Modeling Credit Reservation Procedure for UMTS Online Charging System

Sok-Ian Sou, *Student Member, IEEE,* Hui-Nien Hung, Yi-Bing Lin, *Fellow, IEEE,* Nan-Fu Peng, and Jeu-Yih Jeng

***Abstract*— The *IP Multimedia Core Network Subsystem* (IMS) provides real-time multimedia services for *Universal Mobile Telecommunications System* (UMTS). Through Recharge Threshold-based Credit Reservation (RTCR) mechanism, prepaid IMS services can be supported by the *Online Charging System* (OCS) in UMTS. In RTCR, when the remaining amount of prepaid credit is below a threshold, the OCS reminds the user to recharge the prepaid account. It is essential to choose an appropriate recharge threshold to reduce the probability that the in-progress service sessions are forced-terminated. An analytic model is developed to investigate the performance of RTCR for the OCS. Based on our study, the network operator can select the appropriate parameter values for various traffic conditions.**

***Index Terms*— Charging, IP multimedia core network subsystem (IMS), prepaid services, universal mobile telecommunications system (UMTS).**

I. Introduction

PREPAID telecommunications service requires a user to make an advanced payment before enjoying the service. Usage of prepaid service does not require deposit and monthly bill. Instead the usage fee is directly deducted from the user's prepaid account. In the *Global System for Mobile Communications* (GSM), prepaid voice service is implemented as a circuit-switched domain service. In the *General Packet Radio Service* (GPRS), prepaid data service is also offered in the packet-switched domain. Four billing technologies have been used in mobile prepaid service: hot billing approach [4], service node approach, intelligent network (IN) approach [8] and handset-based approach. Details of these approaches can be found in Chapter 17 in [10].

Manuscript received May 12, 2006; revised July 17, 2006; accepted August 18, 2006. The associate editor coordinating the review of this paper and approving it for publication was D. Wu. This work was sponsored in part by the NSC Excellence Project NSC 94-2752-E-009-005-PAE, NSC 94-2219-E-009-001, NSC 94-2213-E-009-104, the NTP VoIP Project under grant number NSC 94-2219-E-009-002, the NTP Service IOT Project under grant number NSC 94-2219-E-009-024, Intel, Chung Hwa Telecom, IIS/Academia Sinica, the ITRI/NCTU Joint Research Center, and MoE ATU.

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

H.-N. Hung and N.-F. Peng are with the Institute of Statistics, National Chiao Tung University, Hsinchu 30010, Taiwan, R.O.C. (e-mail: {hhung, nanfu}@stat.nctu.edu.tw). The work of H.-N. Hung was supported in part by the National Science Council of Taiwan under Grant NSC 94-2118-M-009-003.

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.2007.060250.

In *Universal Mobile Telecommunications System* (UMTS), real-time multimedia services are supported by the *IP Multimedia Core Network Subsystem* (IMS). In IMS, the Diameter Protocol [6] is utilized for *Authentication, Authorization, and Accounting* (AAA) functions such as authentication and online charging [1]. Based on Diameter, the Diameter Credit Control protocol [6] is adopted in the *Online Charging System* (OCS) to provide IMS prepaid services [2]. The Diameter Credit Control protocol supports functionality for service charging with direct debiting and credit reservation. To support prepaid service, the OCS follows the IN approach. In this approach, when a mobile user subscribes to the prepaid service, an amount of prepaid credit is purchased and is maintained in the prepaid account. The prepaid credit units are deducted at the OCS in real time when the prepaid service is delivered. When the amount of the remaining prepaid credit is below a threshold, the OCS reminds the user (through short message or interactive voice response) to recharge the prepaid account. It is essential to choose an appropriate threshold to determine when to send the recharge messages. If the recharge threshold is set too small, the prepaid credit units may be depleted before the prepaid account is actually recharged, and it is likely that the in-progress IMS service sessions are forced to terminate before the recharge operation is complete. If the recharge threshold is set too large, the user will receive the recharge message too frequently and the network will experience heavy traffic. This paper proposes an analytic model to investigate the effects of the recharge threshold on the performance of the OCS.

II. Online Charging System for IMS Services

Fig. 1 shows the OCS architecture for IMS services [2]. In this architecture, online charging for the IMS services is performed by using the Diameter Credit Control (DCC) protocol (see Fig. 1 (a)) [1]. The OCS provides the *Session Based Charging Function* (SBCF; Fig. 1 (b)) responsible for online charging of network bearer and user sessions.

In the OCS, the *Account Balance Management Function* (ABMF; Fig. 1 (c)) keeps a user's balance and other account data. When the prepaid user's credit depletes, the ABMF connects the Recharge Server (Fig. 1 (f)) to trigger the recharge account function. The SBCF interacts with the Rating Function (Fig. 1 (e)) to determine the price of the requested service. The rating function handles a wide variety of rateable instances, such as data volume, session connection time and event service (e.g. for web content charging). The SBCF

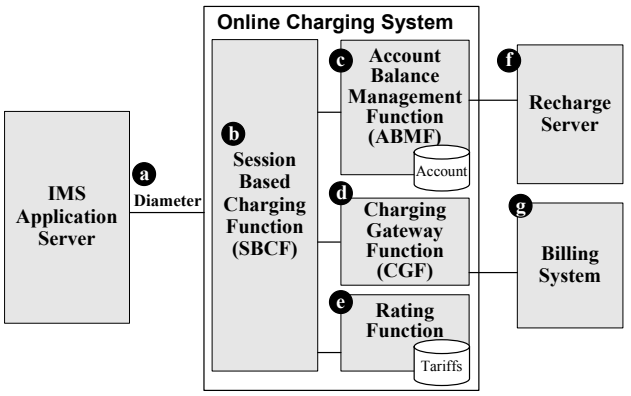


Fig. 1. Online charging system architecture for IMS services.

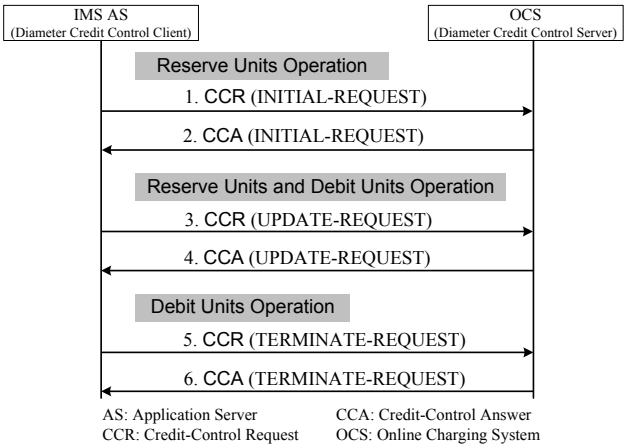


Fig. 2. Diameter credit control message flow.

interacts through the ABMF to query and update the user's account. The *Charging Data Records* (CDRs) generated by the charging functions are transferred to the Charging Gateway Function (CGF; Fig. 1 (d)) immediately. The CGF acts as a gateway between the 3GPP network and the Billing System (Fig. 1 (g)) [7].

A. Diameter Credit Reservation Procedure

In online charging services, the Diameter Credit Control (DCC) protocol is used for communications between an IMS network element and the OCS. The IMS network element acts as a DCC client and the OCS acts as a DCC server. The OCS credit control is achieved by exchanging the `Credit Control Request (CCR)` and the `Credit Control Answer (CCA)` messages. A credit control message can be one of the following types:

- INITIAL-REQUEST initiates a credit control session.
- UPDATE-REQUEST contains update credit control information for an in-progress session. This request is sent when the credit units currently allocated for the session are completely consumed.
- TERMINATION-REQUEST terminates an in-progress credit control session.
- EVENT-REQUEST is used for one-time credit control, which can be DIRECT_DEBITING, CHECK_BALANCE or PRICE_ENQUIRY.

The credit reservation procedure for session-based online charging includes three types of credit control operations: Reserve Units operation (Steps 1 and 2 in Fig. 2), Reserve Units and Debit Units operation (Steps 3 and 4 in Fig. 2) and Debit Units operation (Steps 5 and 6 in Fig. 2). Consider the scenario where a prepaid user requests an IMS session-based service from the *Application Server* (AS). The following operations are executed.

Step 1. [Reserve Units (request)] To start the service delivery with credit reservation, the IMS AS sends the INITIAL-REQUEST `CCR` message to the OCS. This message indicates the amount of requested credit.

Step 2. [Reserve Units (response)] Upon receipt of the `CCR` message, the OCS determines the price of the requested service and then reserves an amount of credit. After the reservation is performed, the OCS acknowledges the IMS AS with the `CCA` message including credit reserving information.

Step 3. [Reserve Units and Debit Units (request)] During the service session, the granted credit units may be depleted. If so, the IMS sends a UPDATE-REQUEST `CCR` message to the OCS. The IMS AS reports the amount of used credit and requests for additional credit units.

Step 4. [Reserve Units and Debit Units (response)] When the OCS receives the `CCR` message, it deducts the used credit units and reserves extra credit units for the IMS AS. The OCS acknowledges the IMS AS with the `CCA` message with the amount of the reserved credit. Note that the Reserve Units and Debit Units operation (i.e., Steps 3 and 4) may repeat many times before a service session is complete.

Step 5. [Debit Units (request)] When the service session is complete, the IMS AS sends the TERMINATE-REQUEST `CCR` message to the OCS. This action terminates the session and reports the amount of the consumed credit.

Step 6. [Debit Units (response)] The OCS releases the unused reserved credit. The OCS acknowledges the IMS AS with the `CCA` message. This message may contain the total cost of the service.

For the discussion purpose, we refer an *RU* operation as a Reserve Units operation (Steps 1 and 2) or a Reserve Units and Debit operation (Steps 3 and 4).

B. Recharge Threshold-Based Credit Reservation (RTCR) Mechanism

Upon receipt of an RU operation request (Steps 1 or 3 in Fig. 2), the OCS needs to determine when to send the recharge message, and how to allocate the credit units when the remaining credit left in the prepaid account is too small. These issues can be addressed by a simple mechanism called Recharge Threshold-based Credit Reservation (RTCR). Let C_r be the amount of the remaining prepaid credit in the OCS. Define C_{min} as the recharge threshold. When $C_r < C_{min}$, the OCS reminds the user to recharge the prepaid account by sending a recharge message. If C_{min} is too small, then the amount of the remaining credit C_r may not be

large enough to support all in-progress service sessions, and some of them will be forced to terminate. Note that force-termination will significantly degrade user satisfaction. For example, when force-termination occurs during a multimedia file downloading, the file transmission is not complete and the credit units that have already been consumed for downloading may be wasted. In order to avoid force-termination in RTCR, C_{min} should be appropriately selected. Also, after the recharge message has been sent, the OCS will not accommodate any new session, and all remaining credit units are reserved for the in-progress service sessions. In the next section, an analytic model is proposed to study the impact of parameter C_{min} on the RTCR mechanism.

III. Analytic Model for RTCR Mechanism

This section proposes an analytic model to investigate the RTCR mechanism. Assume that there are n types of session-based IMS services. Each type has its own traffic characteristics and communications parameters. For example, the average call holding time for a Voice over IP (VoIP) call session is 3 minutes, the average session holding time for the interactive mobile gaming sessions may range from 10 to 30 minutes. Details of the UMTS/IMS services characteristics and communications parameters can be found in [3], [5]. Note that the service sessions can be charged according to time (duration) or volumes of packets transmitted. From the characteristics of the volume-based service session (e.g., the inter-packet arrival distribution), the distribution of the packet volumes that transmitted in a service session can be mapped to a specific service session holding time distribution [8]. For the discussion purpose, we only consider session/call holding time distribution in this paper.

For $1 \leq i \leq n$, the sessions for type-i service can be activated from time to time. In Fig. 3, the current type-i service session starts at t_0 and completes at t_4, and the next type-i service session starts at t_5. Let the service session holding time be $t_{h,i} = t_4 - t_0$ and the inter-arrival time be $t_{a,i} = t_5 - t_4$. For VoIP call session services, $t_{h,i}$ represents the call holding time. For mobile data downloading service, $t_{h,i}$ represents the file transmission time. Assume that $t_{h,i}$ and $t_{a,i}$ are exponentially distributed with rates μ_i and λ_i, respectively (the exponential assumptions will be relaxed in the simulation experiments). We assume that each time unit of the type-i service session is charged for α_i credit units. Without loss of generality, let $\alpha_i = 1$ (i.e., the time unit is equal to the credit unit). Let C denote the amount of the initial prepaid credit for a mobile user. Let θ_i be the amount of credit that the OCS grants in each RU operation for a type-i session. It is essential to select appropriate C_{min} and θ_i values to "optimize" the performance of the RTCR mechanism in terms of the following output measures:

- $E[N_{r,i}]$: the expected number of the RU operations executed during a type-i session. The larger the $E[N_{r,i}]$ value, the higher the DCC control message overhead.
- P_f: the probability that an in-progress session is forced to terminate (for all service type-i). The smaller the P_f value, the better the user satisfaction.
- $E[C_d]$: the expected amount of unused credit units in the user account at the end of RTCR execution (before recharging). Note that $C_d = 0$ if any in-progress session is forced to terminate at the end of RTCR execution. It is apparent that the smaller the $E[C_d]$ value, the better the credit utilization in the user account.

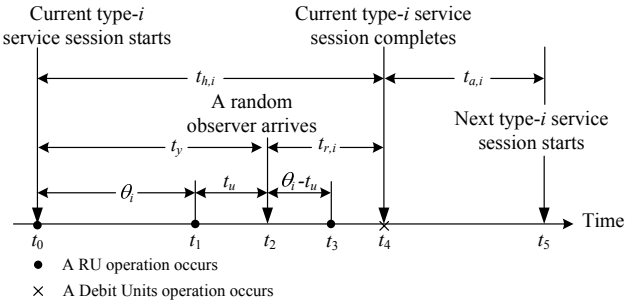


Fig. 3. Timing diagram for the RTCR mechanism.

A. Derivation for $E[N_{r,i}]$

This subsection derives the expected number $E[N_{r,i}]$ that the RU operations are executed in a type-i service session. Suppose that the OCS grants θ_i credit units to the AS each time. When the granted credit units θ_i are depleted, the AS requests the next θ_i credit units from the OCS. Then

$$E[N_{r,i}] = 1 + \sum_{j=1}^{\infty} \Pr[t_{h,i} > j\theta_i] \tag{1}$$

Eq. (1) says that the first RU operation (i.e., the Reserve Units operation) is always executed, and for $j > 0$, the j-th RU operation (i.e., the Reserve Units and Debits Units operation) will be executed with probability $\Pr[t_{h,i} > j\theta_i]$. Since $t_{h,i}$ is exponentially distributed with the density function

$$f_{h,i}(t_{h,i}) = \mu_i e^{-\mu_i t_{h,i}} \tag{2}$$

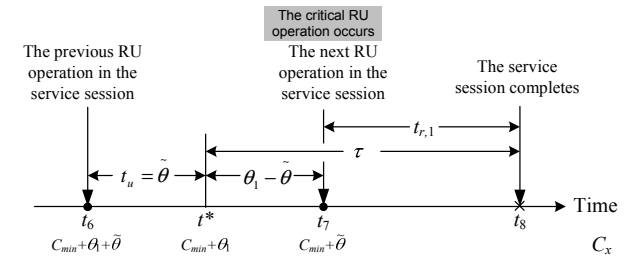
From (2), Eq. (1) is derived as

$$\begin{aligned} E[N_{r,i}] &= 1 + \sum_{j=1}^{\infty} \int_{t_{h,i}=j\theta_i}^{\infty} f_{h,i}(t_{h,i}) dt_{h,i} \\ &= 1 + \sum_{j=1}^{\infty} e^{-\mu_i j\theta_i} = \frac{1}{1 - e^{-\mu_i \theta_i}} \end{aligned} \tag{3}$$

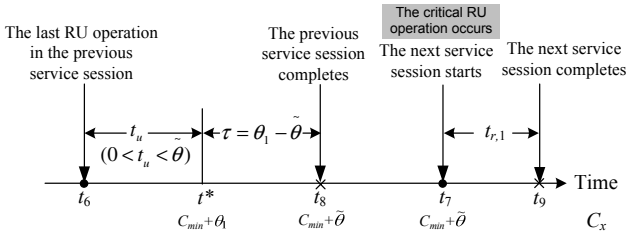
B. Exact Analytic Model for Single-type Service

This subsection derives the exact force-termination probability P_f and the expected credit $E[C_d]$ for single-type service (i.e., $n = 1$). Approximate P_f and $E[C_d]$ for multiple-type services (i.e., $n \geq 2$) will be derived in the next subsection.

1) Derivation for P_f ($n = 1$): Consider the timing diagram in Fig. 3 where the RU operations in the current service session occur at t_0, t_1 and t_3, respectively. Assume that the OCS grants θ_1 units to the AS in every RU operation and $\theta_1 \leq C_{min}$. Then $t_3 - t_1 = t_1 - t_0 = \theta_1$. Suppose that a random observer arrives at t_2. Let $t_y = t_2 - t_0$ and $t_{r,1} = t_4 - t_2$ be the elapsed holding time (i.e., the age) and the residual holding time of the service session, respectively. Let $t_u = t_2 - t_1$ denote the consumed credit (time) units in the AS with density function $f_u(t_u)$. Then $0 \leq t_u \leq \theta_1$ and the unused credit units left in the AS is $\theta_1 - t_u$.



(a) Case I: The critical RU operation is issued from an existing session.



(b) Case II: The critical RU operation is issued from a new session.

- A RU operation occurs
- × A Debit Units operation occurs

Fig. 4. Timing diagram for deriving $\tilde{\theta}$.

For an in-progress service session at a particular time point t, the exact unused credit units for the user is $C_x(t) = C_r(t) + (\theta_1 - t_u)$, which is the sum of the remaining credit units $C_r(t)$ in the OCS and the unused credit units $(\theta_1 - t_u)$ in the AS. Note that at times t_0, t_1 and t_3, no unused credit units are left in the AS and therefore $C_x(t_0) = C_r(t_0)$, $C_x(t_1) = C_r(t_1)$ and $C_x(t_3) = C_r(t_3)$. Define "critical" time t^* as the time when $C_x(t^*) = C_{min} + \theta_1$. The first RU operation occurs after t^* is referred to as the "critical" RU operation. Immediately after the critical RU operation is performed, the OCS will send the recharge message, and newly incoming session requests will be rejected. The critical RU operation may occur during a session execution (Case I; see Fig. 4 (a)) or when a new session arrives (Case II; see Fig. 4 (b)). In Fig. 4, let $\tilde{\theta} = \theta_1 - (\min(t_7, t_8) - t^*)$. Note that when the critical RU operation arrives at t_7, $C_x(t_7) = C_r(t_7) = C_{min} + \tilde{\theta}$. Furthermore, in Case II, no credit units are consumed during $[t_8, t_7]$, and $C_x(t_7) = C_x(t_8)$. The $\tilde{\theta}$ value will be derived later. After the OCS has granted θ_1 units to this request, the remaining credit left in the prepaid account becomes $C_r(t_7^+) = C_{min} - (\theta_1 - \tilde{\theta}) < C_{min}$. At this point, the OCS will send a recharge message to the prepaid user, and the maximum service time that the remaining credit can support is $C_{min} + \tilde{\theta}$. Therefore P_f is computed as

$$P_f = \Pr[t_{r,1} > C_{min} + \tilde{\theta}] \tag{4}$$

Since the session holding time $t_{h,1}$ is exponentially distributed with rate μ_1, and from the residual time theorem [11], $t_{r,1}$ has the same density function as $t_{h,1}$; i.e.,

$$f_{r,1}(t_{r,1}) = \mu_1 e^{-\mu_1 t_{r,1}} \tag{5}$$

In (4), $\tilde{\theta}$ is derived as follows.

Case I. In Fig. 4 (a), the critical RU operation is issued by an in-progress session at t_7. In this case, the previous RU operation is issued at $t_6 = t_7 - \theta_1$ where $C_r(t_6) = C_x(t_6) = C_{min} + \theta_1 + \tilde{\theta}$, and $t_6 < t^* < t_7$. At the critical time t^*, the consumed time units t_u must be $\tilde{\theta}$ (with density $f_u(\tilde{\theta})$) and the residual holding time $\tau = t_8 - t^*$ for the session must be longer than $\theta_1 - \tilde{\theta}$ (with probability $\int_{\tau=\theta_1-\tilde{\theta}}^{\infty} f_{r,1}(\tau)d\tau$). Therefore, the density of $\tilde{\theta}$ for Case I is expressed as

$$f_{\tilde{\theta},I}(\tilde{\theta}) = f_u(\tilde{\theta}) \int_{\tau=\theta_1-\tilde{\theta}}^{\infty} f_{r,1}(\tau)d\tau \tag{6}$$

Case II. In Fig. 4 (b), the critical RU operation is issued by a new session at t_7. The last RU operation in the previous service session occurs at t_6 and the previous service session completes at t_8, and $C_x(t_8) = C_{min} + \tilde{\theta}$. Note that $t^* > t_6 \geq t_8 - \theta_1$. Therefore, $C_{min} + \theta_1 + \tilde{\theta} \geq C_x(t_6) > C_{min} + \theta_1$. At t^*, the AS consumes t_u credit units, where $0 < t_u < \tilde{\theta}$ (with probability $\int_{t_u=0}^{\tilde{\theta}} f_u(t_u)dt_u$), and the residual holding time $\tau = t_8 - t^*$ for the session is $\theta_1 - \tilde{\theta}$ (with density $f_{r,1}(\theta_1 - \tilde{\theta})$). Therefore, the density of $\tilde{\theta}$ for Case II is expressed as

$$f_{\tilde{\theta},II}(\tilde{\theta}) = \int_{t_u=0}^{\tilde{\theta}} f_u(t_u)dt_u f_{r,1}(\theta_1 - \tilde{\theta}) \tag{7}$$

Combining (6) and (7), we have

$$\begin{aligned} f_{\tilde{\theta}}(\tilde{\theta}) &= f_u(\tilde{\theta}) \int_{\tau=\theta_1-\tilde{\theta}}^{\infty} f_{r,1}(\tau)d\tau \\ &\quad + \int_{t_u=0}^{\tilde{\theta}} f_u(t_u)dt_u f_{r,1}(\theta_1 - \tilde{\theta}) \end{aligned} \tag{8}$$

In (8), $f_u(t_u)$ is derived as follows. Let t_y be the elapsed holding time of the in-progress session with density function $f_y(t_y)$. According to the reverse residual time theorem [11], t_y has the same distribution as the residual of $t_{h,1}$. That is, $f_y(t_y) = f_{r,1}(t_y) = \mu_1 e^{-\mu_1 t_y}$. Since $t_u = t_y - \left\lfloor \frac{t_y}{\theta_1} \right\rfloor \theta_1$, the density function for t_u (for $0 \leq t_u \leq \theta_1$) is

$$f_u(t_u) = \sum_{j=0}^{\infty} f_y(t_u + j\theta_1) = \frac{\mu_1 e^{-\mu_1 t_u}}{1 - e^{-\mu_1 \theta_1}} \tag{9}$$

Substituting (5) and (9) into (8), we have

$$f_{\tilde{\theta}}(\tilde{\theta}) = \frac{\mu_1 e^{\mu_1 \tilde{\theta}}}{e^{\mu_1 \theta_1} - 1} \tag{10}$$

From (4), (5) and (10), P_f is derived as

$$\begin{aligned} P_f &= \int_{\tilde{\theta}=0}^{\theta_1} \int_{t_{r,1}=C_{min}+\tilde{\theta}}^{\infty} \mu_1 e^{-\mu_1 t_{r,1}} \left(\frac{\mu_1 e^{\mu_1 \tilde{\theta}}}{e^{\mu_1 \theta_1} - 1} \right) dt_{r,1} d\tilde{\theta} \\ &= \frac{\mu_1 \theta_1 e^{-\mu_1 C_{min}}}{e^{\mu_1 \theta_1} - 1} \end{aligned} \tag{11}$$

2) Derivation for $E[C_d]$ $(n = 1)$: It is clear that $E[C_d] = E[\lim_{t\to\infty} C_r(t)]$ assuming that the prepaid account is not recharged at the end of the RTCR execution. In Fig. 4, the critical RU operation occurs at t_7 when $C_x(t_7) = C_{min} + \tilde{\theta}$. If $C_{min} + \tilde{\theta} > t_{r,1}$, then $C_d = C_{min} + \tilde{\theta} - t_{r,1}$. Otherwise, $C_d = 0$. Therefore, $E[C_d]$ is expressed as

$$E[C_d] = E[\max\{C_{min} + \tilde{\theta} - t_{r,1}, 0\}] \tag{12}$$

From (5) and (10), Eq. (12) is derived as

$$\begin{aligned} E[C_d] &= \int_{\tilde{\theta}=0}^{\theta_1}\int_{t_{r,1}=0}^{C_{min}+\tilde{\theta}} (C_{min}+\tilde{\theta}-t_{r,1}) \\ &\quad \times f_{r,1}(t_{r,1}) f_{\tilde{\theta}}(\tilde{\theta}) dt_{r,1} d\tilde{\theta} \\ &= C_{min} + \frac{\theta_1(e^{\mu_1\theta_1}+e^{-\mu_1 C_{min}})}{e^{\mu_1\theta_1}-1} - \frac{2}{\mu_1} \quad (13) \end{aligned}$$

C. Approximate Analytic Model for Multiple-Type Services

In this subsection, we propose an approximate analytic model for multiple-type services (i.e., $n \geq 2$). This model is accurate when θ_i value is small. We present the derivations of P_f and $E[C_d]$ for $n = 2$. The derivations for $n > 2$ can be directly extended and will be briefly described at the end of this subsection.

1) Derivation of P_f ($n = 2$): We first derive the force-termination probability P_f for $n = 2$. When an RU operation is performed, one of the following three cases occurs.

Case A. There is one active type-1 service session, and the residual session holding time is $t_{r,1}$.

Case B. There is one active type-2 service session, and the residual session holding time is $t_{r,2}$.

Case C. Both type-1 and type-2 service sessions are active, and the residual session holding times are $t_{r,1}$ and $t_{r,2}$, respectively.

When the critical RU operation occurs, let P_A, P_B and P_C be the probabilities of Cases A, B and C, respectively. For a sufficiently small θ_i value, P_f can be computed as

$$\begin{aligned} P_f &= P_A \Pr[t_{r,1} > C_{min}] + P_B \Pr[t_{r,2} > C_{min}] \\ &\quad + P_C \Pr[t_{r,1} + t_{r,2} > C_{min}] \quad (14) \end{aligned}$$

In (14) we assume that $C_x(t_7) = C_{min}$, where t_7 is the time that the critical RU operation occurs. Note that $C_{min} \leq C_x(t_7) < C_{min} + \sum_i \theta_i$. Therefore, (14) incurs error when θ_i is large. Substituting (5) into (14) to yield

$$\begin{aligned} P_f &= P_A e^{-\mu_1 C_{min}} + P_B e^{-\mu_2 C_{min}} \\ &\quad + P_C \left(\frac{\mu_1 e^{-\mu_2 C_{min}} - \mu_2 e^{-\mu_1 C_{min}}}{\mu_1 - \mu_2} \right) \quad (15) \end{aligned}$$

In (15), the probabilities P_A, P_B and P_C are derived as follows. We first compute the probability p_i that a type-i service session is active at a random observation point. In Fig. 3, a random observation point occurs at t_2 in the renewal period $[t_0, t_5]$. From the alternating renewal theory [11], p_i is expressed as

$$p_i = \frac{E[t_{h,i}]}{E[t_{h,i}] + E[t_{a,i}]} = \frac{\lambda_i}{\mu_i + \lambda_i} \quad (16)$$

At t_2, there is only one active type-1 service session with probability $p_1(1-p_2)$, there is only one active type-2 service session with probability $p_2(1-p_1)$, and both type-1 and type-2 service sessions are active with probability $p_1 p_2$. The critical RU operation is issued by a type-1 and a type-2 service sessions in Cases A and B, respectively. In Case C, the critical RU operation may be issued by a type-1 or a type-2 services. Since the sending of recharge message can be modeled as a random observer for sufficiently small θ_i value in a service session, from (16), the ratio $P_A : P_B : P_C$ can be computed as

$$\begin{aligned} P_A : P_B : P_C &\approx p_1(1-p_2) : p_2(1-p_1) : 2p_1p_2 \\ &= \lambda_1\mu_2 : \mu_1\lambda_2 : 2\lambda_1\lambda_2 \quad (17) \end{aligned}$$

Since $P_A + P_B + P_C = 1$ and from (17), we have

$$\begin{aligned} &P_A \approx \tfrac{\lambda_1\mu_2}{\lambda_1\mu_2+\mu_1\lambda_2+2\lambda_1\lambda_2}, \quad P_B \approx \tfrac{\mu_1\lambda_2}{\lambda_1\mu_2+\mu_1\lambda_2+2\lambda_1\lambda_2} \\ &\text{and} \quad P_C \approx \tfrac{2\lambda_1\lambda_2}{\lambda_1\mu_2+\mu_1\lambda_2+2\lambda_1\lambda_2} \quad (18) \end{aligned}$$

Substitute (18) into (15) to yield

$$\begin{aligned} P_f &= \big[\lambda_1\mu_2 e^{-\mu_1 C_{min}} + \mu_1\lambda_2 e^{-\mu_2 C_{min}} \\ &\quad + 2\lambda_1\lambda_2 \left(\frac{\mu_1 e^{-\mu_2 C_{min}} - \mu_2 e^{-\mu_1 C_{min}}}{\mu_1 - \mu_2} \right) \big] \\ &\quad \times (\lambda_1\mu_2 + \mu_1\lambda_2 + 2\lambda_1\lambda_2)^{-1} \quad (19) \end{aligned}$$

For $n > 2$, Eq. (14) can be extended by including all active session combinations (there are $2^n - 1$ combinations). Then P_f can be computed following the same derivations for (15)-(19).

2) Derivation of $E[C_d]$ ($n = 2$): For $n = 2$, $E[C_d]$ is derived as follows. For sufficiently small θ_i values, $C_x(t^*) \approx C_{min}$. Therefore, the C_d values are $\max\{C_{min} - t_{r,1}, 0\}$, $\max\{C_{min} - t_{r,2}, 0\}$ and $\max\{C_{min} - t_{r,1} - t_{r,2}, 0\}$ in Cases A, B and C, respectively. We have

$$\begin{aligned} E[C_d] &= P_A E[\max\{C_{min} - t_{r,1}, 0\}] \\ &\quad + P_B E[\max\{C_{min} - t_{r,2}, 0\}] \\ &\quad + P_C E[\max\{C_{min} - t_{r,1} - t_{r,2}, 0\}] \quad (20) \end{aligned}$$

Substitute (5) and (18) into (20) to yield (21) (see next page). For $n > 2$, $E[C_d]$ can be computed through the same derivations for (20)-(21) by considering $2^n - 1$ active session combinations.

The analytic model developed in this section is validated against the simulation experiments. The discrepancies between analytic analysis (specifically, Eqs. (11), (13), (19) and (21)) and simulation are within 2%. The simulation model follows the discrete event approach described in [9], and the details are omitted. The input parameter θ_i is normalized by the mean $1/\mu_i$ of the service session holding time. The input parameter C_{min} and output measure $E[C_d]$ are normalized by the expected credit units c consumed in a session.

IV. Numerical Examples

This section uses numerical examples to investigate the performance of the RTCR mechanism. For the examples in Figs. 5 and 6, $n = 2$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$. Similar results are observed for other parameter values and are not presented. The effects of the input parameters are described as follows.

[Effects of C_{min}.] Fig. 5 plots the force-termination probability P_f and the expected credit $E[C_d]$ against θ_i and C_{min}, where $n = 2$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$. Fig. 5 (a) shows that P_f decreases as C_{min} increases. When the critical RU operation occurs, more unused credit units are available in the prepaid account when C_{min} increases. Therefore, the possibility of force-termination reduces. For $\theta_i = 1/\mu_i$, when C_{min} increases from $2c$ to $4c$, P_f decreases from 18.78% to 4.99%. Fig. 5 (b) shows that $E[C_d]$ increases as C_{min}

$$E[C_d] = C_{min} - \left(\frac{1}{\lambda_1\mu_2 + \mu_1\lambda_2 + 2\lambda_1\lambda_2}\right)\left\{\frac{\lambda_1\mu_2(1-e^{-\mu_1 C_{min}})}{\mu_1} + \frac{\mu_1\lambda_2(1-e^{-\mu_2 C_{min}})}{\mu_2} + \frac{2\lambda_1\lambda_2\left[{\mu_1}^2(1-e^{-\mu_2 C_{min}}) - {\mu_2}^2(1-e^{-\mu_1 C_{min}})\right]}{\mu_1\mu_2(\mu_1-\mu_2)}\right\} \quad (21)$$

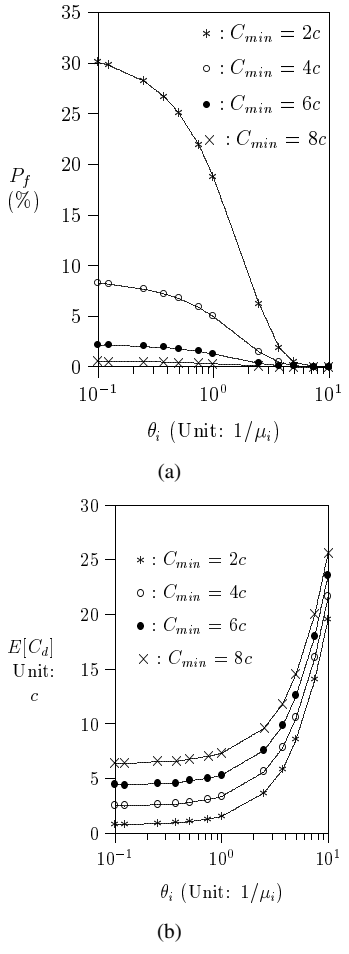


Fig. 5. Effects of θ_i and C_{min} ($n = 2$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$)

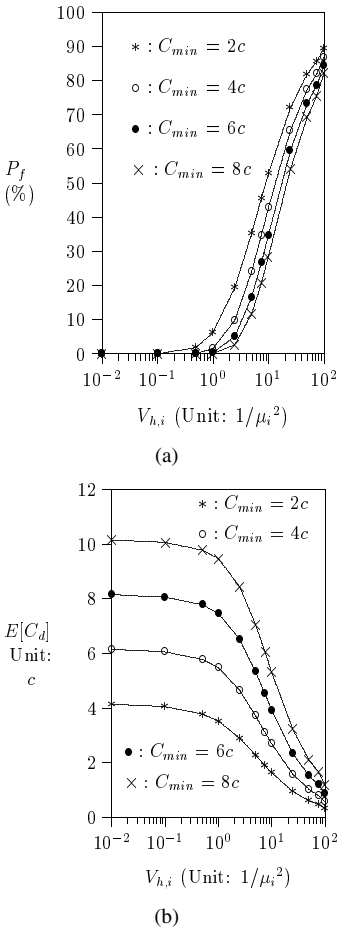


Fig. 6. Effects of $V_{h,i}$ ($n = 2$, $\theta_i = 2.5\mu_i$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$)

increases. It is apparent that when the critical RU operation occurs, the exact unused credit units $C_x(t^*)$ for the user increases as C_{min} increases. That is, the amount of consumed credit reduces, and the expected credit $E[C_d]$ increases. For $\theta_i = 1/\mu_i$, when C_{min} increases from $2c$ to $4c$, $E[C_d]$ increases from $1.60c$ to $3.41c$. In this scenario, we expect that $3.41 - 1.60 = 1.81$ more sessions are complete when $C_{min} = 2c$ than when $C_{min} = 4c$.

[Effects of θ_i.] Fig. 5 (a) shows that P_f is a decreasing function of θ_i. When θ_i increases, more credit units are granted to the AS. Therefore, the possibility of force-termination reduces. For $C_{min} = 6c$, when θ_i increases from $1/\mu_i$ to $2.5/\mu_i$, P_f decreases from 1.32% to 0.37%. This effect becomes insignificant when θ_i is large (e.g., $\theta_i \geq 5/\mu_i$). Fig. 5 (b) shows that $E[C_d]$ is an increasing function of θ_i. For $C_{min} = 6c$, when θ_i increases from $1/\mu_i$ to $2.5/\mu_i$, the $E[C_d]$ increases from $5.37c$ to $7.64c$. Fig. 5 (b) also quantitatively indicates how the θ_i and C_{min} values affect $E[C_d]$. When $\theta_i \leq 1/\mu_i$, $E[C_d] \approx C_{min}$. On the other hand, $E[C_d] >> C_{min}$ as θ_i increases. For example, when $C_{min} = 6c$ and $\theta_i = 10/\mu_i$, $E[C_d] = 23.63c >> 6c$.

[Effects of $V_{h,i}$.] Fig. 6 plots P_f and $E[C_d]$ against C_{min} and the variance $V_{h,i}$ of the Gamma service session holding time $t_{h,i}$, where $n = 2$, $\theta_i = 2.5\mu_i$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$. Fig. 6 (a) shows that P_f increases as $V_{h,i}$ increases. This phenomenon is explained as follows. As $V_{h,i}$ increases, more long and short $t_{h,i}$ periods are observed. The recharge message is more likely to be sent in the long $t_{h,i}$ periods than the short $t_{h,i}$ periods, and larger residual service session holding time $t_{r,i}$ are expected. Therefore, P_f increases as $V_{h,i}$ increases. Fig. 6 (b) shows that $E[C_d]$ decreases as $V_{h,i}$ increases. As $V_{h,i}$ increases, the recharge message is likely to be sent in the long $t_{h,i}$ periods. Then the possibility that $t_{r,i} \geq C_{min}$ (i.e., $C_d = 0$) increases. Therefore $E[C_d]$ decreases as $V_{h,i}$ increases.

V. Conclusions

This paper studied the Recharge Threshold-based Credit Reservation (RTCR) mechanism for UMTS Online Charging System (OCS). In RTCR, when the remaining amount of prepaid credit is below a threshold, the OCS reminds the user to recharge the prepaid account. It is essential to choose an appropriate recharge threshold to reduce the probability that the in-progress service sessions are forced-terminated. An analytic model is developed to compute the expected number $E[N_{r,i}]$ of the RU operations executed in a type-i service session, the force-termination probability P_f and the expected credit $E[C_d]$ left in the user account. We make the following observations:

- P_f decreases as the recharge threshold C_{min} or θ_i increases. This effect becomes insignificant when θ_i is large (e.g., $\theta_i \geq 5/\mu_i$). $E[C_d]$ increases as C_{min} or θ_i increases.
- P_f increases as the variance $V_{h,i}$ of the service session holding time increases, and $E[C_d]$ decreases as $V_{h,i}$ increases.

Based on the above discussion, a mobile operator can select the appropriate C_{min} and θ_i values for various traffic conditions.

Acknowledgment

The authors would like to thank the anonymous reviewers. Their valuable comments have significantly improved the quality of this paper. Their efforts are highly appreciated.

References

[1] 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 6.3.0 (2005-06), 2005.

[2] 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 6.1.0 (2005-06), 2005.

[3] 3GPP, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; IP Multimedia Subsystem (IMS); Stage 2. Technical Specification 3G TS 23.228 version 6.9.0 (2005-03), 2005.

[4] M.-F. Chang, W.-Z. Yang, and Y.-B. Lin, Performance of hot billing mobile prepaid service, *IEEE Trans. Veh. Technol.*, vol. 51, no. 3, pp. 597–612, 2002.

[5] M. A. Clemente and F. J. Velez "Parameters for tele-traffic characterisation in enhanced UMTS," [Online]. *http://seacorn.ptinovacao.pt/*

[6] H. Hakala, L. Mattila, J.-P. Koskinen, M. Stura, and J. Loughney, "Diameter credit-control application," IETF RFC 4006, Aug. 2005.

[7] H.-N. Hung, Y.-B. Lin, N.-F. Peng, and and S.-I. Sou, "Connection failure detection mechanism of UMTS charging protocol," *IEEE Trans. Wireless Commun.*, vol. 5, no. 5, pp. 1180–1186, 2006.

[8] P. Lin, Y.-B. Lin, C.-S. Yen, and J.-Y. Jeng, "Credit allocation for UMTS prepaid service," *IEEE Trans. Veh. Technol.*, vol. 55, no. 1, pp. 306–316, 2006.

[9] 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.

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

[11] S. M. Ross, *Stochastic Processes*. New York: John Wiley & Sons, 1996.

Sok-Ian Sou (S'06) received the B.S.CSIE. and M.S.CSIE degrees from National Chiao Tung University (NCTU), Taiwan, in 1997 and 2004, respectively. She is currently a Ph.D. candidate of the Department of Computer Science, NCTU.

Her current research interests include design and analysis of personal communications services networks, mobile computing, and performance modeling.

Hui-Nien Hung received the B.S.Math. degree from National Taiwan University, Taiwan, in 1989, the M.S.Math. degree from National Tsin-Hua University, Taiwan, in 1991, and the Ph.D. degree in statistics from The University of Chicago in 1996. He is a Professor at the Institute of Statistics, National Chiao Tung University, Taiwan.

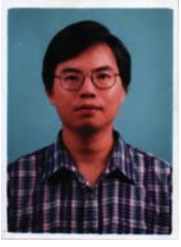
His current research interests include applied probability, financial calculus, bioinformatics, statistical inference, statistical computing, and industrial statistics.

Yi-Bing Lin (M'96-SM'96-F'03) is Chair Professor and Vice President of Research and Development, National Chiao Tung University.

His current research interests include wireless communications and mobile computing. Dr. Lin has published over 190 journal articles and more than 200 conference papers. Lin is the co-author of the book *Wireless and Mobile Network Architecture* (with Imrich Chlamtac; published by John Wiley & Sons). Lin is an IEEE Fellow, an ACM Fellow, an AAAS Fellow, and an IEE Fellow.

Nan-Fu Peng received the B.S. degree in applied mathematics from National Taiwan University, Hsinchu, Taiwan, R.O.C., in 1981, and the Ph.D. degree in statistics from The Ohio State University, Columbus, in 1989.

He is currently an Associate Professor with the Institute of Statistics, National Chiao Tung University. His research interests include Markov chains, population dynamics, and the queueing theory.



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 the design and analysis of personal communications services networks, the development of telecommunication operation support systems, and performance modeling.