such that $\Delta(t) = k + 1$ is $\tau_{d,k+1}$. Since t_B is uniformly distributed over $[t_1^*, t_2^*]$, and from the alternating renewal theory [8], $P_{i,k}$ is computed as

$$P_{i,k} = \frac{\Pr[t_2^* > t_{n+k} | K = i] E[\tau_{d,k+1} | K = i]}{E[t_2^* - t_1^* | K = i]} \tag{26}$$

From (5), (17) and (19), $\Pr[t_2^* > t_{n+k} | K = i]$ in (26) is derived as

$$\begin{aligned} & \Pr[t_2^* > t_{n+k} | K = i] \\ = \; & \Pr[\tau_{a,1} > \tau_r | K = i] \prod_{l=2}^{k} \Pr[\tau_{a,l} > x_{n+l-1} | K = i] \\ = \; & \begin{cases} \frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil} (\lambda+\hat{\alpha}\mu)^{\lfloor k/2 \rfloor}} & , K = 1 \\ \frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil} (\lambda+\mu/\hat{\alpha})^{\lfloor k/2 \rfloor}} & , K = 2 \end{cases} \end{aligned} \tag{27}$$

In Fig. 4, $t_2^* - t_1^* = \sum_{l=1}^{n_a+1} \tau_{d,l}$. Therefore, the expected length of renewal interval $[t_1^*, t_2^*]$ can be computed as

$$\begin{aligned} & E[t_2^* - t_1^* | K = i] \\ = \; & \sum_{l=1}^{\infty} \Pr[t_2^* > t_{n+l-1} | K = i] E[\tau_{d,l} | K = i] \end{aligned} \tag{28}$$

Following the derivation for (27), we have

$$\begin{aligned} & \Pr[t_2^* > t_{n+l-1} | K = i] \\ = \; & \begin{cases} \frac{\lambda^{l-1}}{(\lambda+\mu)^{\lceil (l-1)/2 \rceil} (\lambda+\hat{\alpha}\mu)^{\lfloor (l-1)/2 \rfloor}} & , K = 1 \\ \frac{\lambda^{l-1}}{(\lambda+\mu)^{\lceil (l-1)/2 \rceil} (\lambda+\mu/\hat{\alpha})^{\lfloor (l-1)/2 \rfloor}} & , K = 2 \end{cases} \end{aligned} \tag{29}$$

Following the derivation for (23)-(25), we have

$$E[\tau_{d,l} | K = i] = \begin{cases} 1/(\lambda + \mu) & , l \text{ is odd} \\ 1/(\lambda + \hat{\alpha}\mu) & , K = 1;\ l \text{ is even} \\ 1/(\lambda + \mu/\hat{\alpha}) & , K = 2;\ l \text{ is even} \end{cases} \tag{30}$$

Then, $E[t_2^* - t_1^* | K = i]$ can be obtained by substituting (29) and (30) into (28).

From (25) and (27), (26) is derived as (31). Finally, by substituting (22), (25) and (31) in (18), $C_{T,i}$ is obtained and C_T can be computed from Eqs. (15), (16) and (18).

The analytic model developed in this section is used to validate against the simulation experiments. The input parameter τ_g and the output measures C_B and C_T are normalized by the mean $1/\lambda$ of the subsession holding time. The discrepancies between analytic analysis (specifically, Eqs. (1), (14) and (15))

$$\begin{aligned} E[\tau_{d,l}|K=i, t_{n+l} < t_2^*] &= E[x_{n+l-1}|x_{n+l-1} < \tau_{a,l}, K=i] \\ &= \frac{\int_{x_{n+l-1}=0}^{\infty} x_{n+l-1} f_x(x_{n+l-1}) \int_{\tau_{a,l}=x_{n+l-1}}^{\infty} f_{i,l}(\tau_{a,l}) d\tau_{a,l} dx_{n+l-1}}{\int_{x_{n+l-1}=0}^{\infty} f_x(x_{n+l-1}) \int_{\tau_{a,l}=x_{n+l-1}}^{\infty} f_{i,l}(\tau_{a,l}) d\tau_{a,l} dx_{n+l-1}} \end{aligned} \quad (21)$$

$$\begin{aligned} E[y|K=i, t_{n+k} < t_B < t_{n+k+1}] &= \int_{x_{n+k}=0}^{\infty} x_{n+k} f_x(x_{n+k}) \int_{\tau_{a,k+1}=x_{n+k}}^{\infty} f_{i,k+1}(\tau_{a,k+1}) d\tau_{a,k+1} dx_{n+k} \\ &\quad + \int_{\tau_{a,k+1}=0}^{\infty} \tau_{a,k+1} f_{i,k+1}(\tau_{a,k+1}) \int_{x_{n+k}=\tau_{a,k+1}}^{\infty} f_x(x_{n+k}) dx_{n+k} d\tau_{a,k+1} \end{aligned} \quad (24)$$

$$\left.\begin{aligned} P_{1,k} &= \begin{cases} \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)}\right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)}\right] \left[\frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil} (\lambda+\hat{\alpha}\mu)^{\lfloor k/2 \rfloor + 1}}\right] & , k \text{ is odd} \\ \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)}\right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)}\right] \left[\frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil + 1} (\lambda+\hat{\alpha}\mu)^{\lfloor k/2 \rfloor}}\right] & , k \text{ is even} \end{cases} \\ P_{2,k} &= \begin{cases} \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})}\right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})}\right] \left[\frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil} (\lambda+\mu/\hat{\alpha})^{\lfloor k/2 \rfloor + 1}}\right] & , k \text{ is odd} \\ \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})}\right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})}\right] \left[\frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil + 1} (\lambda+\mu/\hat{\alpha})^{\lfloor k/2 \rfloor}}\right] & , k \text{ is even} \end{cases} \end{aligned}\right\} \quad (31)$$

and simulation are within 1%. The simulation model follows the discrete event approach described in [6], and the details are omitted.

IV. Numerical Examples

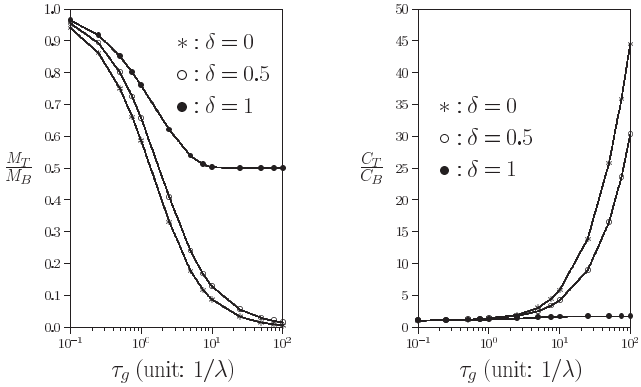
This section uses simulation experiments to investigate the performance of the credit re-authorization procedure. In the simulation, the granted time units τ_g is a fixed value and the performance is investigated by the following output measures:

- M_B and M_T represent the numbers of ABMF message exchanges performed by the basic scheme and the threshold-based scheme, respectively. The smaller the M_T/M_B value, the better the threshold-based scheme.
- C_B and C_T represent the inaccuracies of the credit information reported by the basic scheme and the threshold-based scheme, respectively. The larger the C_T/C_B value, the larger the inaccuracy of the account balance reported by the threshold-based scheme.

Effects of the granted time units τ_g. Fig. 5 plots M_T/M_B and C_T/C_B against the threshold parameter δ and the granted time units τ_g, where $N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$. Fig. 5 shows the trivial result that M_T/M_B is a decreasing function of τ_g, and C_T/C_B is an increasing function of τ_g. The non-trivial result is that when $\tau_g \leq 5/\lambda$, the threshold-based scheme significantly reduces the ABMF signaling overhead while the inaccuracy of the credit information insignificantly increases.

Effects of the threshold parameter δ. Fig. 6 plots M and C against the threshold parameter δ, where $N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$. The output measures for the basic scheme are not affected by δ. For the threshold-based scheme, M_T increases and C_T decreases as δ increases. When δ is large, the basic scheme and the threshold-based scheme have the same performance. In Fig. 6, the performance of the threshold-based scheme is similar to that of the basic scheme when $\delta \geq 2.5$.

Fig. 6 (b) quantitatively indicates how δ and τ_g affect C. In the OCS, the guideline for selecting the granted time units



(a) Effect on M_T/M_B (b) Effect on C_T/C_B

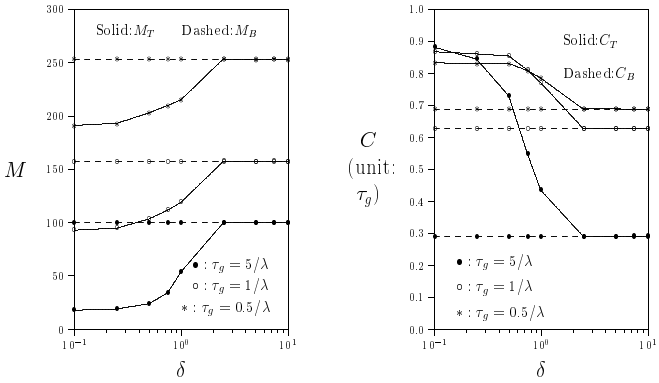
Fig. 5. Effects of τ_g ($N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$)

τ_g can be found in [10]. When a specific τ_g is selected, for example, $\tau_g = 5/\lambda$, we observe that $C_B = 0.29\tau_g$ in the basic scheme. If the mobile operator can tolerate a larger inaccuracy of the credit information, e.g., $C \leq 0.5\tau_g$, then we can choose $\delta = 1$ in the threshold-based scheme. In this case, $C_T = 0.43\tau_g$ and M is decreased by 46.45% as compared with the basic scheme.

Also, as described in [9], p_0 and N only have insignificant effects on M_T/M_B and C_T/C_B. The details are omitted.

V. Conclusions

This paper studied *Online Charging System* (OCS) in UMTS. We proposed a threshold-based scheme with parameter δ to reduce the traffic signaling for the OCS credit re-authorization procedure. An analytic model is developed to investigate the performance on the basic scheme [9] and our proposed threshold-based scheme. Basically, the threshold-based scheme reduces the number M of ABMF message exchanges during a session at the cost of increasing the inaccuracy of credit information C when a balance check occurs. These two conflicting output measures are affected



(a) Effect on M (b) Effect on C

Fig. 6. Effects of δ ($N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$)

by the threshold parameter δ and the granted time units τ_g. We make the following observations, where the subscripts "B" and "T" in the output measures M and C represent the basic scheme and the threshold-based scheme, respectively.

- The ratio M_T/M_B is a decreasing function of τ_g, and the ratio C_T/C_B is an increasing function of τ_g. When τ_g is small, the threshold-based scheme significantly reduces the ABMF signaling overhead while the inaccuracy of the credit information insignificantly increases.
- As the threshold parameter δ increases, M_T increases and C_T decreases. When δ is large, the basic scheme and the threshold-based scheme have the same performance.

Based on the above discussion, the mobile operator can select the appropriate δ and τ_g values for various traffic conditions based on our model.

REFERENCES

[1] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Service and System Aspects; Telecommunication management; Charging management; Online Charging System (OCS): Applications and interfaces. Technical Specification 3G TS 32.296 Version 6.3.0 (2006-09), 2006.

[2] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Service and System Aspects; Telecommunication management; Charging management; Diameter charging applications. Technical Specification 3G TS 32.299 Version 6.10.0 (2007-03), 2007.

[3] B. Bhushan *et al*, "OSS functions for flexible charging and billing of mobile services in a federated environment," in *Proc. 9th IFIP/IEEE International Symposium on Integrated Network Management*, May 2005.

[4] F. Ghys and A. Vaaraniemi, "Component-based charging in a next-generation multimedia network," *IEEE Commun. Mag.*, vol. 41, no. 1, pp. 99–102, 2003.

[5] L. Kleinrock, *Queueing Systems: Volume I–Theory*. John Wiley & Sons, 1976.

[6] Y.-B. Lin and Y.-K. Chen, "Reducing authentication signaling traffic in third generation mobile network," *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, pp. 493–504, 2003.

[7] Y.-B. Lin and A.-C. Pang, *Wireless and Mobile All-IP Networks*. Wiley, 2005.

[8] S. M. Ross, *Stochastic Processes*. John Wiley & Sons, 1996.

[9] S.-I. Sou, "Performance analysis of credit re-authorization schemes in UMTS online charging system," accepted for publication in *International Wireless Communications and Mobile Computing Conference*, 2007.

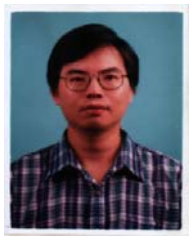
[10] S.-I. Sou, H.-N. Hung, Y.-B. Lin, N.-F. Peng, and J.-Y. Jeng, "Modeling credit reservation procedure for UMTS online charging system," accepted for publication in *IEEE Trans. Wireless Commun.*

[11] W.-Z. Yang, F.-S. Lu, and M.-F. Chang, "Performance modeling of integrated mobile prepaid services," *IEEE Trans. Veh. Technol,* vol. 56, no. 2, pp. 899–906, 2007.

Sok-Ian Sou received B.S., M.S., and Ph.D. degrees in computer science and information engineering from National Chiao Tung University (NCTU) in 1997, 2004 and 2008, respectively. She joined the Department of Electrical Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, as an assistant professor in 2008. Her current research interests include design and analysis of personal communications services networks, mobile computing and performance modeling.

Yi-Bing Lin (M'96-SM'96-F'03) is Chair Professor and Dean of College of Computer Science, National Chiao Tung University.

His current research interests include mobile computing and cellular telecommunications services. Dr. Lin has published over 200 journal articles and more than 200 conference papers. He is the co-author of the books *Wireless and Mobile Network Architecture* (with Imrich Chlamtac, published by Wiley, 2001) and *Wireless and Mobile All-IP Networks* (with Ai-Chun Pang, published by Wiley, 2005). Dr. Lin is an IEEE Fellow, ACM Fellow, AAAS Fellow, and IEE Fellow.



Jeu-Yih Jeng received the B.S. degree in mathematics from Fu-Jen University in 1983, the M.S. degree in applied mathematics from National Chiao Tung University in 1985, and the Ph.D. degree in computer science and information engineering from National Chiao-Tung University in 1998. Since 1985, he has been with the Information Technology Laboratory of Telecommunication Laboratories, Chunghwa Telcom Co., Ltd, where he is currently a Distinguished Researcher and a project manager. His research interests include design and analysis of personal communications services network, development of telecommunication operation support systems, and performance modeling.