Transactions Papers

Credit Pre-Reservation Mechanism for UMTS Prepaid Service

Hsin-Yi Lee and Yi-Bing Lin, *Fellow, IEEE*

***Abstract*—*Online Charging System* (OCS) supports multiple prepaid and postpaid sessions simultaneously. Through credit reservation, the OCS assigns some credit units to a session. These credit units are decremented based on the traffic volume or the duration time. If the assigned credit units are consumed before the session is completed, an *reserve units* (RU) operation is executed to obtain more credit units from the OCS. If the credit at the OCS is depleted, the prepaid session is forced to terminate. During the RU operation, packet delivery is suspended until extra credit units are granted from the OCS. To avoid session suspension during credit reservation, we propose the *credit pre-reservation mechanism* (CPM) that reserves credit earlier before the credit at the GGSN is actually depleted. Analysis and simulation experiments are conducted to investigate the performance of the mechanism. Our study indicates that the CPM can significantly improve the performance of the OCS prepaid mechanism.**

***Index Terms*—Credit reservation, diameter protocol, online charging, prepaid service, Universal Mobile Telecommunications System (UMTS).**

I. Introduction

PRICING, charging and billing are important activities in telecommunications [1], [2], [3]. Advanced mobile telecommunications operation incorporates data applications (specifically, mobile Internet applications [4], [5]) with real-time control and management, which can be archived by a convergent and flexible *Online Charging System* (OCS) [6], [7], [8]. Such convergence is essential to mitigate fraud and credit risks, and provide more personalized advice to users about charges and credit limit controls. The OCS allows simultaneous prepaid and postpaid sessions to be charged in real-time. This feature is important for a telecom operator to deliver multiple sessions simultaneously. Through online charging, the operator can ensure that credit limits are enforced and resources are authorized on a per-transaction basis.

Manuscript received March 8, 2008; accepted May 14, 2008. The associate editor coordinating the review of this paper and approving it for publication was M. Guizani.

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

This work was supported in part by NSC 97-2221-E-009-143-MY3, NSC 98-2221-E-009-059-MY2, NSC 98-2219-E-009-016-, Intel, Chunghwa Telecom, IBM, ITRI and NCTU joint research center, and MoE ATU plan.

Digital Object Identifier 10.1109/TWC.2010.06.080342

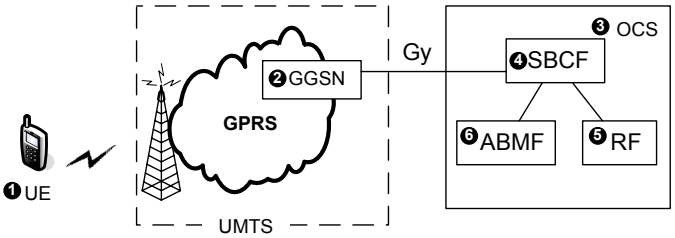


Fig. 1. OCS architecture for mobile telecommunications services.

By merging the prepaid and postpaid methods, the OCS proposed in Universal Mobile Telecommunications System (UMTS) provides two-way communications between network nodes and the OCS to transfer information about rating, billing and accounting. Fig. 1 illustrates how UMTS integrates with the OCS [9]. When a User Equipment (UE; Fig. 1 (1)) originates a prepaid session through the General Packet Radio Service (GPRS), the Gateway GPRS Support Node (GGSN; Fig. 1 (2)) requests the prepaid credit from the OCS (Fig. 1 (3)) through the Gy interface.

In the OCS, the Session Based Charging Function (SBCF; Fig. 1 (4)) module is responsible for online charging of user sessions. The SBCF interacts with the Rating Function (Fig. 1 (5)) to determine the tariff of the requested session. The rating function handles a wide variety of ratable instances, such as data volume, session connection time, and event service (e.g., for web content charging). The SBCF also interacts with the Account Balance Management Function (ABMF; Fig. 1 (6)) for credit control. The ABMF maintains user balances and other account data to check if the users have enough credit. The SBCF assigns some prepaid credit units to the GGSN for a user session. These credit units are decremented at the GGSN in real-time based on either the traffic volume or the duration time. After the assigned credit units are consumed, the GGSN may execute a *reserve units* (RU) operation to ask for more credit from the OCS. If the credit of the user account at the OCS is depleted, the prepaid session is forced to terminate. During the RU operation, packet delivery is suspended until extra credit units are granted from the OCS. To avoid suspension of packet delivery, we propose the *credit pre-reservation mechanism* (CPM) that reserves extra credit earlier

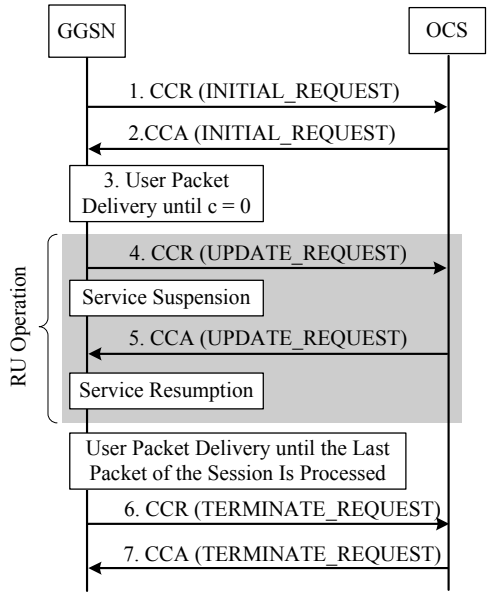


Fig. 2. Message flow of the diameter credit control mechanism.

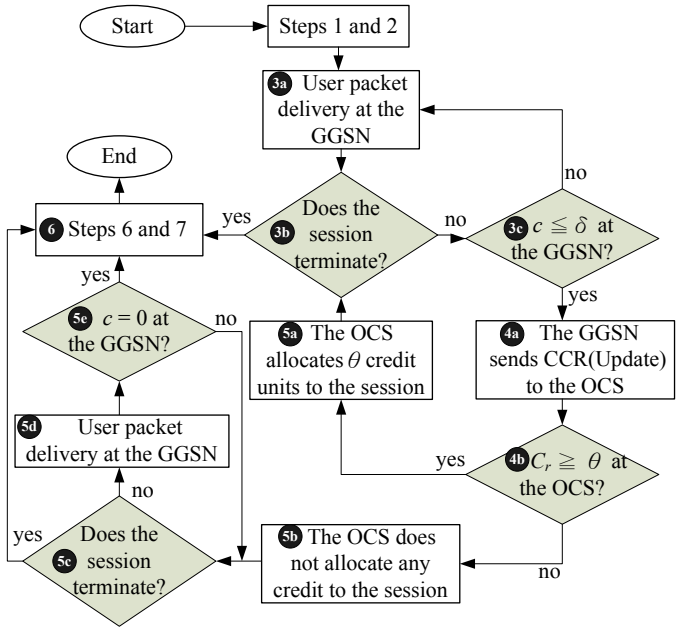


Fig. 3. Flowchart of the CPM (steps 1, 2, 6, and 7 are the same as those in the diameter credit control mechanism).

before the credit units at the GGSN are actually depleted.

The paper is organized as follows. Section 2 describes the Diameter credit control mechanism exercised at the Gy interface. Section 3 proposes the CPM. Section 4 develops an analytical model for the CPM. Section 5 uses numerical examples to investigate the performance of the CPM. Section 6 gives our conclusions.

II. Diameter Credit Control Mechanism

In online charging services, credit control is achieved by exchanging Diameter Credit Control Request (CCR) and Credit Control Answer (CCA) messages through the Gy interface defined between the GGSN and the OCS [10]. The message flow of the online charging is illustrated in Fig. 2 with the following Steps:

Step 1. To initiate the service session, the GGSN sends a CCR message with type INITIAL_REQUEST to the OCS.

Step 2. Through the interactions among the SBCF, the ABMF and the RF, the OCS determines the tariff of the requested service session and then reserves the amount θ of credit units for the service session if the credit balance C_r at the OCS is no less than θ (i.e., $C_r \geq \theta$). Then the OCS returns a CCA message with type INITIAL_REQUEST to the GGSN. This message indicates the amount θ of granted credit to the session.

Step 3. After receiving the CCA message, the GGSN starts to deliver the user packets. Let c be the amount of the remaining credit of the service session. This credit is decremented at the GGSN. At the beginning of this Step, $c = \theta$. During the session, the granted credit units may be depleted (i.e., $c = 0$). If so, the session is suspended and Step 4 is executed.

Step 4. The GGSN sends a CCR message with type UPDATE_REQUEST to the OCS to report the used credit units and request for additional credit units. Packet delivery is suspended, and any newly arriving packets are buffered at the GGSN.

Step 5. The OCS deducts the amount of used credit from the user account and reserves extra amount θ of credit units for the session. Then the OCS acknowledges the GGSN with a CCA message to indicate that θ credit units have been reserved. The session is resumed at the GGSN and the buffered packets are delivered. Note that Steps 3 - 5 may repeat many times before the service session is complete. At Steps 2 and 4, if $C_r < \theta$ at the OCS, then the session is forced to terminate. For the discussion purpose, Steps 4 and 5 are called an *RU operation*.

Step 6. When the session is completed, the GGSN sends a CCR message with type TERMINATE_REQUEST to terminate the session and report the amount of used credit.

Step 7. The OCS deducts the amount of the used credit from the account. Then the OCS acknowledges the reception of the CCR message by sending a CCA message with type TERMINATE_REQUEST.

In the above procedure, when the credit units are depleted at the GGSN, an RU operation is executed to request for additional credit units. The session must be suspended until extra credit units are granted from the OCS. Delayed processing of user packets at the GGSN may seriously degrade the quality of service. To resolve this issue, we propose the credit pre-reservation mechanism in the next section.

III. Credit Pre-reservation Mechanism

In the *credit pre-reservation mechanism* (CPM), we define a threshold δ. When the amount c of the remaining credit for the session is not larger than δ (i.e., $c \leq \delta$), the GGSN conducts an RU operation to request extra credit from the OCS. During the RU operation, the GGSN continues to process the user packets. Hopefully, the GGSN will receive the extra amount θ of credit

units before $c = 0$, and therefore the user packets need not be buffered (i.e., they are not suspended for processing) at the GGSN. Fig. 3 illustrates the flowchart of CPM, which modifies Steps 3-5 in Fig. 2 as follows:

Step 3a. The GGSN delivers the user packets and deducts the reserved credit units.

Step 3b. If the processed packet is the last one of the service session, then Step 6 is executed to terminate the session (the session is successfully completed). Otherwise, the execution proceeds to Step 3c.

Step 3c. Let δ be the CPM threshold. If $c \leq \delta$, Step 4a is executed. Otherwise, the execution proceeds to Step 3a.

Step 4a. The GGSN sends a CCR message with type UPDATE_REQUEST to request for additional credit. During the RU operation, if $c > 0$, the user packets are continuously delivered at the GGSN. When $c = 0$, the session is suspended and the newly arriving packets are buffered.

Step 4b. If the OCS does not have enough credit units (i.e., $C_r < \theta$), Step 5b is executed. Otherwise, Step 5a is executed.

Step 5a. The OCS sends the CCA message to the GGSN to indicate that extra amount θ of credit units have been reserved for the session. Then the execution proceeds to Step 3b. If the last packet arrives during the RU operation, the termination operation (Steps 3b and 6) is executed after the GGSN has received the CCA message. (This is called *delayed termination*) In this case, the session is successfully completed.

Step 5b. The OCS sends the CCA message to the GGSN. This message indicates that no credit is reserved for the session.

Step 5c. If the previously processed packet is the last one of the session, then the session is successfully completed. Step 6 is executed.

Step 5d. The GGSN continues to deliver the user packets.

Step 5e. If $c = 0$, then the session is forced to terminate, and Step 6 is executed. Otherwise, the execution proceeds to Step 5c.

Step 6. The session terminates by executing Steps 6 and 7 in Fig. 2.

In the CPM, if δ is set too small, then the credit units for a session are likely to be depleted and the session must be suspended during the RU operation. On the other hand, if δ is set too large, many credit units are reserved in the active sessions, and the credit in the OCS is consumed fast. In this case, an incoming session has less chance to be served, and an in-progress session is likely to be force-terminated. Therefore, it is important to select an appropriate δ value to "optimize" the CPM performance.

IV. Analytic Model for the CPM

We describe an analytic model to investigate the CPM performance. We assume that the prepaid session arrivals for a user form a Poisson process with rate γ. The inter-arrival time between two packet arrivals has the mean $1/\lambda$. The round-trip transmission delay for the RU operation (i.e., the round-trip message delay for the CCR and CCA message pair) has the mean $1/\mu$. An arrival packet is the last one of the session with probability α; in other words, the session continues with probability $1 - \alpha$, and the expected number of packets delivered in a session is $1/\alpha$.

Initially, a user has C credit units at the prepaid account in the OCS. Without loss of generality, we assume that each user packet consumes one credit unit. Define a *low credit* (LC) *period* as an interval such that during this interval, $c \leq \delta$ for a session. At the beginning of an LC period, the session initiates an RU operation. If more than θ packets arrive during this RU operation, then $\theta - \delta$ packets will be buffered at the GGSN. Consequently, at the end of the RU operation, another RU operation must be issued to obtain more credit units to absorb the buffered packets and to ensure that $c > \delta$ after the buffer is empty. Before an LC period ends, the RU operation may be executed for several times until the session has reserved more than δ credit units. The output measures investigated in our study are listed below.

B: the average number of packets buffered during an RU operation

W: the average packet waiting time

P_r: the probability that during an LC period, two or more RU operations are executed

P_{nc}: the probability that a session is not completely served; i.e., the probability that a new session request is blocked or an in-progress session is forced to terminate

X_s: the average number of the RU operations performed in a session

To derive P_r and B, we first consider the case where $\alpha = 0$; i.e., a session is never terminated. Let K be the number of packets arriving in one RU operation (excluding the first packet arrival that triggers the RU operation). It is clear that

$$P_r = \Pr[K \geq \theta] \tag{1}$$

We assume that an RU operation delay has the Erlang density function $f(t)$ with the shape parameter $b = 2$ and the scale parameter $h = 1/\mu$. (I.e., t is the summation of two Exponential delays. This assumption will be relaxed, and more general distributions will be considered in the simulation model.) Therefore the Laplace-Stieltjes Transform $f^*(s)$ of the RU operation delay is

$$f^*(s) = \left(\frac{\mu}{\mu + s}\right)^2 \tag{2}$$

For $\alpha = 0$, the probability that $K = k$ can be calculated as follows:

$$\Pr[K = k, \alpha = 0] = \int_{t=0}^{\infty} \left[\frac{(\lambda t)^k}{k!}\right] e^{-\lambda t} f(t) dt \tag{3}$$

$$= \left(\frac{\lambda^k}{k!}\right) \int_{t=0}^{\infty} t^k e^{-\lambda t} f(t) dt \tag{4}$$

$$= \left(\frac{\lambda^k}{k!}\right) (-1)^k \left[\frac{d^k f^*(s)}{ds^k}\right]\Bigg|_{s=\lambda} \tag{5}$$

$$= \left(\frac{\lambda^k}{k!}\right) (-1)^k \left[\frac{d^k}{ds^k}\left(\frac{\mu}{\mu + s}\right)^2\right]\Bigg|_{s=\lambda} \tag{6}$$

$$= \frac{\lambda^k (k+1) \mu^2}{(\lambda + \mu)^{k+2}} \tag{7}$$

In (3), the RU operation delay is t with the probability $f(t)dt$. During period t, there are k packet arrivals following the Poisson distribution with the rate λ. Eq. (5) is derived from (4) using Rule P.1.1.9 in [11]. Substitute (2) in (5), we obtain (7).

Now we consider the case when $\alpha \geq 0$. If a session is terminated during an RU operation, $\Pr[K = k]$ is derived by considering the following cases:

(I) When $k = 0$, we have $\Pr[K = 0] = \Pr[K = 0, \alpha = 0]$

(II) When $k > 0$, there are two subcases:

- **(IIa)** There are exactly k packet arrivals during an RU operation (with probability $\Pr[K = k, \alpha = 0]$) and the session is not terminated by any of the first $k-1$ packets (with the probability $(1-\alpha)^{k-1}$). Note that the k-th packet can be the last one of the session.
- **(IIb)** There are more than k packet arrivals during an RU operation if the session is never terminated (with probability $\sum_{i=k+1}^{\infty} \Pr[K = i, \alpha = 0]$), and the session is actually terminated at the k-th packet arrival (with probability $(1-\alpha)^{k-1}\alpha$).

Based on the above cases, $\Pr[K = k]$ is derived as

$$\Pr[K = k] = \begin{cases} \Pr[K = 0, \alpha = 0] & , \text{ k=0} \quad (8) \\ \Pr[K = k, \alpha = 0]\,(1-\alpha)^{k-1} + \displaystyle\sum_{i=k+1}^{\infty} \Pr[K = i, \alpha = 0]\,(1-\alpha)^{k-1}\alpha, & \text{k>0} \quad (9) \end{cases}$$

From (7), (8) can be derived as

$$\Pr[K = 0] = \left(\frac{\mu}{\lambda+\mu}\right)^2 \quad (10)$$

For $k > 0$, Eq. (9) is simplified as

$$\begin{aligned} \Pr[K = k] &= \left[\frac{\lambda^k(k+1)\mu^2}{(\lambda+\mu)^{k+2}}\right](1-\alpha)^{k-1} + \sum_{i=k+1}^{\infty}\left[\frac{\lambda^i(i+1)\mu^2}{(\lambda+\mu)^{i+2}}\right](1-\alpha)^{k-1}\alpha \\ &= \left[\frac{(1-\alpha)^{k-1}\lambda^k}{(\lambda+\mu)^{k+2}}\right] \left\{(k+1)\mu^2 + \alpha\lambda\left[(k+2)\mu+\lambda\right]\right\} \quad (11) \end{aligned}$$

When $\theta > 0$, from (1) and (11), P_r is derived as

$$\begin{aligned} P_r &= \Pr[K \geq \theta] \\ &= \sum_{k=\theta}^{\infty}\left[\frac{(1-\alpha)^{k-1}\lambda^k}{(\lambda+\mu)^{k+2}}\right]\left\{(k+1)\mu^2 + \alpha\lambda\left[(k+2)\mu+\lambda\right]\right\} \\ &= \left[\frac{\lambda^\theta(1-\alpha)^{\theta-1}}{(\lambda+\mu)^{\theta+1}}\right]\left[\mu(\theta+1)+\lambda\right] \quad (12) \end{aligned}$$

When $\theta = 0$,

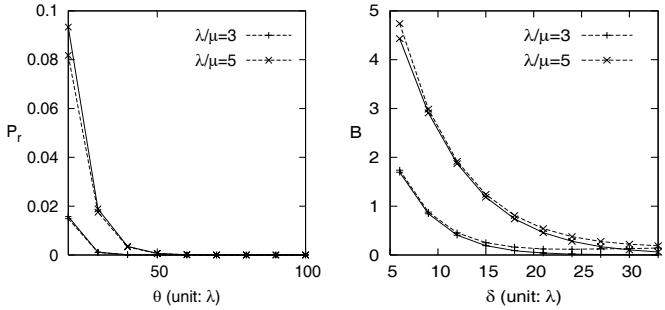
$$P_r = \Pr[K \geq 0] = \left(\frac{\mu}{\mu+\lambda}\right)^2 + \sum_{k=1}^{\infty}\Pr[K \geq k] = 1 \quad (13)$$

The expected number B of buffered packets is derived as follows: When the number k (of packet arrivals during an RU operation) is no less than the threshold δ (i.e., $k \geq \delta$), the session will have $k - \delta$ buffered packets. Therefore, from (11)

$$\begin{aligned} B &= \sum_{k=\delta}^{\infty}(k-\delta)\Pr[K = k] \\ &= \sum_{k=\delta}^{\infty}(k-\delta)\left[\frac{(1-\alpha)^{k-1}\lambda^k}{(\lambda+\mu)^{k+2}}\right] \left\{(k+1)\mu^2 + \alpha\lambda\left[(k+2)\mu+\lambda\right]\right\} \\ &= \left[\frac{(1-\alpha)\lambda}{\lambda+\mu}\right]^{\delta+1} \left[\frac{\delta\mu^2 + \alpha\delta\lambda\mu + 2\mu^2 + 2\lambda\mu + \alpha\lambda\mu + \alpha\lambda^2}{(1-\alpha)(\mu+\alpha\lambda)^2}\right] \quad (14) \end{aligned}$$

The purpose of the analytic model is two folds: First, it partially verifies that the simulation model is correct. Second, it sheds light on the effects of the input parameters on P_r and B. Our simulation model is based on an event-driven approach widely adopted in mobile network studies [12], and the details are elaborated in [13].

Eqs. (12) and (14) validate against the simulation experiments as illustrated in Fig. 4. In this figure, the dashed curves represent the simulation results, and the solid curves represent the analytical results. These curves indicate that the analytic and the simulation results are consistent.



(a) Effect on P_r (b) Effect on B

Fig. 4. Validation of simulation and analytical results on P_r and B ($\alpha = 0.01$, $\gamma/\mu = 1/20$, and $\delta = 0.3\theta$; solid curves: analytic results; dashed curves: simulation results).

V. Numerical Examples

This section uses numerical examples to investigate the performance of the CPM. For the presentation purpose, we assume that the packet termination probability is α = 0.01 and the session arrival rate (normalized by the message delivery rate) is γ = μ/20. For other α and γ/μ values, we observe similar results, which are not presented in this paper. In Figs. 4 and 5, the packet arrivals in a session have a Possion distribution and the RU operation delay has an Erlang distribution. These Expontial-like assumptions are relaxed in Figs. 6 and 7 by considering the Pareto and the Gamma distributions. The effects of the input parameters λ/μ, C, θ and δ are described as follows.

Effects on P_r. Fig. 4 (a) shows how P_r is affected by θ and λ/μ. From (12) and (13), we have

$$\lim_{\theta\to 0} P_r = 1 \quad \text{and} \quad \lim_{\theta\to\infty} P_r = 0 \quad (15)$$

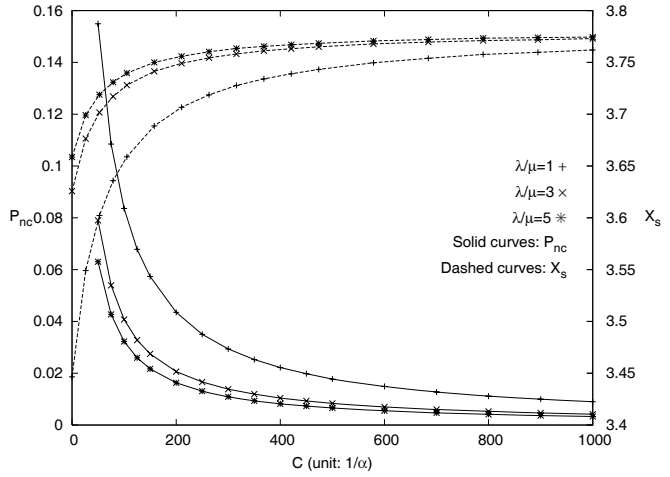


Fig. 5. Effects of λ/μ and C on P_{nc} and X_s (α=0.01,γ/μ=1/20, δ=0.3θ, and θ=50λ).

Therefore, it is obvious that P_r is a decreasing function of θ.

When λ/μ is very small or vary large, we have

$$\lim_{\lambda/\mu\to 0} P_r = (1-\alpha)^{\theta-1} \quad \text{and} \quad \lim_{\lambda/\mu\to\infty} P_r = 0 \quad (16)$$

When λ/μ increases, more packets arrive during an RU operation. When more than θ+δ packets arrive during one RU operation, an extra RU operation will be immediately executed. Consequently, P_r increases. Therefore P_r is an increasing function of λ/μ.

Effects of λ/μ. Fig. 4 (b) shows that B is increasing functions of λ/μ. From (14), we have

$$\lim_{\lambda/\mu\to 0} B = 0 \quad \text{and} \quad \lim_{\lambda/\mu\to\infty} B = \frac{(1-\alpha)^{\delta}}{\alpha} \quad (17)$$

When λ/μ increases, it is likely that more than δ packets will arrive during the interval of the RU operation. Note that W positively co-relates with B. Therefore both B and W increase as λ/μ increases.

When $\lambda/\mu \to \infty$, we found that the B value in (17) is higher than the simulation result (not shown in this Figure), which is explained as follows: When $\lambda/\mu \to \infty$, all packets for a session will arrive before the end of the first RU operation, and therefore the B value in (17) is determined by α. In simulation, the session is always in the "low-credit" status (in the LC period), and every time an RU operation is performed, the number of buffered packets at the end of the operation is reduced. Therefore the expected value B is smaller than that shown in (17). To avoid buffer overflow, the B value in (17) should be considered in system setup.

Fig. 5 shows that P_{nc} increases as λ/μ decreases. Since α is fixed, when λ/μ decreases, the session holding times become longer, and it is likely that more new sessions will arrive during the holding time of an existing session. Therefore, when λ/μ decreases, more sessions will exist at the same time. Suppose that the credit in the OCS suffices to support these sessions if they are sequentially delivered. It is clearly that the OCS may not be able to support these sessions if they are delivered simultaneously. In this case, a newly incoming session is rejected because the credit in the OCS is depleted (while there are unused credit units held in the multiple in-progress sessions). Therefore, P_{nc} increases as λ/μ decreases.

In Fig. 5, X_s is a decreasing function of P_{nc} because the number of RU operations performed in a force-terminated session is less than that in a complete session. Therefore, X_s increases as λ/μ increases.

Effects of *C*. Fig. 5 shows that the output measures (P_{nc} and X_s) are only affected by the "end effect" of C. As C increases, it is more likely that the remaining credit units in the OCS suffice to support one RU operation and such end effect becomes insignificant. Similar to the λ/μ impact, P_{nc} is a decreasing function of C, and X_s is an increasing function of C. Fig. 5 indicates that when C is sufficiently large (e.g. $C \geq 600/\alpha$), the end effect of C can be ignored. Same phenomenon is observed for B and W, and the results are not shown.

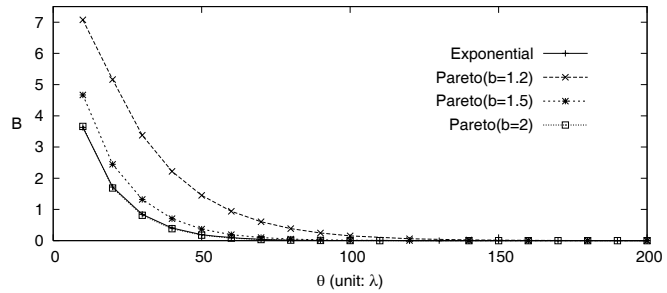
Effects of θ. Fig. 6 intuitively shows that B, W, and X_s are decreasing functions of θ. Since $\delta = 0.3\theta$ in Fig. 6 (b), from (13) and (14), we have

$$\lim_{\theta\to\infty} B = \lim_{\delta\to\infty} B = 0 \quad \text{and}$$
$$\lim_{\theta\to 0} B = \lim_{\delta\to 0} B = \frac{\lambda(2\mu+\alpha\lambda)}{(1-\alpha)(\mu+\alpha\lambda)^2} \quad (18)$$

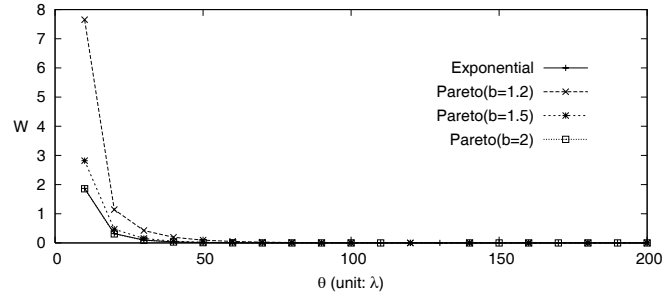
The non-trivial result is that there is a threshold θ value ($\theta \approx 100\lambda$ in Fig. 6) such that beyond this threshold value, increasing θ does not improve the performance. On the other hand, Fig. 6 (c) shows that P_{nc} linearly increases as θ increases. When θ increases, more credit units are reserved in an RU operation, and the credit in the OCS is consumed fast. Therefore, a newly incoming session has less chance to be served, and an in-progress session is likely to be force-terminated.

Effects of packet interarrival time distribution. Fig. 6 considers the packet arrival times with the Exponential and the Pareto distributions with mean $1/\lambda$. In the Pareto distribution, the shape parameter b describes the "heaviness" of the tail of the distribution. It has been shown that the Pareto distribution with $1 \leq b \leq 2$ can approximate the packet traffic very well [14], [15].

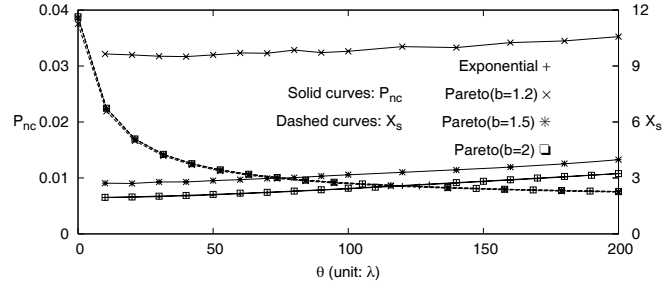
Fig. 6 shows that B, W and P_{nc} are decreasing functions of b. When b decreases, the tail of the distribution becomes longer, and more long packet interarrival times are observed. Since the mean value $1/\lambda$ is fixed for the Pareto distribution in Fig. 6, more long packet interarrival times also imply more short packet interarrival times. The number of short interarrival times must be larger than that of long interarrival times because the minimum of the interarrival time is fixed but the maximum of the interarrival time is infinite. Thus, it is likely that more packets will arrive during an RU operation, and B and W increase. With a small b, it is likely that the last session for a user accommodated by the OCS is a very long session. Before the session is completed, new sessions continue to arrive, and are rejected by the OCS, which



(a) Effect on B

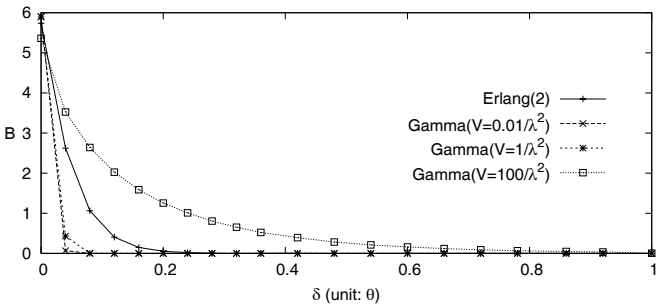


(b) Effect on W

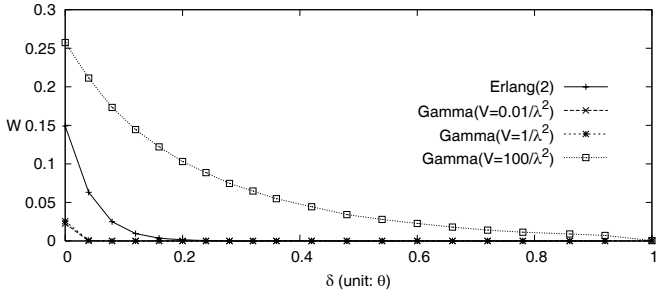


(c) Effect on P_{nc} and X_s

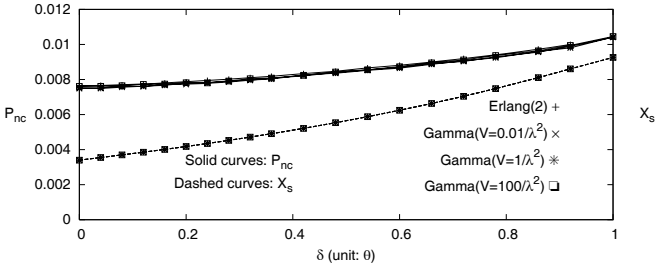
Fig. 6. Effects of θ and the packet interarrival time distribution (α=0.01, γ/μ=1/20, δ=0.3θ, $C = 600/\alpha$, and λ/μ=3).



(a) Effect on B



(b) Effect on W



(c) Effect on P_{nc} and X_s

Fig. 7. Effects of δ and the RU opeartion delay distribution (α=0.01, γ/μ=1/20, θ=100λ, $C = 600/\alpha$, λ/μ=3, and the packet arrival times have the Pareto distribution with the mean $1/\lambda$ and $b = 2$).

contribute to P_{nc}. Therefore P_{nc} increases as b decreases.

Effects of δ. Similar to the effect of θ, Fig. 7 shows that B and W are decreasing functions of δ (see (18)), and P_{nc} is a linearly increasing function of δ. A non-trivial observation is that when $\delta \geq 0.6\theta$, B and W approach to zero. It implies that selecting δ value larger than 0.6θ will not improve the performance. Fig. 7 (c) shows that X_s is an increasing function of δ. When the amount of the credit in a session is less than δ, a CCR message is sent to the OCS. Therefore, for a fixed θ, when δ is increased, X_s increases.

Effects of RU operation delay distribution. Fig. 7 considers the Erlang with $b = 2$ (which is a Gamma distribution with variance $V = 18/\lambda^2$) and Gamma distributed RU operation delays with variances $V = 0.01/\lambda^2, 1/\lambda^2$, and $100/\lambda^2$, respectively.

The figure indicates that P_{nc} and X_s are not significantly affected by the RU operation delay distribution. On the other hand, B and W increase as V increases. As V increases, more long and short RU operation delays are observed. In long RU operation delays, it is likely that more than δ packets arrive. Therefore the packets are more likely to be buffered and delayed processed.

VI. Conclusions

This paper investigated the prepaid services for the UMTS network where multiple prepaid and postpaid sessions are simultaneously supported for a user. We described the prepaid network architecture based on UMTS, and proposed the credit pre-reservation mechanism (CPM) that reserves extra credit earlier before the credit at the GGSN is actually depleted.

An analytic model was developed to compute the average number B of packets buffered during an reserve units (RU) operation and the probability P_r that more than one RU operation is executed during a low credit (LC) period. Simulation experiments are conducted to investigate the performance of CPM. We have the following observations:

- B and W increase as λ/μ increases. The probability P_{nc} that a session is not completely served decreases as λ/μ increases. The avarage number X_s of RU operations performed in a session is an increasing function of λ/μ.
- B, W, X_s and P_{nc} are only affected by the end effect of C. When C is sufficiently large (e.g., $C \geq 600/\alpha$, where α is the probability that an arrival packet is the last one of the session), the end effect can be ignored.
- B, W and X_s decrease but P_{nc} increases as θ increases. There is a threshold θ value (e.g., $\theta \approx 100\lambda$) such

that beyond this threshold value, increasing θ does not improve the CPM performance.
- B and W decrease but X_s increase as δ increases. When δ is large (e.g., $\delta \geq 0.6\theta$), both B and W approach to zero.
- P_r increases when λ/μ increases or θ decreases.
- B, W and P_{nc} increase as the tail of the packet arrival time distribution becomes longer.
- B and W increase as the variance of the RU operation delay increases.

Our study provides guidelines to select the CPM parameters. Specifically, it is appropriate to select $C \geq 600/\alpha$, $\theta \approx 100\lambda$, and $\delta \approx 0.6\theta$.

REFERENCES

[1] P. Lin, H.-Y. Chen, Y. Fang, J.-Y. Jeng, and F.-S. Lu, "A secure mobile electronic payment architecture platform for wireless mobile networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2705-2713, 2008.

[2] J. Lau, and B. Liang, "Optimal pricing for selfish users and prefetching in heterogeneous wireless networks," in *Proc. IEEE International Conference on Communications*, June 2007.

[3] Y.-B. Lin and S.-I. Sou, *Charging for Mobile All-IP Telecommunications*. Wiley, 2008.

[4] D. Wu, T. Hou, and Y.-Q. Zhang, "Transporting real-time video over the Internet: challenges and approaches," *Proc. IEEE*, vol. 88, no. 12, pp. 1855-1875, Dec. 2000.

[5] A.-C. Pang and Y.-K. Chen, "A multicast mechanism for mobile multimedia messaging service," *IEEE Trans. Veh. Technol.*, vol. 53, no. 6, pp. 1891-1902, Nov. 2004.

[6] S.-T. Sou, H.-N. Hung, Y.-B. Lin, N.-F. Peng, and J.-Y. Jeng, "Modeling credit reservation procedure for UMTS online charging system," *IEEE Trans. Wireless Commun.*, vol. 6, no. 11), pp. 4129-4135, Nov. 2007.

[7] S.-I. Sou, Y.-B. Lin, Q. Wu, and J.-Y. Jeng, "Modeling prepaid application server of VoIP and messaging services for UMTS," *IEEE Trans. Veh. Technol.*, vol. 56, no. 3, pp. 1434-1441, May 2007.

[8] S.-I. Sou, Y.-B. Lin, and J.-Y. Jeng, "Reducing credit re-authorization cost in UMTS online charging system," *IEEE Trans. Wireless Commun.*, vol. 7, no. 9, pp. 3629-3635, 2008.

[9] 3GPP, 3rd Generation Partnership Project; Technical Specification Group Service and System Aspects; Telecommunication management; Charging management; Online Charging System (OCS): Applications and interfaces (Release 10), 3G TS 32.296 version 10.0.0 (2010-03), 2010.

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

[11] E. J. Watson, *Laplace Transforms and Applications*. Birkhauserk, 1981.

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

[13] H.-Y. Lee and Y.-B. Lin, "The simulation model for credit pre-reservation mechanism," Technical Report NCTU-08-01, 2008.

[14] M. Cheng and L.-F. Chang, "Wireless dynamic channel assignment performance under packet data traffic," *IEEE J. Sel. Areas Commun.*, vol. 17, pp. 7, pp. 1257-1269, July 1999.

[15] S.-R. Yang and Y.-B. Lin, "Performance evaluation of location management in UMTS," *IEEE Trans. Veh. Technol.*, vol. 52, no. 6. pp. 1603-1615, Nov. 2003.

Hsin-Yi Lee received her B.S.C.S.I.E degree from National Chiao-Tung University in 2003. She is currently a Ph.D student of the Department of Computer Science, National Chiao-Tung University. Her current research interests include wireless and mobile computing systems, and performance modeling.

Yi-Bing Lin (M'95-SM'95-F'03) is Chair Professor of Computer Science, National Chiao Tung University. His current research interests include wireless communications and mobile computing. Lin is the author of the book *Wireless and Mobile Network Architecture* (co-authored with Imrich Chlamtac; published by John Wiley & Sons) and the book *Wireless and Mobile All-IP Networks* (co-authored with Ai-Chun Pang; published by John Wiley & Sons). Lin is an ACM Fellow, an AAAS Fellow, and an IET(IEE) Fellow. He is also an adjunct research fellow of the Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan, and consultant professor of Beijing Jiaotong University, Beijing, China.