Connection Failure Detection Mechanism of UMTS Charging Protocol

Hui-Nien Hung, Yi-Bing Lin, *Fellow, IEEE,* Nan-Fu Peng, and Sok-Ian Sou

***Abstract*— In Universal Mobile Telecommunications System (UMTS), the extension of GPRS tunneling protocol called GTP' is utilized to transfer the Charging Data Records (CDRs) from GPRS Support Nodes (GSNs) to Charging Gateways (CGs). To ensure that the mobile operator receives the charging information, availability for the GTP' transmission is essential. One important issue on GTP' availability is connection failure detection. It is desirable to select appropriate parameter values to avoid false failure detections (e.g., temporary network congestions) while to detect the true failures quickly. We propose an analytic model to compute the false failure detection probability and the expected true failure detection time. Based on our study, the network operator can select the appropriate parameter values for various traffic conditions to reduce the probability of false failure detection and/or true failure detection time.**

***Index Terms*— GPRS Tunneling Protocol extension (GTP'), charging protocol, connection failure detection, Charging Data Record (CDR).**

I. Introduction

UNIVERSAL Mobile Telecommunications System (UMTS) [1], [7] supports high-speed *Packet Switched* (PS) data for accessing versatile multimedia services. The PS *Core Network* is an IP-based backbone network [8]. This core network consists of *GPRS Support Nodes* (GSNs) such as *Serving GSNs* (SGSNs) and *Gateway GSNs* (GGSNs). The *Charging Gateway* (CG) collects the billing and charging information from the GSNs. The GTP' protocol [3] is utilized to transfer the *Charging Data Records* (CDRs) from GSNs to CGs. When a *Mobile Station* is receiving a UMTS PS service, the CDRs are generated based on the charging characteristics (data volume limit, duration limit and so on) of the subscription information for that service. A CG analyzes and possibly consolidates the CDRs from various GSNs, and passes the consolidated data to a billing system.

A CG maintains a *GSN list*. An entry in the list represents a GTP' connection to a GSN. This entry consists of pointers to a *CDR database* and the sequence numbers of possibly duplicated packets. A GSN maintains a list of CGs in the priority order (typically ranges from 1 to 100). If a GSN unexpectedly loses its connection to the current CG, it may send the CDRs to the next CG in the priority list. An entry in the CG list describes parameters for GTP' transmission. After sending a GTP' request, a GSN may not receive a response from the CG due to network failure, network congestion or temporary node unavailability. In this case, 3GPP TS 29.060 [2] defines a mechanism for request retry, where the GSN will retransmit the message until either a response is received within a timeout period or the number of a retry threshold is reached. In the latter case, the GSN-CG communication link is considered disconnected. This paper studies the availability issues for GTP'. Specifically we propose an analytic model to investigate the GTP' connection failure detection mechanism. Our study will provide guidelines for the mobile operators to select the parameters for GTP' connection manipulation.

Manuscript received May 13, 2004; revised December 22, 2004; accepted March 27, 2005. The associate editor coordinating the review of this letter and approving it for publication was Z. Zhang. This work was sponsored in part by NSC Excellence project NSC93-2752-E-0090005-PAE, ITRI/NCTU Joint Research Center, and IIS/Academia Sinica. The work of H.-N. Hung was supported in part by the National Science Council of Taiwan under Grant NSC-92-2118-M-009-013.

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@stat.nctu.edu.tw; nanfu@stat.nctu.edu.tw).

Y.-B. Lin and S.-I. Sou are with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan, R.O.C. (e-mail: liny@csie.nctu.edu.tw; sisou@csie.nctu.edu.tw).

Digital Object Identifier 10.1109/TWC.2006.05026.

II. GTP' Failure Detection Mechanism

This section describes the *Path Failure Detection Algorithm* (PFDA) that detects path failure between the GSN and the CG. In a GSN, an entry in the CG list represents a GTP' connection to a CG. We describe the entry attributes related to PFDA as follow:

- The *CG address* attribute identifies the CG connected to the GSN.
- The *Status* attribute indicates if the connection is "active" or "inactive".
- The *Charging Packet Ack Wait Time* (T_r) is the maximum elapsed time the GSN is allowed to wait for the acknowledgement of a charging packet; typical allowed values range from 1 millisecond to 65 seconds.
- The *Maximum Number of Charging Packet Tries* (L) is the number of attempts (including the first attempt and the retries) the GSN is allowed to send a charging packet; typical L range is $1 - 16$. When $L = 1$, it means that there is no retry.
- The *Maximum Number of Unsuccessful Deliveries* (K) is the maximum number of consecutive failed deliveries that are attempted before the GSN considers a connection failure occurs. Note that a *delivery* is considered failed (or timed out if it has been attempted for L times without receiving any acknowledgement from the CG).
- The *Unsuccessful Delivery Counter* (N_K) attribute records the number of the consecutive failed delivery attempts.
- The *Unacknowledged Buffer* stores a copy of each GTP' message that has been sent to the CG but has not been acknowledged. A record in the unacknowledged

buffer consists of an *Expiry Timestamp* t_e, the *Charging Packet Try Counter* (N_L) and an unacknowledged GTP' message. The expiry timestamp t_e is equal to T_r plus the time when the GTP' message was sent, which represents the expiry of the message. The counter N_L counts the number of the first attempt and retries that have been performed for this charging packet transmission.

PFDA works as follows:

Step 1. After the connection setup procedure is complete, both N_L and N_K are set to 0, and the *Status* is set to "active". At this point, the GSN can send GTP' messages to the CG.

Step 2. When a GTP' message is sent from the GSN to the CG at time t, a copy of the message is stored in the unacknowledged buffer, where the expiry timestamp is set to $t_e = t + T_r$.

Step 3. If the GSN has received the acknowledgement from the CG before t_e, both N_L and N_K are set to 0.

Step 4. If the GSN has not received the acknowledgement from the CG before t_e, N_L is incremented by 1. If $N_L = L$, then the charging packet delivery is considered failed. N_K is incremented by 1.

Step 5. If $N_K = K$, then the GTP' connection is considered failed. The *Status* is set to "inactive".

When Step 5 of PFDA is encountered, it is assumed that the path between the GSN and the CG is no longer available, and the GSN is switched to another CG. However, besides link failure, unacknowledged packet transfers may also be caused by temporary network congestion. In this case, it is not desirable to perform CG switching (which is a very expensive operation). A simple way to avoid this kind of "*false*" *failure detection* is to set large values for parameters T_r, L and K. On the other hand, large parameter values may result in delayed detection of "*true*" failures. Therefore, it is important to select appropriate parameter values so that true failures can be quickly detected while false failures can be avoided. Based on the GTP' mechanism described in this section, we derive the probability of false failure detection in Section III, and compute the expected detection time of true failure in Section IV.

III. Probability of False Failure Detection

Let random variable t_f be the lifetime between when the GTP' connection is established and when a true failure occurs. During this period, undesirable false failures (temporary network congestions) may be detected, and the GSN is unnecessarily switched to another CG. Let α be the probability that the PFDA detects a false failure (and therefore the GSN is switched to another CG before a true failure occurs). Suppose that t_f has the density function $f_f(t_f)$. Let the arrivals of charging packets be a Poisson stream with rate λ_c. Note that the charging packets delivered between a GSN and the CG are generated by all users in this GSN. Each CDR stream of an individual user may have an arbitrary distribution, but the net traffic of all users becomes a Poisson stream [11]. We observe that the charging packets forms a Poisson stream when there are more than 20 users. Let the Echo message arrivals be a deterministic stream with the fixed interval T_e. For any reasonable setting, an Echo message should not be issued before the previous one is acknowledged or timed out. Thus, in CG configuration, we set

$$T_e \geq LT_r \quad (1)$$

Let random variable $N_c(t_f)$ be the number of charging packet arrivals (excluding retries) during the lifetime t_f of the GTP' connection. Then

$$\Pr[N_c(t_f) = n] = \left[\frac{(\lambda_c t_f)^n}{n!}\right] e^{-\lambda_c t_f} \quad (2)$$

Let random variable $N_e(t_f)$ denote the number of Echo message arrivals (excluding retries) during t_f. That is

$$N_e(t_f) = \lfloor t_f/T_e \rfloor \quad (3)$$

Let $N(t_f)$ be the number of GTP' messages (excluding retries) that the GSN attempts to deliver to the CG during t_f. That is, $N(t_f) = N_e(t_f) + N_c(t_f)$. From (3), $N(t_f) = \lfloor t_f/T_e \rfloor + N_c(t_f)$. Therefore, for a given t_f , (2) can be re-written as

$$\Pr[N(t_f) = \lfloor t_f/T_e \rfloor + n] = \left[\frac{(\lambda_c t_f)^n}{n!}\right] e^{-\lambda_c t_f} \quad (4)$$

Let random variable t_r be the round-trip transmission delay (between the GSN and the CG) for a GTP' message attempt. We assume that t_r has a distribution $F_r(t_r)$ and the density function $f_r(t_r)$. From Step 4 of PFDA, a transmission is timed out with probability $\Pr[t_r \geq T_r]$. From Step 5 of PFDA, a delivery is timed out (after it has been tried for L times) with probability p, where

$$p = (\Pr[t_r \geq T_r])^L = [1 - F_r(T_r)]^L \quad (5)$$

The GTP' connection is considered disconnected after K consecutive delivery timeouts where each of the deliveries fails for L attempts (see Step 5 of PFDA). Since the GTP' path is connected during t_f, a false failure is detected if Step 5 of PFDA is executed when the j-th GTP' message delivery is timed out, where $j \leq N(t_f)$. Let $\theta(j)$ denote the probability that such false failure is detected at the j-th delivery. Assume that the delivery results (i.e., a success or a failure) are independent. Based on the relationship between j and K, $\theta(j)$ is derived in three cases:

Case I. $0 \leq j < K$. It is clear that $\theta(j) = 0$.

Case II. $j = K$. It is clear that $\theta(j) = p^K$.

Case III. $j > K$. In this case, no false failure is detected before the $(j-K-1)$-th delivery, the $(j-K)$-th delivery is a success, and the last K deliveries are timed out. Therefore, $\theta(j) = \left[1 - \sum_{i=0}^{j-K-1} \theta(i)\right](1-p)p^K$.

From (5) and the three cases described above, we have

$$\theta(j) = \begin{cases} 0 & ,0 \leq j < K \\ p^K & ,j = K \\ \left[1 - \sum_{i=0}^{j-K-1} \theta(i)\right](1-p)p^K & ,j > K \end{cases} \quad (6)$$

For $K = 1$ and $j \geq 1$, (6) is simplified as $\theta(j) = (1-p)^{j-1}p$. In this case, $\theta(j)$ becomes a geometric distribution. Let $\bar{\theta}(j)$ be the probability that no false failure is detected before (and including) the j-th GTP' message delivery. Then

$$\bar{\theta}(j) = 1 - \sum_{i=0}^{j} \theta(i) \quad (7)$$

From (4) and (7), the probability α of false failure detection is

$$\alpha = 1-\int_{t_f=0}^{\infty}\sum_{n=0}^{\infty}\bar{\theta}(\lfloor t_f/T_e\rfloor+n) \times\Pr[N(t_f)=\lfloor t_f/T_e\rfloor+n]f_f(t_f)dt_f \quad (8)$$

The derivation for (8) can be extended by assuming that the lifetime t_f has an exponential distribution with rate λ_f. The exponential distribution is chosen because it has often been used in reliability and lifetime modeling [10]. We note that our result can be easily generalized for t_f with mixed-Erlang distribution with a tedious routine. Eq. (8) is re-written as

$$\alpha = 1-\lambda_f\sum_{k=0}^{\infty}\sum_{n=0}^{\infty}\bar{\theta}(k+n)\left[\frac{\lambda_c^n}{(\lambda_c+\lambda_f)^{n+1}}\right] \times\sum_{j=0}^{n}\left\{\frac{e^{-(\lambda_c+\lambda_f)kT_e}[(\lambda_c+\lambda_f)T_e]^j}{j!}\right\} \times\left[k^j-e^{-(\lambda_c+\lambda_f)T_e}(k+1)^j\right] \quad (9)$$

IV. Expected True Failure Detection Time

This section derives the expected detection time of "*true*" failure. Consider the timing diagram in Fig. 1(a), where a failure occurs at time t_f and is detected at time t_d. The detection time for the failure is $\tau_d = t_d - t_f$. Let random variable $N_K(t)$ represent the N_K value at time t. If $N_K(t_f) = K-n$ (for $0 < n \le K$), then the GTP' connection failure is detected when n more GTP' message deliveries are timed out. Consider a GTP' message sent from the GSN to the CG. The GSN either receives an acknowledgement from the CG or the delivery (i.e., the L-th transmission for this message) is timed out at time t^*. This time t^* is denoted as the *departure time* of the GTP' message delivery. For $1 \le i \le n$, let $t_{d,i}$ be the departure time of the i-th failed GTP' message delivery after t_f. Note that $t_d = t_{d,n}$. In Fig. 1(b), the arrival times $t_{a,i}$ (for $1 \le i \le n$) correspond to the GTP' message deliveries with the departure times $t_{d,i}$ in Fig. 1(a). It is apparent that $t_{a,i} = t_{d,i} - LT_r$. Note that these arrivals may occur before or after t_f. In Fig. 1(b), the first j' deliveries arrive before t_f. If

$$t_{a,n} > t_f \quad (10)$$

then the true failure detection time τ_d is

$$\tau_d = t_{d,n} - t_f = t_{a,n} + LT_r - t_f \quad (11)$$

In this section, we compute the probability that $N_K(t_f) = K-n$ (for $0 < n \le K$). This probability is used to derive $E[\tau_d|t_{a,n} > t_f]$. Then $E[\tau_d]$ is computed from $E[\tau_d|t_{a,n} > t_f]$ derived in the following subsections and $E[\tau_d|t_{a,n} \le t_f]$ derived in [12].

A. Derivation for the $N_K(t_f)$ distribution

We first compute $\Pr[N_K(t_f)=0]$. Then we use this result to derive $\Pr[N_K(t_f)=j]$ (for $1 \le j \le K-1$). It is clear that t_f lies in two consecutive Echo message arrivals. Suppose that these two Echo messages arrive at times t_0 and t_0+T_e, respectively (see Fig. 2). Since t_f is a random observer, it is uniformly distributed over $[t_0, t_0{+}T_e)$. Let random variable $N_{K\to\infty}(t)$ be the N_K value at time t when $K\to\infty$. In interval $[t_0, t_0{+}T_e)$, $\{N_{K\to\infty}(t); t\in[t_0,t_0{+}T_e)\}$ is a continuous time, discrete state stochastic process (the state space is $0,1,2,...$). There exists j such that for $1\le i\le j$ the interval $[t_0, t_0+T_e)$ consists of j alternative periods (x_i, y_i), where

$$N_{K\to\infty}(t)\begin{cases} =0 & \text{, for } t \text{ in one of the } x_i \text{ periods}\\ >0 & \text{, for } t \text{ in one of the } y_i \text{ periods}\end{cases}$$

If $N_{K\to\infty}(t_0)\neq 0$, then x_1=0. Similarly, if $N_{K\to\infty}(t_0{+}T_e)$=0, then y_j=0. Let $X=\sum_{i=1}^{j}x_i$ and $Y=\sum_{i=1}^{j}y_i$. Then

$$\Pr[N_{K\to\infty}(t)=0]=\frac{E[X]}{E[X]+E[Y]}=\frac{E[X]}{T_e} \quad (12)$$

From (12), $\Pr[N_{K\to\infty}(t)=j]$ (for $j>0$) is expressed as

$$\Pr[N_{K\to\infty}(t)=j]=(1-p)p^{j-1}(1-E[X]/T_e) \quad (13)$$

In (13), the last GTP' message arrival before t is timed out with probability $(1-E[X]/T_e)$, and the probability that there are exact $j-1$ delivery timeouts before this last GTP' message delivery is $(1-p)p^{j-1}$. Suppose that no false failure is detected before t_f. Under this condition, $N_K(t_f)$ ranges from 0 to $K-1$. From (12) and (13), we have

$$\Pr[N_K(t_f)=j]=\begin{cases}\frac{E[X]}{T_e-p^{K-1}(T_e-E[X])} & ,j=0\\ \frac{(1-p)p^{j-1}(T_e-E[X])}{T_e-p^{K-1}(T_e-E[X])} & ,0<j<K\end{cases} \quad (14)$$

In (14), $E[X]$ is derived as follows. Let t_l $(0<t_l\le LT_r)$ be the delivery delay for a GTP' message delivery (including retries). In Fig. 2, $k>0$ departures occur in $[t_0, t_0{+}T_e)$, where the i-th departure occurs at t_i (for $1\le i\le k$). Let $t_{k+1}=t_0{+}T_e$ be the arrival time of the next Echo message. According to (1), the departure of the previous Echo message must occur in $(t_0, t_0{+}T_e)$. Suppose that this departure is the j-th departure where $j\le k$. By considering whether the previous Echo message delivery fails or successes, we express $E[X]$ as

$$E[X] = E[X|t_l=LT_r]\Pr[t_l=LT_r] + E[X|t_l<LT_r]\Pr[t_l<LT_r] \quad (15)$$

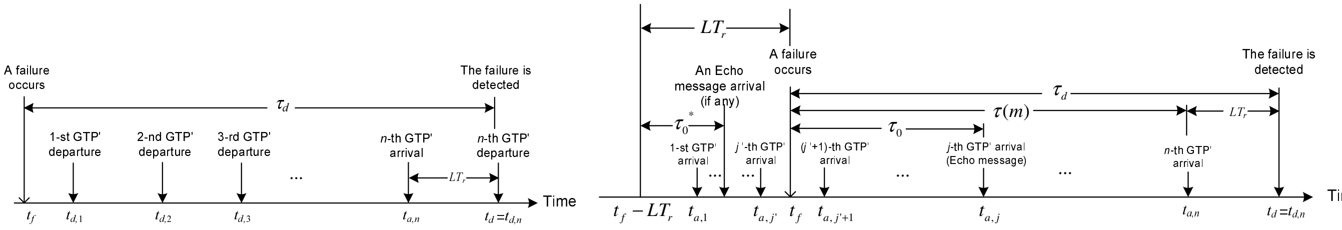
$E[X|t_l{=}LT_r]$ is derived as follows. When $t_l{=}LT_r$, the previous Echo message delivery fails. That is, $t_j{=}t_0{+}LT_r$ and $N_{K\to\infty}(t_j)\neq 0$. Let $z_i{=}t_{i+1}-t_i$ for $0\le i\le k$. Since the N_K value is only changed at times when departures occur, z_i contributes to $E[X|t_l{=}LT_r]$ if $N_{K\to\infty}(t_i)$=0. Let $C=\Pr[N_{K\to\infty}(t_0)=0]$. For $j\le k$, we have

$$E[X|t_l=LT_r] = (1-p)\left\{E\left[\sum_{i=0}^{j-1}z_i\right]+E\left[\sum_{i=j+1}^{k}z_i\right]\right\} +CE[z_0] \quad (16)$$

Since $\sum_{i=1}^{j-1}z_i=LT_r-z_0$ and $\sum_{i=j+1}^{k}z_i=T_e-LT_r-z_j$, (16) is re-written as

$$E[X|t_l=LT_r] = (1-p)(T_e-E[z_j])+(C+p-1)E[z_0] \quad (17)$$

In (17), $C=\Pr[N_{K\to\infty}(t_0)=0]$ is derived in [12]. $E[z_0]$ is derived as follows. If the first charging packet departure



(a) Departures after a true failure

(b) Arrivals corresponding to the departures in (a) where $t_{a,n} > t_f$

Fig. 1. Timing Diagram for Detecting True Failure $(n \le K)$

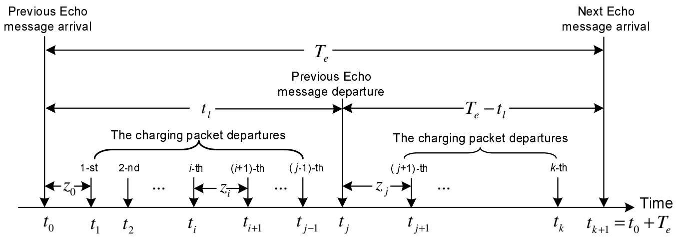


Fig. 2. Timing Diagram for Deriving $E[X]$

occurs before $t_0 + LT_r$, then z_0 is exponentially distributed under the condition that $z_0 < LT_r$. That is

$$\begin{aligned} & E[z_0 | z_0 < LT_r] \Pr[z_0 < LT_r] \\ = & \frac{1}{\lambda_c}\left(1 - e^{-\lambda_c LT_r}\right) - LT_r e^{-\lambda_c LT_r} \end{aligned} \quad (18)$$

If the first charging packet departure occurs after $t_0 + LT_r$, then $z_0 = LT_r$. In this case

$$E[z_0 | z_0 = LT_r] \Pr[z_0 = LT_r] = LT_r e^{-\lambda_c LT_r} \quad (19)$$

Combining (18) and (19) to yield

$$E[z_0] = \frac{1}{\lambda_c}\left(1 - e^{-\lambda_c LT_r}\right) \quad (20)$$

Following similar derivation, $E[z_j]$ can be expressed as

$$E[z_j] = \frac{1}{\lambda_c}\left[1 - e^{-\lambda_c (T_e - LT_r)}\right] \quad (21)$$

From (17), (20) and (21), we have

$$\begin{aligned} & E[X | t_l = LT_r] \Pr[t_l = LT_r] \\ = & \; p\left\{(1-p)T_e + \frac{1}{\lambda_c}\left\{(C+p-1)\left(1 - e^{-\lambda_c LT_r}\right)\right.\right. \\ & \left.\left. -(1-p)\left[1 - e^{-\lambda_c (T_e - LT_r)}\right]\right\}\right\} \end{aligned} \quad (22)$$

$E[X|t_l < LT_r]$ is derived as follows. When $0 < t_l < LT_r$, the previous Echo message delivery successes. That is, $t_j = t_0 + t_l < t_0 + LT_r$ and $N_{K\to\infty}(t_j) = 0$. Let $z_i(t_l)$ be the z_i value for a specific $t_l < LT_r$. Then for $t_l < LT_r$,

$$\begin{aligned} E[X|t_l] = & (1-p)\left\{E\left[\sum_{i=1}^{j-1} z_i(t_l)\right] + E\left[\sum_{i=j+1}^{k} z_i(t_l)\right]\right\} \\ & + CE[z_0(t_l)] + E[z_j(t_l)] \end{aligned} \quad (23)$$

Following similar derivation for (22), for $t_l < LT_r$,

$$\begin{aligned} E[X|t_l] = & \frac{1}{\lambda_c}\left\{(C+2p-1) - (C+p-1)e^{-\lambda_c t_l}\right. \\ & \left. - p e^{-\lambda_c T_e} e^{\lambda_c t_l}\right\} + (1-p)T_e \end{aligned} \quad (24)$$

Suppose that t_l has the density function $f_l(t_l)$ and the distribution function $F_l(t_l)$. If the previous Echo message is successfully delivered, the delivery delay is $0 < t_l < LT_r$ with probability $f_l(t_l)dt_l$. Therefore,

$$E[X|t_l < LT_r] \Pr[t_l < LT_r] = \int_{t_l=0}^{LT_r} E[X|t_l] f_l(t_l) dt_l \quad (25)$$

where $E[X|t_l]$ is expressed in (24), and $f_l(t_l)$ is derived in [12]. Then $E[X]$ can be obtained from (15), (22) and (25). Finally, $\Pr[N_K(t_f) = j]$ can be computed by using (14) and (15).

B. Derivation for $E[\tau_d]$

For $t_{a,n} > t_f$, let $m > 0$ denote the number of failed GTP' message arrivals occurring after t_f. Note that m is not necessarily equal to $K - N_K(t_f)$ because some GTP' message arrivals may occur before t_f and are timed out after t_f. Such messages are denoted as *cross* messages ("cross" means that the delivery delay "crosses" the time point t_f). Therefore, the departures of cross messages are not accurately counted in $N_K(t_f)$. Fortunately, we know that these departures must occur by $t_f + LT_r$, and therefore $m = K - N_K(t_f + LT_r)$. $N_K(t_f + LT_r)$ can be derived from $N_K(t_f)$ as follows. Let n_c and n_e denote the numbers of cross charging packets and cross Echo messages, respectively (in Fig. 1(b); $j' = n_c + n_e$). It can be observed that

$$N_K(t_f + LT_r) = \min\{N_K(t_f) + n_c + n_e, K\} \quad (26)$$

Note that when $m = K - N_K(t_f + LT_r) = 0$, we have $t_{a,n} \le t_f$. In this special case, $m = 0$ and $E[\tau_d | m = 0]$ is derived in [12]. Now assume that $m > 0$. Since the deliveries of charging packets can be modeled by the M/G/∞ system and t_f is a random observer of the system, n_c can be represented by a Poisson random variable with parameter ρ (see Chapter 2.4 in [9]), where

$$\rho = \lambda_c \int_{t_l=0}^{LT_r} [1 - F_L(t_l)] dt_l \quad (27)$$

and the probability mass function of n_c is given by

$$\Pr[n_c = i] = \left(\frac{\rho^i}{i!}\right) e^{-\rho} \quad (28)$$

In Fig. 1(b), let $t_{a,j}$ (for $n_c + n_e < j$) be the arrival time of the first Echo message occurring after t_f, and $\tau_0 = t_{a,j} - t_f$. Since $T_e \ge LT_r$, the n_e value is either 0 or 1. Let $\Pr[n_e =$

$1|\tau_0]$ be the probability that $n_e = 1$ for a specific τ_0. Then $\Pr[n_e = 1|\tau_0]$ can be expressed as

$$\Pr[n_e = 1|\tau_0] = \begin{cases} 0 & ,\tau_0 \le T_e - LT_r \\ 1 - F_L(T_e - \tau_0) & ,\tau_0 > T_e - LT_r \end{cases} \quad (29)$$

where $F_L(t)$ is derived in [12]. In (29), when $\tau_0 \le T_e - LT_r$, there is no undelivered Echo message before t_f. When $\tau_0 > T_e - LT_r$, an Echo message arrival occurs in period $[t_f - LT_r, t_f)$. This Echo message delivery fails before t_f with probability $\Pr[n_e = 1|\tau_0] = 1 - F_L(T_e - \tau_0)$. From (28) and (29), $\Pr[n_c + n_e = j'|\tau_0]$ can be expressed as

$$\begin{aligned} & \Pr[n_c + n_e = j'|\tau_0] \\ = & \begin{cases} e^{-\rho}\left(1 - \Pr[n_e = 1|\tau_0]\right) & ,j' = 0 \\ e^{-\rho}\left[\frac{\rho^{j'-1}}{(j'-1)!}\right]\Pr[n_e = 1|\tau_0] & \\ \quad +\frac{\rho^{j'}}{j'!}\left(1 - \Pr[n_e = 1|\tau_0]\right) & ,j' > 0 \end{cases} \end{aligned} \quad (30)$$

Therefore, for $i \le j < K$, $\Pr[N_K(t_f + LT_r) = j|\tau_0]$ can be computed from $\Pr[N_K(t_f) = i]$ and (30) as

$$\begin{aligned} & \Pr[N_K(t_f + LT_r) = j|\tau_0] \\ = & \sum_{i=0}^{j} \Pr[N_K(t_f) = i]\Pr[n_c + n_e = j - i|\tau_0] \end{aligned} \quad (31)$$

For $m > 0$, let $\tau(m) = t_{a,n} - t_f$ (see Fig. 1(b)). $E[\tau(m)]$ is derived as follows. Let m_c and m_e denote the numbers of charging packet arrivals and Echo message arrivals occurring in period $\tau(m)$. That is, $m = m_c + m_e = n - (n_c + n_e) > 0$. We have

$$m_e = \lfloor(\tau(m) - \tau_0)/T_e\rfloor + 1 \quad (32)$$

If $\tau_0 > \tau(m)$, then $m_e = 0$. Let τ_e be the interval between t_f and the arrival time of the m_e-th Echo message after t_f. By convention, $\tau_e = 0$ for $m_e = 0$. Let τ_c be the interval between t_f and the arrival time of the m_c-th charging packet after t_f. Then $\tau(m)$ =max$\{\tau_c, \tau_e\}$. Note that m_e is determined by $\tau(m)$ and τ_0 (see (32)), and therefore τ_e and τ_c are dependent of each other. Since the arrivals of charging packets are a Poisson stream, τ_c has the Erlang distribution with mean m_c/λ_c and shape parameter m_c. For $m > 0$, the distribution function $F_c(\tau_c)$ of τ_c is

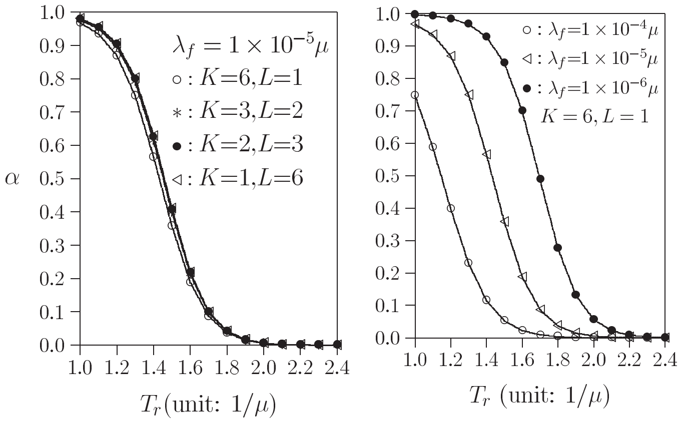
$$F_c(\tau_c) = 1 - \sum_{i=0}^{m_c-1}\left[\frac{(\lambda_c\tau_c)^i}{i!}\right]e^{-\lambda_c\tau_c} \quad (33)$$

For $m > 0$, let $F_m(\tau(m))$ be the distribution function of $\tau(m)$. From (32) and (33), we have

$$\begin{aligned} & F_m(\tau(m)|\tau_0) = F_c(\tau(m)|\tau_0) \\ = & 1 - \sum_{i=0}^{m-\left\lfloor\frac{(\tau(m)-\tau_0)}{T_e}\right\rfloor-2}\left\{\frac{[\lambda_c\tau(m)]^i}{i!}\right\}e^{-\lambda_c\tau(m)} \end{aligned} \quad (34)$$

Note that $F_m(\tau(m)|\tau_0)$ is discontinuous at points $\tau(m) = \tau_0 + jT_e$, for $j = 0, 1, ..., m_e - 1$. From (34) we have

$$\begin{aligned} & \Pr[\tau(m) = \tau_0 + jT_e|\tau_0] \\ = & F_m(\tau_0 + jT_e|\tau_0) - F_m(\tau_0 + jT_e{}^-|\tau_0) \\ = & \left\{\frac{[\lambda_c(\tau_0 + jT_e)]^{m-j-1}}{(m-j-1)!}\right\}e^{-\lambda_c(\tau_0+jT_e)} \end{aligned} \quad (35)$$



(a) Effect of L (b) Effect of λ_f

Fig. 3. Effects of T_r, L and λ_f on α ($\lambda_c = \mu/18$)

Eq. (35) says that the m-th GTP' message arrival is the $(j+1)$-th Echo message, and there are $m - j - 1$ charging packets occurring in period $\tau(m)$, which has the Poisson distribution with parameter λ_c. For a given τ_0 and $m > 0$, the expected value of $\tau(m)$ is

$$\begin{aligned} E[\tau(m)|\tau_0] &= \int_{\tau(m)=0}^{\infty}[1 - F_m(\tau(m)|\tau_0)]d\tau(m) \\ = & \left(\frac{1}{\lambda_c}\right)\sum_{i=0}^{m-1}\left\{1 - e^{-\lambda_c[\tau_0+(m-i-1)T_e]}\right\} \\ & \times\left\{\sum_{j=0}^{i}\frac{\{\lambda_c[\tau_0 + (m-i-1)T_e]\}^j}{j!}\right\} \end{aligned} \quad (36)$$

Since t_f is a random observer of the inter-Echo arrival times, τ_0 is uniformly distributed over $(0, T_e]$. From (11), (31) and (36), the expected value of $E[\tau_d]$ is expressed as

$$\begin{aligned} E[\tau_d] &= E[\tau_d|m > 0]\Pr[m > 0] + E[\tau_d|m = 0]\Pr[m = 0] \\ &= \left(\frac{1}{T_e}\right)\sum_{m=1}^{K}\int_{\tau_0=0}^{T_e}(E[\tau(m)|\tau_0] + LT_r) \\ &\quad \times \Pr[N_K(t_f + LT_r) = K - m|\tau_0]d\tau_0 \\ &\quad + E[\tau_d|m = 0]\Pr[m = 0] \end{aligned} \quad (37)$$

where $E[\tau_d|m = 0]$ and $\Pr[m = 0]$ are derived in [12].
The analytic model developed in this paper is validated against the simulation experiments. The discrepancies between analytic analysis (specifically, Eqs. (9) and (37)) and simulation are within 3% in most cases. The simulation technique used in this paper is similar to the one described in [6], and the details are omitted.

V. Numerical Examples

Based on the analytic model developed in the previous section, we show how K, L and T_r affect the probability α of false failure detection and the expected time $E[\tau_d]$ of true failure detection. We assume that the round-trip transmission delay t_r between a GSN and a CG has a hyper-Erlang distribution with the expected value $1/\mu = \sum_{i=1}^{M}\beta_i/\mu_i$ and the distribution function

$$F_r(t_r) = 1 - \sum_{i=1}^{M} \beta_i \left\{ \sum_{j=0}^{m_i-1} \left[\frac{(m_i\mu_i t_r)^j}{j!} \right] e^{-m_i\mu_i t_r} \right\} \quad (38)$$

where $M, m_1, m_2, ..., m_M$ are nonnegative integers, $\mu_i > 0$, $\beta_i > 0$, and $\sum_{i=1}^{M} \beta_i = 1$. The hyper-Erlang distribution is selected because this distribution has been proven as a good approximation to many distributions as well as measured data [4], [5]. From (5) and (38)

$$p = \left\{ \sum_{i=1}^{M} \beta_i \left\{ \sum_{j=0}^{m_i-1} \left[\frac{(m_i\mu_i T_r)^j}{j!} \right] e^{-m_i\mu_i T_r} \right\} \right\}^L \quad (39)$$

In our study, the input parameters λ_c, λ_f, T_r and the output measure $E[\tau_d]$ are normalized by the mean $1/\mu$ of the round-trip transmission delay. For purposes of demonstration, we consider t_r with a 2-Erlang distribution and $KL = 6$. The Echo message arrivals is a deterministic stream with fixed interval $T_e = 18/\mu$.

A. *Effects of input parameters on* α

Based on (9), Fig. 3(a) plots α against T_r and the (K, L) pair, where $\lambda_c = \mu/18$ and $\lambda_f = 1 \times 10^{-5}\mu$. It is trivial that α is a decreasing function of T_r. The non-trivial result is that Fig. 3(a) quantitatively indicates how the T_r value affects α. When $T_r < 2/\mu$, increases T_r significantly reduces α. On the other hand, when $T_r > 2/\mu$, increasing T_r does not improve the performance. Also, for small T_r, $L = 1$ outperforms other L setups. Same effect is observed for other λ_c values. When T_r is large, the L (and thus K) values have same impact on α.

Fig. 3(b) plots α as a function of T_r and λ_f, where $K = 6$, $L = 1$ and $\lambda_c = \mu/18$. This figure shows that α increases as λ_f decreases. When λ_f decreases (i.e., the system reliability improves but the transmission delay distribution remains the same as before), the GTP' connection lifetime becomes longer. Therefore, the opportunity for false failure detection increases. For $T_r = 1.6/\mu$, when the system reliability increases from $\lambda_f = 1 \times 10^{-5}\mu$ to $\lambda_f = 1 \times 10^{-6}\mu$, α increases by 2.72 times. This effect becomes insignificant when T_r is large (e.g., $T_r > 2.2/\mu$).

Fig. 4(a) plots α as a function of T_r and λ_c, where $K = 6$, $L = 1$ and $\lambda_f = 1 \times 10^{-5}\mu$. This figure shows that α increases as λ_c increases. When there are more GTP' message arrivals, it is more likely that false failure detection occurs. This effect is insignificant when T_r becomes large (e.g., $T_r > 2/\mu$).

B. *Effects of input parameters on* $E[\tau_d]$

Based on (37), Fig. 4(b) plots $E[\tau_d]$ as a function of T_r and λ_c, where $K = 6$, $L = 1$. This figure shows that $E[\tau_d]$ significantly increases as λ_c decreases.

Figs. 5(a) and 5(b) plot $E[\tau_d]$ as functions of T_r and the (K, L) pair, where $\lambda_c = \mu$ and $\lambda_c = \mu/36$, respectively. These figures show that $E[\tau_d]$ is an increasing function of T_r and $E[\tau_d]$ is more sensitive to the change of T_r when L is large than when L is small. When $\lambda_c = \mu$, $E[\tau_d]$ is larger for $L = 6$ than for $L = 1$. When $\lambda_c = \mu/36$, the opposite results

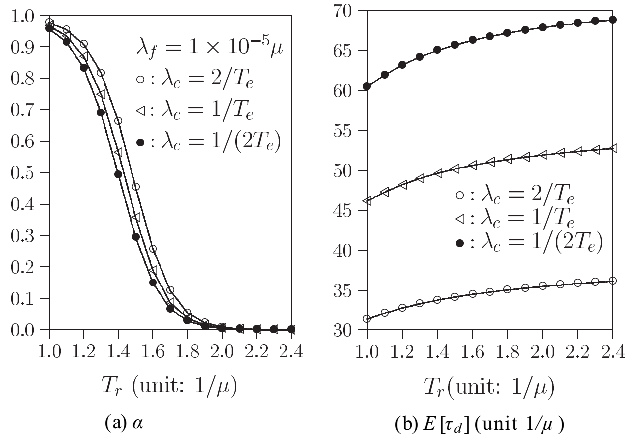


Fig. 4. Effects of T_r and λ_c $(K = 6, L = 1)$

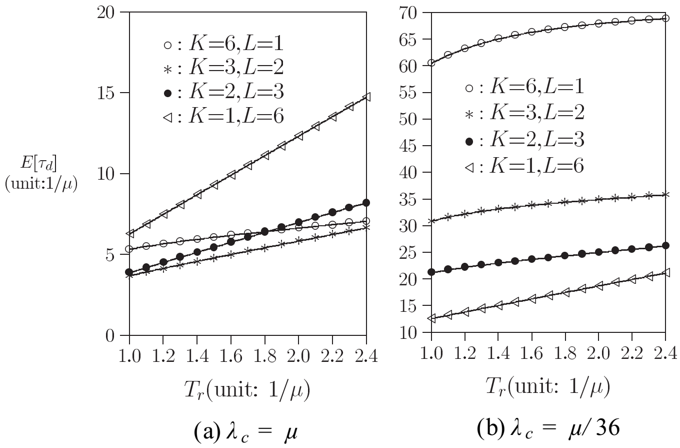


Fig. 5. Effects of T_r and L on $E[\tau_d]$

are observed. This phenomenon can be explained as follows. Without loss of generality, assume that $t_{a,1} \geq t_f$. Consider an extreme case that λ_c is very large, and many GTP' charging packets arrive in a very short period $(t', t'+dt')$ where $t' \geq t_f$. For $L = 1 (K = 6)$, $t_{a,6} \approx t'$ and $t_{d,6} \approx t' + T_r$. Therefore, the true failure detection time is $t_d \approx t' + T_r$. For $L = 6 (K = 1)$, we have $t_{a,1} \approx t'$, but the true failure detection time is $t_d = t_{d,1} \approx t' + 6T_r$. Therefore, $E[\tau_d]$ is larger for $L = 6$ than for $L = 1$ in Fig. 5(a).

On the other hand, when λ_c is small, the charging packets rarely occur in a short period, and it is likely that $t_{a,i+1} - t_{a,i} > T_r$ (for $i > 0$). For $L = 1$, the failure is detected at $t_{a,6} + T_r$. For $L = 6$, the failure is detected at $t_{a,1} + 6T_r$. Under the situation that $t_{a,i+1} - t_{a,i} > T_r$, we have $t_{a,6} - t_{a,1} > 5T_r$. Therefore, we expect that $E[\tau_d]$ is smaller for $L = 6$ than for $L = 1$ in Fig. 5(b).

VI. Conclusions

In UMTS, the GTP' protocol is used to deliver the CDRs from GSNs to CGs. To ensure that the mobile operator receives the charging information, availability for the charging system is essential. One of the most important issues on GTP' availability is connection failure detection. This paper studied the GTP' connection failure detection mechanism specified in 3GPP TS 29.060 and 3GPP TS 32.215. The output measures considered are the false failure detection probability α and the expected time $E[\tau_d]$ of true failure detection. We proposed an analytic model to investigate how these two output measures are affected by input parameters including the Charging Packet Ack Wait Time T_r, the Maximum Number L of Charging Packet Tries and the Maximum Number K of Unsuccessful Deliveries. We make the following observations.

- When T_r is small, increasing T_r reduces α significantly. When T_r is sufficiently large, increasing T_r only has insignificant impact on α. On the other hand, increasing T_r always non-negligibly increases $E[\tau_d]$.
- α increases as the charging packet arrival rate λ_c increases. This effect is insignificant when T_r becomes large. On the other hand, the effects of λ_c on $E[\tau_d]$ are not the same for different (K, L) setups. In our examples, when λ_c is large, $E[\tau_d]$ is larger for $L = 6$ than for $L = 1$. When λ_c is small, $E[\tau_d]$ is smaller for $L = 6$ than for $L = 1$. Therefore, the effects of λ_c should be considered when we select the L value.

In summary, the network operator can select the appropriate T_r, L and K values for various traffic conditions based on our study.

REFERENCES

[1] 3rd Generation Partnership Project, Technical Specification Group Services and Systems Aspects, "Architectural Requirements for Release" 1999 (Release 1999), 3G TS 23.121 version 3.6.0 (2002-06), 2002.

[2] 3rd Generation Partnership Project, Technical Specification Group Core Network, General Packet Radio Service (GPRS), "GPRS Tunneling Protocol (GTP) across the Gn and Gp Interface" (Release 5), 3G TS 29.060 version 5.9.0 (2004-03), 2004.

[3] 3rd Generation Partnership Project, Technical Specification Group Services and Systems Aspects, Telecommunication management, Charging management, "Charging data description for the Packet Switched (PS) domain" (Release 5), 3G TS 32.215 version 5.5.0 (2003-12), 2003.

[4] Y. Fang, and I. Chlamtac, "Teletraffic analysis and mobility modeling for PCS networks,"*IEEE Trans. Commun.*, vol. 47, no.7, pp. 1062-1072, July 1999.

[5] F. P. Kelly, *Reversibility And Stochastic Networks*. John Wiley & Sons, 1979.

[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. 493-501, May 2003.

[7] Y.-B. Lin and I. Chlamtac, *Wireless and Mobile Network Architectures*. JohnWiley & Sons, 2001.

[8] Y.-B. Lin, Y.-R. Haung, A.-C. Pang, and I. Chlamtac, "All-IP approach for UMTS third generation mobile networks," *IEEE Network*, vol. 16, no.5, pp. 8-19, Sept. 2002.

[9] R. G. Gallager, *Discrete Stochastic Processes*. Kluwer Academic Publishers, 1999.

[10] S. M. Ross, *A First Course in Probability*. Prentice Hall, 2001.

[11] S. M. Ross, *Stochastic processes*. JohnWiley & Sons, 1996.

[12] H.-N. Hung, Y.-B. Lin, N.-F. Peng, and S.-I. Sou, Connection Failure Detection Mechanism of UMTS Charging Protocol. Technical Report, 2004.

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'95-SM'95-F'03) received the B.S.E.E. degree from the National Cheng Kung University in 1983 and the Ph.D. degree in computer science from the University of Washington in 1990. He is chair professor in the Department of Computer Science and Information Engineering (CSIE), National Chiao Tung University (NCTU). Dr. Lin is a fellow of the IEEE and the ACM.

Nan-Fu Peng received the B.S. degree in the 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.

Sok-Ian Sou 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 working toward the Ph.D. degree at NCTU. Her current research interests include personal communications services network, Voice over IP technology and performance modeling.