國立交通大學

資訊工程系

碩士論文

UMTS計費協定之連線失敗偵測機制

Connection Failure Detection Mechanism of UMTS Charging Protocol

研究生：蘇淑茵

指導教授：林一平　教授

洪慧念　教授

中華民國九十三年六月

UMTS計費協定之連線失敗偵測機制

Connection Failure Detection Mechanism of UMTS Charging Protocol

研究生：蘇淑茵　　Student： Sok-Ian Sou

指導教授：林一平　　Advisor： Yi-Bing Lin

洪慧念　　Hui-Nien Hung

國立交通大學

資訊工程系

碩士論文

A Thesis

Submitted to Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science and Information Engineering

June 2004

Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

UMTS計費協定之連線失敗偵測機制

Student：蘇淑茵　　　　　　　　Advisors：　林一平教授

洪慧念教授

國立交通大學資訊工程系碩士班

中文摘要

在Universal Mobile Telecommunications System (UMTS)，GPRS Tunneling (GTP)協定的延伸稱為GTP’協定，負責把計費資料紀錄從GPRS服務節點傳送到計費閘道。為了確保行動營運商能收到計費資料，GTP’協定的傳輸可靠性(Reliability及Availability)是很重要的。而GTP’協定的性能評估之重要指標是連線失敗偵測。在本論文，我們研究在第三代通訊規格TS29.060及TS 32.215所提出的GTP’連線錯誤偵測機制。我們希望選取適當的參數值，避免偵測出錯誤的連線失敗(False Failure Detection；例如因暫時性的網路擁塞所引致)。同時，我們希望可以儘快地偵測出真正的連線失敗，當真正的連線失敗被偵測後，GPRS服務節點可立即導向到另一台計費閘道。我們提出一個分析模型去計算錯誤的連線失敗偵測機率以及偵測出真正的連線失敗所需之期望時間。這個分析模型所導出的分析結果已和模擬實驗所得到的數據互相驗證。依據我們的研究結果，行動營運商可針對不同的狀況去調整參數，以降低錯誤的連線失敗偵測和/或偵測真正連線失敗所需的時間。

Connection Failure Detection Mechanism of UMTS Charging Protocol

Student：Sok-Ian Sou

Advisors： Dr. Yi-Bing Lin
Dr. Hui-Nien Hung

Department of Computer Science and Information Engineering
National Chiao Tung University

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 charging system is essential. One of the most important issues on GTP' availability is connection failure detection. This paper studies the GTP' connection failure detection mechanism specified in 3GPP TS 29.060 and 3GPP TS 32.215. It is desirable to select appropriate parameter values to avoid false failure detections (e.g., temporary network congestions). It is also important to detect the true failures quickly, and after a true failure is detected, the GSNs can immediately re-direct to another CG. In this paper, we propose an analytic model to compute the false failure detection probability and the expected true failure detection time. The analytic model is validated against simulation experiments. 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.

Acknowledgment

I would especially like to thank my advisors, Prof. Yi-Bing Lin and Prof. Hui-Nien Hung. Without their supervision and perspicacious advices, I cannot complete this thesis. I have learned a lot from them. I would like to thank my committee members, Prof. Wei-Ru Lai and Dr. Yuan-Kai Chen for their valuable comments. I am very grateful to my colleagues in Laboratory 117 for the support I received while I was writing this thesis. Also, I like to express my thanks to all my dear friends for their friendship.

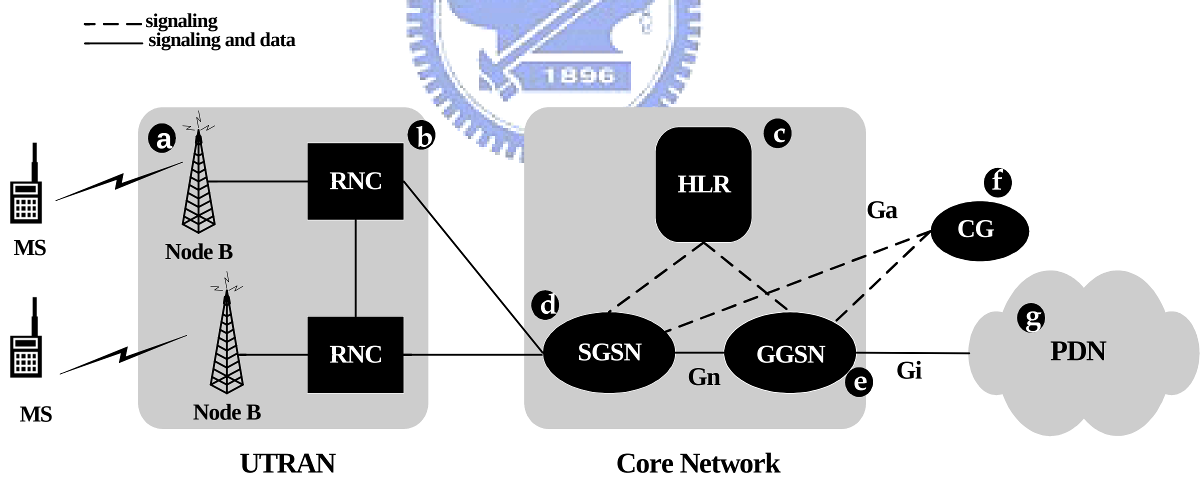
Lastly, I want to thank my dear parents, my dear sisters and my love, Yinman Lee for their unfailing love and firmly support in these years.

Contents

List of Figures

Chapter 1
Introduction

Universal Mobile Telecommunications System (UMTS) [2,11] supports high-speed *Packet Switched* (PS) data for accessing versatile multimedia services anytime and anywhere. Fig. 1 shows the architecture for the UMTS PS service domain [12]. In this figure, the dashed lines represent signaling links, and the solid lines represent data and signaling links. The PS *Core Network* is an Internet Protocol (IP)-based backbone network. This core network consists of *GPRS Support Nodes* (GSNs) such as *Serving GPRS Support Nodes* (SGSNs; see Fig. 1 (d)) and *Gateway GPRS Support Nodes* (GGSNs; see Fig. 1 (e)).



CG: Charging Gateway
GGSN: Gateway GPRS Support Node
HLR: Home Location Register
MS: Mobile Station
PDN: Packet Data Network

UTRAN: UMTS Terrestrial Radio Access Network
RNC: Radio Network Controller
SGSN: Serving GPRS Support Node
Node B: Base Station

Figure 1: The UMTS Network Architecture

A SGSN connecting to the *UMTS Terrestrial Radio Access Network* (UTRAN) plays a role in the PS service domain similar to a mobile switching center in the circuit switched service domain. The GGSN interworks to the external *Packet Data Network* (PDN; see Fig. 1 (g)).

The *Home Location Register* (HLR; see Fig. 1 (c)) communicates with the GSNs for mobility management and session management [11,12]. The UTRAN consists of *Node B*s (the UMTS term for base stations; see Fig. 1 (a)) and *Radio Network Controller*s (RNCs; see Fig. 1 (b)) connected by an ATM network. A *Mobile Station* (MS) communicates with one or more Node Bs through the radio interface *Uu* based on the *Wideband CDMA* (WCDMA) radio technology [8]. The *Charging Gateway* (CG; see Fig. 1 (f)) collects the billing and charging information from the GSNs.

Several IP-based interfaces are defined among the GSNs, CGs and the external PDN. In the Gn interface, the *GPRS Tunneling Protocol* (GTP) [3] transports user data and control signals among the GSNs. The GGSN connects to the PDN through the Gi interface. In the Ga interface, the GTP' protocol is utilized to transfer the *Charging Data Records* or *Call Detail Records* (CDRs) from GSNs to CGs. When an MS 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. Each GSN will only send the CDRs to the CG(s) in the same UMTS network. A CG analyzes and possibly consolidates the CDRs from various GSNs, and passes the consolidated data to a billing system.

For the purposes of this paper, GSN and CG merit further discussion. 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. The CDR database is a non-volatile storage. Data stored in this database are analyzed and consolidated before the CG sends them to the billing system. The CG is associated with a *Restart Counter* that records the number of restarts performed at the CG. Details of this counter will be elaborated in Section 2.1. For redundancy reasons, a CG may also maintain a configurable list of peer CG addresses (e.g., to be able to recommend other CGs to the GSNs).

A GSN maintains a list of CGs in the priority order (typically ranges from 1 to 100). This *CG* *list* can be configured by the *Operation and Management* (O&M) system. 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 to be elaborated in Sections 3 and 4. The entry includes pointers to buffers containing the unacknowledged CDR packets and the sequence numbers of possibly duplicated packets. The entry also stores the restart counter of the corresponding CG.

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 [3] 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, and an alarm is sent to the O&M system. For a GSN-CG link failure, the O&M system may cancel CDR packets in the CG and unacknowledged sequence numbers in the GSN.

This paper studies the availability issues for GTP'. Specifically we propose an analytic model to investigate the GTP' connection failure detection mechanism. This analytic model is validated against simulation experiments. Our study will provide guidelines for the mobile operators to select the parameters for GTP' connection manipulation.

Chapter 2

The GTP' Protocol

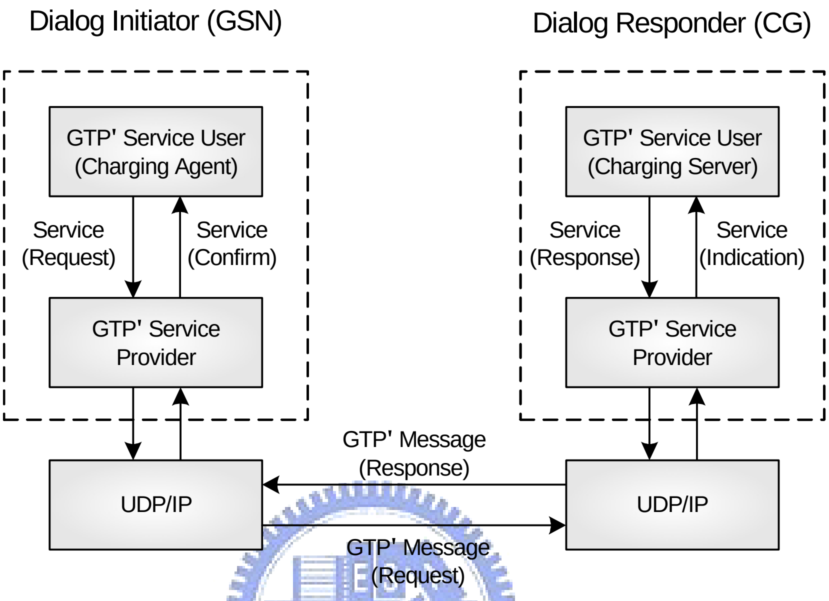


Figure 2: The GTP' Service Model

The GTP' protocol is used for communications between a GSN and a CG, which can be implemented over UDP/IP or TCP/IP. GTP' utilizes some aspects of GTP defined in 3GPP TS 29.060 [3]. Specifically, GTP control plane (GTP-C) is partly reused. Fig. 2 illustrates a GTP' service model for a WLAN and GPRS integration system developed in National Chiao Tung University (NCTU) [7].

In our design, the GTP' protocol is built on top of UDP/IP. Above the GTP' protocol, a *Charging Agent* (or *CDR sender*) is implemented in the GSN and a *Charging Server* is implemented in the CG. Our GTP' service model follows the GSM Mobile Application Part (MAP) service model (see Chapter 10 in [11]). In this model, a GSN communicates with a CG through a *dialog* by invoking GTP' *service primitives*. A service primitive can be one of four types: Request (REQ), Indication (IND), Response (RSP) and Confirm (CNF). A service primitive is initiated by a *GTP' service user* of the *dialog initiator*. In Fig.2, the dialog initiator is a GSN and the service user is a charging agent. The charging agent issues a service primitive with type REQ. This service request is sent to the *GTP’ service provider* of the GSN. The service provider sends the request to the *dialog responder* (the CG in Fig. 2) by creating a GTP’ message. This GTP’ message is delivered through lower layer protocols; i.e., UDP/IP. When the GTP’ service provider of the CG receives the request, it invokes the same service primitive with type IND to the charging server (GTP’ service user). The charging server then performs appropriate operations, and invokes the same service primitive with type RSP. This response primitive is a service acknowledgement sent from the CG to the GSN. After the GTP’ service provider of the GSN receives this response, it invokes the same service primitive with type CNF. The parameters of the CNF and the RSP primitives are identical in most cases except that the CNF primitive may include an extra provider error parameter to indicate a protocol error.

If a dialog is initiated by the CG, then the roles of the CG and the GSN are exchanged in Fig. 2. Based on the above GTP’ service model, this section describes the GTP’ message format, the GTP’ connection setup procedure and the CDR transfer procedure.

2.1 GTP’ Message Format

As defined in 3GPP TS 32.215 [5], the GTP’ header may follow the standard 20-octet GTP header format (Fig. 3 (a)) [1] or a simplified 6-octet format (Fig. 3 (b)). The 6-octet GTP’ header is the same as the first 6 octets of the standard GTP header. Octets 7-20 of the GTP header are used to specify data session between a GSN and the MS. These octets are not needed in GTP’. In Fig. 3, the first bit of octet 1 is used to indicate the header format. If the value is 1, the 6-octet header is used. If the value is 0, the 20-octet standard GTP header is used. Note that better GTP’ performance is expected by using the 6-octet format, because the un-used GTP header fields are eliminated. On the other hand, it is easier to support GTP’ in an existing GTP environment if the standard GTP header format is used. In Fig. 3, the *Protocol Type* (PT) and the *Version* fields are used to specify the protocol being used (GTP or GTP’ in R99, R4, R5 and so on). For a GTP’ message, PT=0. The *Length* field indicates the length of payload. The *Sequence Number* is used as the transaction identity.

	Bits							
Octets	8	7	6	5	4	3	2	1
1	Version			PT	Spare ‘ 1 1 1 ‘			‘0/1’
2	Message Type							
3-4	Length							
5-6	Sequence Number							
7-8	Flow Label							
9	SNDCP N-PDULLC Number							
10	Spare ‘ 1 1 1 1 1 1 1 1 ‘							
11	Spare ‘ 1 1 1 1 1 1 1 1 ‘							
12	Spare ‘ 1 1 1 1 1 1 1 1 ‘							
13-20	TID							

(a) GTP header (Version 0)

	Bits							
Octets	8	7	6	5	4	3	2	1
1	Version			PT	Spare ‘ 1 1 1 ‘			‘0/1’
2	Message Type							
3-4	Length							
5-6	Sequence Number							

(b) 6-octet header

Figure 3: GTP’ Header Formats

Message Type value	GTP’ message
1	Echo Request
2	Echo Response
3	Version Not Supported
4	Node Alive Request
5	Node Alive Response
6	Redirection Request
7	Redirection Response
240	Data Record Transfer Request
241	Data Record Transfer Response

Figure 4: GTP’ Message Types

The GTP’ *Message Type*s are listed in Fig. 4. Three GTP message types are reused in GTP’, including Echo Request, Echo Response and Version Not Supported. The Echo Request/Response message pair is typically used to check if the peer is alive. These path management messages are required if GTP’ is supported by UDP. Specifically, the Echo Request is sent by a GSN to find out if the peer CG is alive. In 3GPP TS 29.060 [3], the Echo Request is periodically sent for more than every 60 seconds on each connection. Whenever a CG receives an Echo Request, it replies with an Echo Response that contains the value of its local restart counter. As we mentioned in the previous section, this counter is maintained in both the GSN and the CG to indicate the number of restarts performed at the CG. If the restart counter value received by the GSN is larger than the value previously stored, the GSN assumes that the CG has restarted since the last Echo Request/Response message pair exchange. In this case, the GSN may retransmit the earlier unacknowledged packets to the CG rather than wait for expiries of their timers.

The Node Alive Request/Response message pair is used to inform that a CG has restarted its service after a service break. The service break may be caused by, e.g., hardware maintenance. When a CG’s service is stopped due to, e.g., outage for maintenance, the CG sends a Redirection Request message to inform a GSN to redirect its CDRs to another CG. This message can also be used to balance the workloads among the CGs.

The Data Record Transfer Request/Response message pair is used for CDR delivery. In a Data Record Transfer Request message, the header is followed by two *Information Elements* (IEs). The first IE is a code indicating “`Send Data Record Packet`”. The next IE consists of one or more CDRs. In a Data Record Transfer Response message, the header is followed by a cause IE. This IE is a code that indicates how a CDR is processed in the CG (e.g., `Request Accepted`, `No Resource Available`, and so on).

2.2 GTP’ Connection Setup Procedure

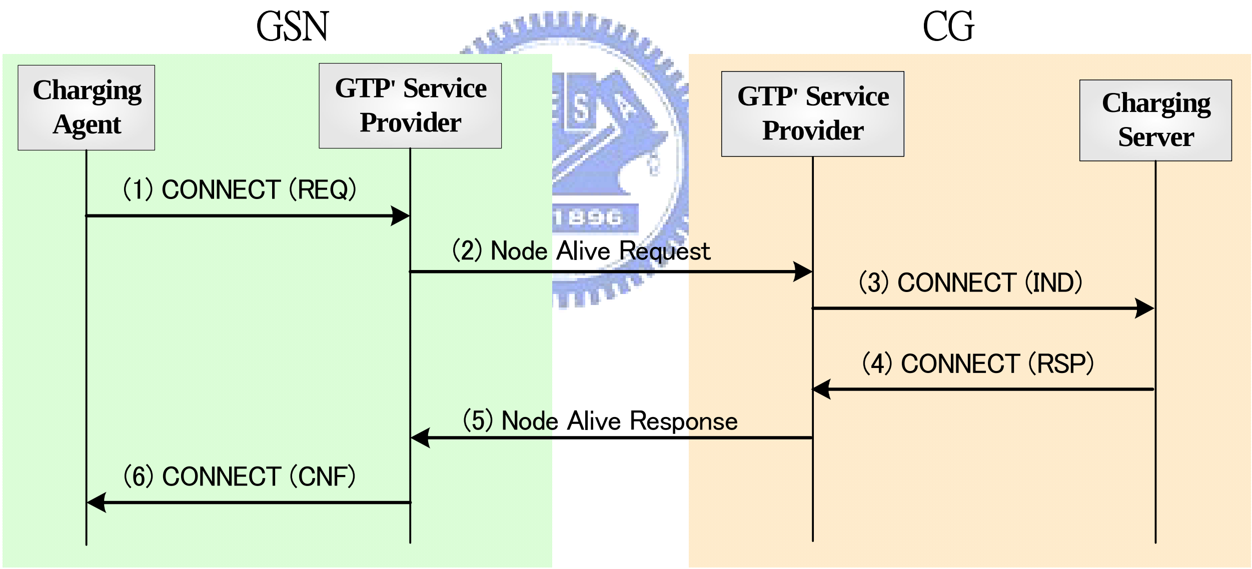


Figure 5: GTP’ Connection Setup Message Flow

Before a GSN can send CDRs to a CG, a GTP’ connection must be established between the charging agent in the GSN and the charging server in the CG. The GTP’ connection setup procedure is described in the following steps (see Fig. 5):

Step 1. The charging agent instructs the GTP’ service provider to set up a GTP’ connection. This task is performed by issuing the CONNECT (REQ) primitive with the CG address.

Step 2. The service provider generates a Node Alive Request message and delivers it to the CG through UDP/IP. The UDP source port number is locally allocated at the GSN. On the CG side, the default UDP destination port number is 3386 reserved for GTP’ [5]. Alternatively, the CG may configure this destination port number.

Step 3. The GTP’ service provider of the CG interprets the Node Alive Request message and reports this connection setup event to the charging server via the CONNECT (IND) primitive.

Step 4. The charging server creates and sets a new entry (for this new connection) in the GSN list, and responds to the service provider with the CONNECT (RSP) primitive. Either the charging server is ready to receive the CDRs or it is not available for this connection. In the latter case, the charging server may include the address of a recommended CG in the CONNECT (RSP) primitive for further redirection request.

Step 5. Suppose that the CG is available. The GTP’ service provider generates a Node Alive Response message, and delivers this message to the GSN.

Step 6. The GTP’ service provider of the GSN receives the Node Alive Response message. It interprets the message and reports this acknowledgement event to the charging agent through the CONNECT (CNF) primitive. The charging agent creates and sets the CG entry’s status as active in the CG list. At this point, the setup procedure is complete.

2.3 GTP' CDR Transfer Procedure

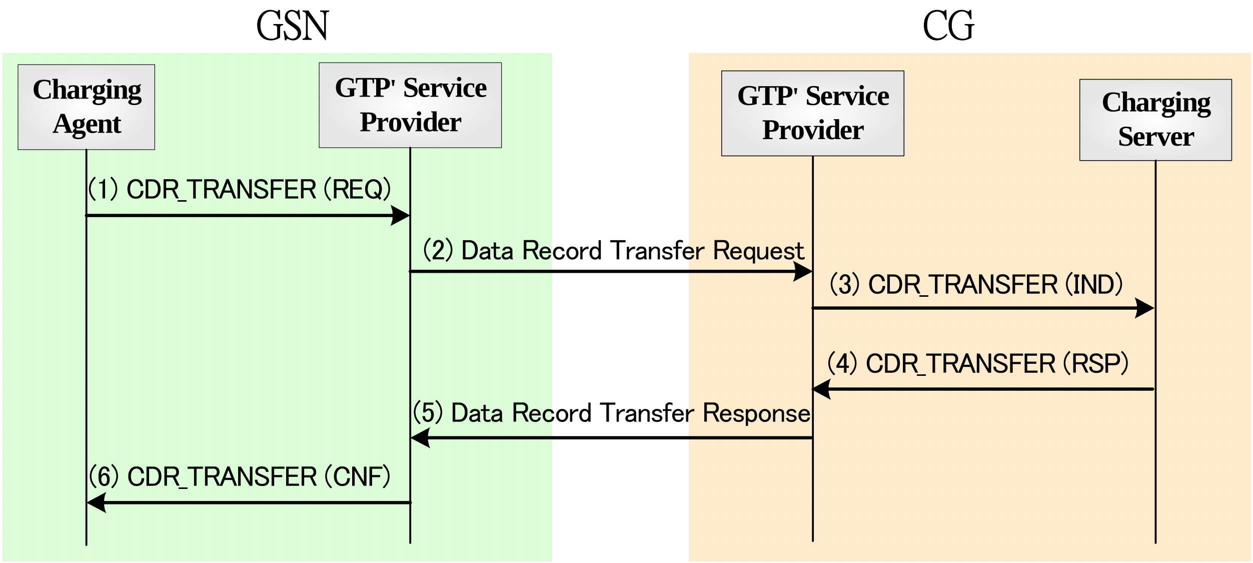


Figure 6: GTP' CDR Transfer Message Flow

The charging agent is responsible for CDR generation in a GSN. The CDRs are encoded using, for example, the ASN.1 format defined in [5]. The charging server is responsible for decoding the CDRs and returns the processing results to the GSN. The CDR transfer procedure is illustrated in Fig. 6 and is described in the following steps:

Step 1. The charging agent encodes the released CDR. Then it invokes the CDR_TRANSFER (REQ) primitive. This primitive instructs the GTP' service provider to generate a Data Record Transfer Request message.

Step 2. The service provider includes the CDR in the Data Record Transfer Request message and sends it to the CG.

Step 3. When the service provider of the CG receives the GTP' message, it issues the CDR_TRANSFER (IND) primitive to inform the charging server that a CDR is received. The charging server decodes the CDR and stores it in the CDR database. This CDR may be consolidated with other CDRs, and is later sent to the billing system.

Steps 4 and 5. The charging server invokes the CDR_TRANSFER (RSP) primitive that requests the GTP' service provider to generate a Data Record Transfer Response message. The cause IE value of the message is "`Request Accepted`". The service provider sends this GTP' message to the GSN.

Step 6. The GTP' service provider of the GSN receives the Data Record Transfer Response message and reports this acknowledgement event to the charging agent via the CDR_TRANSFER (CNF) primitive. The charging agent deletes the delivered CDR from its unacknowledged buffer.

2.4 GTP' Failure Detection

This subsection describes the *Path Failure Detection Algorithm* (PFDA) that detects path failure between the GSN and the CG. Fig. 7 illustrates the data structures utilized to implement PFDA.

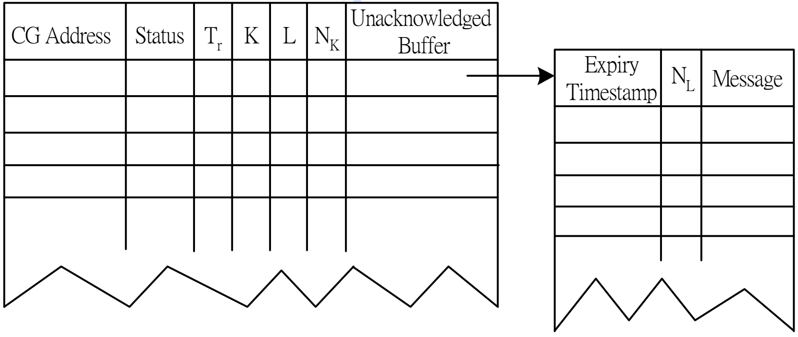


Figure 7: Data Structures for Path Failure Detection Algorithm

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 in Section 2.2 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 (Step 2, Section 2.3), 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 (Step 6, Section 2.3), 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 3, and compute the expected detection time of true failure in Section 4.

Chapter 3
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, and 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

$$T_e \geq LT_r \tag{1}$$

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

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

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

$$N_e(t_f) = \lfloor t_f / T_e \rfloor \tag{3}$$

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 (2) and (3),

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

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

$$p = \left(\Pr[t_r \geq T_r]\right)^L = [1 - F_r(T_r)]^L \tag{5}$$

The GTP' connection is considered disconnected after *K* consecutive delivery timeouts where each of the delivery 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 (with probability $1 - \sum_{i=0}^{j-K-1} \theta(i)$), the $(j$-$K)$-th delivery is a success (with probability 1-*p*), and the last *K* deliveries are timed out (with probability p^K). 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

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

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 $\overline{\theta}(j)$ be the probability that no false failure is detected before (and including) the j-th GTP’ message delivery. Then

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

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

$$\begin{aligned}
\alpha &= 1 - \int_{t_f=0}^{\infty} \sum_{n=0}^{\infty} \overline{\theta}\left(\lfloor t_f / T_e \rfloor + n \right) \Pr\left[N(t_f) = \lfloor t_f / T_e \rfloor + n \right] f_f(t_f) dt_f \\
&= 1 - \int_{t_f=0}^{\infty} \sum_{n=0}^{\infty} \overline{\theta}\left(\lfloor t_f / T_e \rfloor + n \right) \left[\frac{(\lambda_c t_f)^n}{n!} \right] e^{-\lambda_c t_f} f_f(t_f) dt_f \\
&= 1 - \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \overline{\theta}(k+n) \int_{t_f=kT_e}^{(k+1)T_e} \left[\frac{(\lambda_c t_f)^n}{n!} \right] e^{-\lambda_c t_f} f_f(t_f) dt_f \qquad (8)
\end{aligned}$$

The derivation for (8) can be extended by assuming that the lifetime t_f has an exponential distribution with mean $1/\lambda_f$. The exponential distribution is chosen because it has often been used in reliability and lifetime modeling [14]. 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

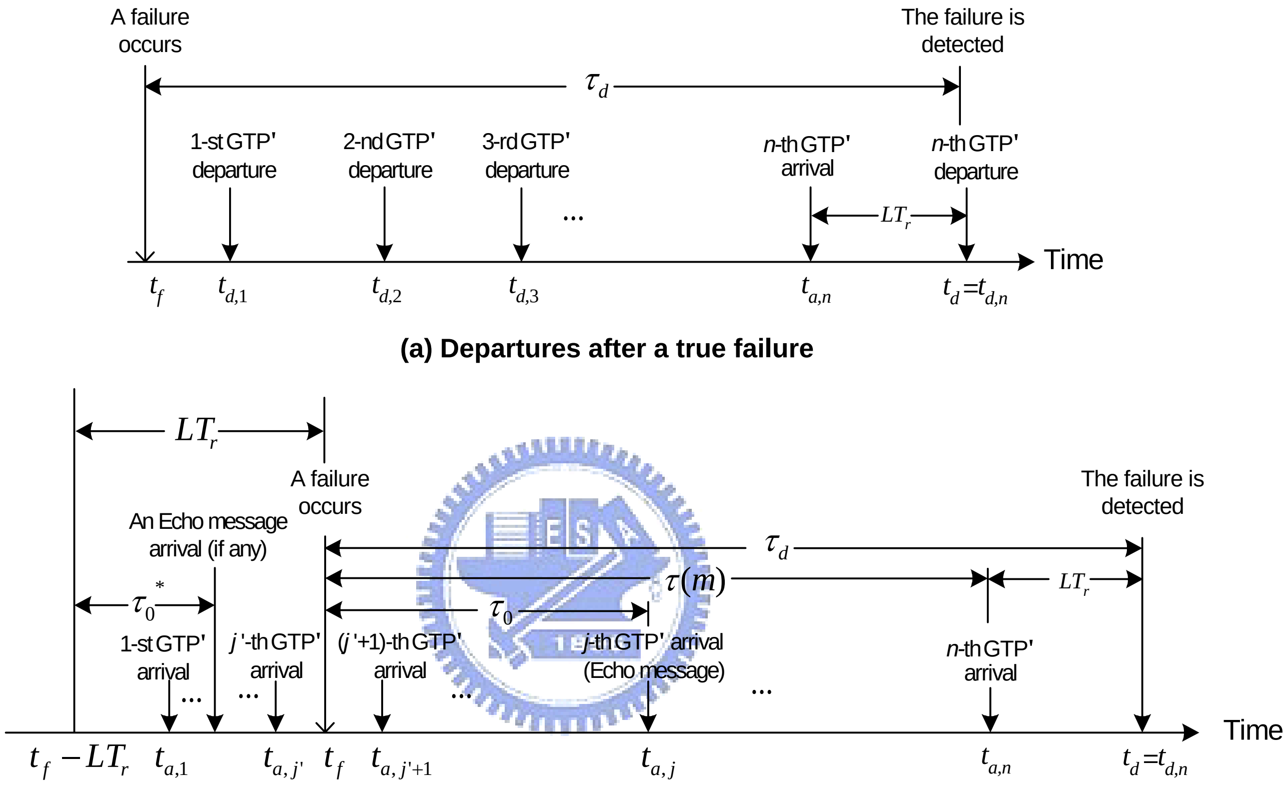
$$\begin{aligned}
\alpha &= 1 - \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \overline{\theta}(k+n) \int_{t_f=kT_e}^{(k+1)T_e} \left[\frac{\lambda_f (\lambda_c t_f)^n}{n!} \right] e^{-(\lambda_c + \lambda_f) t_f} dt_f \\
&= 1 - \lambda_f \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \overline{\theta}(k+n) \left(\frac{\lambda_c{}^n}{n!} \right) \int_{t_f=kT_e}^{(k+1)T_e} t_f{}^n e^{-(\lambda_c + \lambda_f) t_f} dt_f
\end{aligned}$$

$$\begin{aligned}
&= 1 - \lambda_f \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \overline{\theta}(k+n) \left(\frac{\lambda_c{}^n}{n!} \right) \left\{ \left[\frac{n!}{(\lambda_c + \lambda_f)^{n+1}} \right] \left\{ 1 - \sum_{j=0}^{n} \frac{e^{-(\lambda_c+\lambda_f)(k+1)T_e} [(\lambda_c + \lambda_f)(k+1)T_e]^j}{j!} \right\} \right. \\
&\qquad \left. - \left[\frac{n!}{(\lambda_c + \lambda_f)^{n+1}} \right] \left\{ 1 - \sum_{j=0}^{n} \frac{e^{-(\lambda_c+\lambda_f)kT_e} [(\lambda_c + \lambda_f)kT_e]^j}{j!} \right\} \right\}
\end{aligned}$$

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

Chapter 4

Expected True Failure Detection Time



(a) Departures after a true failure

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

Figure 8: Timing Diagram for Detecting True Failure ($n \le K$)

This section proposes an analytic model to derive the expected detection time of "*true*" failure. Consider the timing diagram in Fig. 8 (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. 8 (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. 8 (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. 8 (b), the first *j'* deliveries arrive before t_f. If

$$t_{a,n} > t_f \tag{10}$$

then the true failure detection time τ_d is

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

In this section, we compute the probability that $N_K(t_f) = K\text{-}n$ (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 Appendix C.

4.1 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. 9). 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

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

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

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

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

$$\Pr[N_K(t_f)=j] = \begin{cases} \dfrac{E[X]}{T_e - p^{K-1}(T_e - E[X])} & , j=0 \\[2ex] \dfrac{(1-p)p^{j-1}(T_e - E[X])}{T_e - p^{K-1}(T_e - E[X])} & , 0<j<K \end{cases} \tag{14}$$

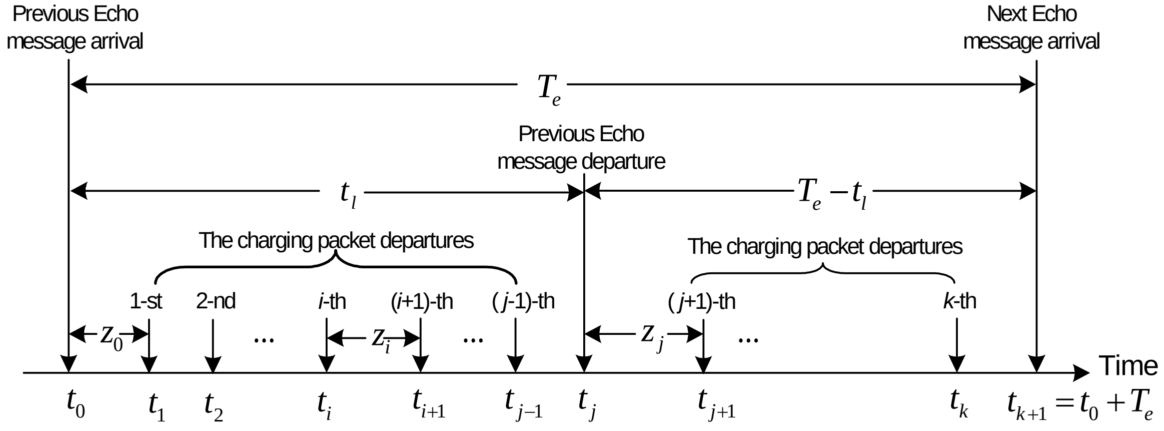


Figure 9: Timing Diagram for Deriving *E*[*X*]

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. 9, 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

$$E[X] = E[X \mid t_l = LT_r]\Pr[t_l = LT_r] + E[X \mid t_l < LT_r]\Pr[t_l < LT_r] \tag{15}$$

$E[X \mid t_l = LT_r]$ 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 $E[X \mid t_l = LT_r]$ if $N_{K\to\infty}(t_i)=0$. For $j \le k$, we have

$$E[X \mid t_l = LT_r] = \Pr[N_{K\to\infty}(t_0) = 0]E[z_0] + (1-p)E\left[\sum_{i=1}^{j-1} z_i\right] + (1-p)E\left[\sum_{i=j+1}^{k} z_i\right] \tag{16}$$

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

$$\begin{aligned} E[X \mid t_l = LT_r] &= \Pr[N_{K\to\infty}(t_0) = 0]E[z_0] + (1-p)\big(LT_r - E[z_0]\big) + (1-p)\big(T_e - LT_r - E[z_j]\big) \\ &= (1-p)T_e + (\Pr[N_{K\to\infty}(t_0) = 0] + p - 1)E[z_0] - (1-p)E[z_j] \end{aligned} \tag{17}$$

In (17), $\Pr[N_{K\to\infty}(t_0)=0]$ is derived in Appendix A. $E[z_0]$ is derived as follows. If the first charging packet departure occurs before t_0+LT_r , then z_0 is exponentially distributed under the condition that $z_0 < LT_r$. That is

$$\begin{aligned} E[z_0 \mid z_0 < LT_r]\Pr[z_0 < LT_r] &= \int_{z_0=0}^{LT_r} z_0 \lambda_c e^{-\lambda_c z_0} dz_0 \\ &= \left(\frac{1}{\lambda_c}\right)\left(1 - e^{-\lambda_c LT_r}\right) - LT_r e^{-\lambda_c LT_r} \end{aligned} \tag{18}$$

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

$$\begin{aligned} E[z_0 \mid z_0 = LT_r]\Pr[z_0 = LT_r] &= \int_{t=LT_r}^{\infty} LT_r \lambda_c e^{-\lambda_c t} dt \\ &= LT_r e^{-\lambda_c LT_r} \end{aligned} \tag{19}$$

Combining (18) and (19) to yield

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

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

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

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

$$E[X \mid t_l = LT_r] = (1-p)T_e + \left(\frac{\Pr[N_{K\to\infty}(t_0) = 0] + p - 1}{\lambda_c}\right)\left(1 - e^{-\lambda_c LT_r}\right) - \left(\frac{1-p}{\lambda_c}\right)\left[1 - e^{-\lambda_c (T_e - LT_r)}\right]$$

(22)

$E[X \mid t_l < LT_r]$ 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 $t_l < LT_r$,

$$E[X \mid t_l] = \Pr[N_{K\to\infty}(t_0) = 0]E[z_0(t_l)] + (1-p)E\left[\sum_{i=1}^{j-1} z_i(t_l)\right] + E[z_j(t_l)] + (1-p)E\left[\sum_{i=j+1}^{k} z_i(t_l)\right]$$

(23)

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

$$E[X \mid t_l] = (1-p)T_e + \left(\Pr[N_{K\leftarrow\infty}(t_0) = 0] + p - 1\right)E[z_0(t_l)] + pE[z_j(t_l)]$$

$$= (1-p)T_e + \left(\frac{\Pr[N_{K\to\infty}(t_0) = 0] + p - 1}{\lambda_c}\right)\left(1 - e^{-\lambda_c t_l}\right) + \left(\frac{p}{\lambda_c}\right)\left[1 - e^{-\lambda_c (T_e - t_l)}\right]$$

$$= (1-p)T_e + \left(\frac{\Pr[N_{K\to\infty}(t_0) = 0] + 2p - 1}{\lambda_c}\right) - \left(\frac{\Pr[N_{K\to\infty}(t_0) = 0] + p - 1}{\lambda_c}\right)e^{-\lambda_c t_l}$$

$$-\left(\frac{pe^{-\lambda_c T_e}}{\lambda_c}\right)e^{\lambda_c t_l} \quad (24)$$

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,

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

$$= (1-p)^2 T_e + \frac{(1-p)\left(\Pr[N_{K\to\infty}(t_0)=0]+2p-1\right)}{\lambda_c}$$

$$-\left(\frac{pe^{-\lambda_c T_e}}{\lambda_c}\right)\int_{t_l=0}^{LT_r} e^{\lambda_c t_l} f_l(t_l)dt_l$$

$$-\left(\frac{\Pr[N_{K\to\infty}(t_0)=0]+p-1}{\lambda_c}\right)\int_{t_l=0}^{LT_r} e^{-\lambda_c t_l} f_l(t_l)dt_l \qquad (25)$$

From (15), (22), (25) and (43) derived in Appendix B, $E[X]$ is expressed as

$$E[X] = pE[X \mid t_l = LT_r] + (1-p)^2 T_e + \frac{(1-p)\left(\Pr[N_{K\to\infty}(t_0)=0]+2p-1\right)}{\lambda_c}$$

$$-\left(\frac{pe^{-\lambda_c T_e}}{\lambda_c}\right)\int_{t_l=0}^{LT_r} e^{\lambda_c t_l} f_l(t_l)dt_l$$

$$-\left(\frac{\Pr[N_{K\to\infty}(t_0)=0]+p-1}{\lambda_c}\right)\int_{t_l=0}^{LT_r} e^{-\lambda_c t_l} f_l(t_l)dt_l$$

$$= pE[X \mid t_l = LT_r] + (1-p)^2 T_e + \frac{(1-p)\left(\Pr[N_{K\to\infty}(t_0)=0]+2p-1\right)}{\lambda_c}$$

$$-\left(\frac{pe^{-\lambda_c T_e}}{\lambda_c}\right)\int_{t_l=0}^{LT_r} e^{\lambda_c t_l}[1-F_r(T_r)]^{\lfloor t_l/T_r \rfloor} f_r(t_l - \lfloor t_l/T_r \rfloor T_r)dt_l$$

$$-\left(\frac{\Pr[N_{K\to\infty}(t_0)=0]+p-1}{\lambda_c}\right)\int_{t_l=0}^{LT_r} e^{-\lambda_c t_l}[1-F_r(T_r)]^{\lfloor t_l/T_r \rfloor} f_r(t_l - \lfloor t_l/T_r \rfloor T_r)dt_l \qquad (26)$$

Finally, $\Pr[N_K(t_f) = j]$ can be computed by using (14) and (26).

4.2 Derivation for $E[\tau_d]$

For $t_{a,n} > t_f$ and *m*>0, let *m* 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. 8 (b); $j' = n_c + n_e$). It can be observed that

$$N_K(t_f+LT_r) = \min\left\{N_K(t_f) + n_c + n_e,\ K\right\} \tag{27}$$

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 $E[\tau_d \mid m=0]$ is derived in Appendix C. 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 [13]), where

$$\rho = \lambda_c \int_{t_l=0}^{LT_r} \left[1 - F_L(t_l)\right] dt_l \tag{28}$$

and the probability mass function of n_c is given by

$$\Pr[n_c = i] = \left(\frac{\rho^i}{i!}\right) e^{-\rho} \tag{29}$$

In Fig. 8 (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 $\Pr[n_e = 1 \mid \tau_0]$ be the probability that n_e=1 for a specific τ_0. Then $\Pr[n_e = 1 \mid \tau_0]$ can be expressed as

$$\Pr[n_e=1\,|\,\tau_0]=\begin{cases}0 & ,\ \tau_0\le T_e-LT_r\\ 1\text{-}\ F_L(T_e-\tau_0) & ,\ \tau_0>T_e-LT_r\end{cases}\qquad(30)$$

where $F_L(t)$ is derived in Appendix B.

In (30), 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 (29) and (30), $\Pr[n_c+n_e$=$j'\,|\,\tau_0]$ can be expressed as

$$\Pr[n_c+n_e=j'\,|\,\tau_0]=\begin{cases}\Pr[n_c=0]\big(1-\Pr[n_e=1\,|\,\tau_0]\big) & ,j'\text{=}0\\ \Pr[n_c=j'-1]\Pr[n_e=1\,|\,\tau_0]+\Pr[n_c=j']\big(1-\Pr[n_e=1\,|\,\tau_0]\big) & ,j'>0\end{cases}$$

$$=\begin{cases}e^{-\rho}\big(1-\Pr[n_e=1\,|\,\tau_0]\big) & ,j'\text{=}0\\ e^{-\rho}\left\{\left[\dfrac{\rho^{j'-1}}{(j'-1)!}\right]\Pr[n_e=1\,|\,\tau_0]+\left(\dfrac{\rho^{j'}}{j'!}\right)\big(1-\Pr[n_e=1\,|\,\tau_0]\big)\right\} & ,j'>0\end{cases}\qquad(31)$$

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 (31) as

$$\Pr[N_K(t_f+LT_r)=j\,|\,\tau_0]=\sum_{i=0}^{j}\Pr[N_K(t_f)=i]\Pr[n_c+n_e=j-i\,|\,\tau_0]\qquad(32)$$

For m>0, let $\tau(m)=t_{a,n}-t_f$ (see Fig. 8(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

$$m_e=\lfloor(\tau(m)-\tau_0)/T_e\rfloor+1\qquad(33)$$

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, τ_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{τ_c, τ_e}. Note that m_e is determined by $\tau(m)$ and τ_0 (see (33)), 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

$$F_c(\tau_c) = 1 - \sum_{i=0}^{m_c-1}\left[\frac{(\lambda_c\tau_c)^i}{i!}\right]e^{-\lambda_c\tau_c} \qquad (34)$$

For m>0, let $F_m(\tau(m))$ be the distribution function of $\tau(m)$. From (33) and (34), we have

$$\begin{aligned} F_m(\tau(m) \mid \tau_0) &= F_c(\tau(m) \mid \tau_0) \\ &= 1 - \sum_{i=0}^{m-\lfloor(\tau(m)-\tau_0)/T_e\rfloor-2}\left\{\frac{[\lambda_c\tau(m)]^i}{i!}\right\}e^{-\lambda_c\tau(m)} \qquad (35) \end{aligned}$$

Note that $F_m(\tau(m) \mid \tau_0)$ is discontinuous at points $\tau(m) = \tau_0 + jT_e$, for j=0, 1, ..., m_e-1. From (35) we have

$$\Pr[\tau(m) = \tau_0 + jT_e \mid \tau_0] = \Pr[\tau(m) \le \tau_0 + jT_e \mid \tau_0] - \Pr[\tau(m) < \tau_0 + jT_e \mid \tau_0]$$

$$= F_m(\tau_0 + jT_e \mid \tau_0) - F_m(\tau_0 + jT_e^- \mid \tau_0)$$

$$= \left\{1 - \sum_{i=0}^{m-j-2}\left\{\frac{[\lambda_c(\tau_0 + jT_e)]^i}{i!}\right\}e^{-\lambda_c(\tau_0+jT_e)}\right\} - \left\{1 - \sum_{i=0}^{m-j-1}\left\{\frac{[\lambda_c(\tau_0 + jT_e)]^i}{i!}\right\}e^{-\lambda_c(\tau_0+jT_e)}\right\}$$

$$= \left\{\frac{[\lambda_c(\tau_0 + jT_e)]^{m-j-1}}{(m-j-1)!}\right\}e^{-\lambda_c(\tau_0+jT_e)} \qquad (36)$$

Eq. (36) 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

$$E[\tau(m) \mid \tau_0] = \int_{\tau(m)=0}^{\infty}[1 - F_m(\tau(m) \mid \tau_0)]\ d\tau(m)$$

$$= \int_{\tau(m)=0}^{\infty}\sum_{i=0}^{m-\lfloor(\tau(m)-\tau_0)/T_e\rfloor-2}\left\{\frac{[\lambda_c\tau(m)]^i}{i!}\right\}\ e^{-\lambda_c\tau(m)}d\tau(m)$$

$$= \sum_{i=0}^{m-1} \int_{\tau(m)=0}^{\tau_0+(m-i-1)T_e} \left\{ \frac{[\lambda_c \tau(m)]^i}{i!} \right\} e^{-\lambda_c \tau(m)} d\tau(m)$$

$$= \sum_{i=0}^{m-1} \left(\frac{\lambda_c{}^i}{i!} \right) \int_{\tau(m)=0}^{\tau_0+(m-i-1)T_e} [\tau(m)]^i e^{-\lambda_c \tau(m)} d\tau(m)$$

$$= \left(\frac{1}{\lambda_c} \right) \sum_{i=0}^{m-1} \left\{ 1 - e^{-\lambda_c[\tau_0+(m-i-1)T_e]} \sum_{j=0}^{i} \frac{\{\lambda_c[\tau_0+(m-i-1)T_e]\}^j}{j!} \right\} \qquad (37)$$

Since t_f is a random observer of the inter-Echo arrival times, τ_0 is uniformly distributed over (0, T_e]. From (11), (32) and (37), the expected value of $E[\tau_d]$ is expressed as

$$\begin{aligned} E[\tau_d] &= E[\tau_d \mid m > 0]\Pr[m > 0] + E[\tau_d \mid m = 0]\Pr[m = 0] \\ &= \sum_{m=1}^{K} \left(E[\tau(m)] + LT_r \right) \Pr[N_K(t_f + LT_r) = K - m] + E[\tau_d \mid m = 0]\Pr[m = 0] \\ &= \sum_{m=1}^{K} \left[\left(\frac{1}{T_e} \right) \int_{\tau_0=0}^{T_e} \left(E[\tau(m) \mid \tau_0] + LT_r \right) \Pr[N_K(t_f + LT_r) = K - m \mid \tau_0]\ d\tau_0 \right] \\ &\quad + E[\tau_d \mid m = 0]\Pr[m = 0] \end{aligned} \qquad (38)$$

where $E[\tau_d \mid m = 0]$ and $\Pr[m = 0]$ are derived in Appendix C.

The analytic model developed in this paper is validated against the simulation experiments. The discrepancies between analytic analysis (specifically, Eqs. (9) and (38)) and simulation are within 3% in most cases. The simulation technique used in this paper is similar to the one described in [10], and the details are omitted.

Chapter 5
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

$$F_r(t_r)=1-\sum_{i=1}^{M}\beta_i\left\{\sum_{j=0}^{m_i-1}\left[\frac{(m_i\mu_i t_r)^j}{j!}\right]e^{-m_i\mu_i t_r}\right\} \tag{39}$$

where *M*, m_1, m_2, …, m_M are nonnegative integers, μ_i> 0, β_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 [6,9]. From (5) and (39)

$$p=\left\{\sum_{i=1}^{M}\beta_i\left\{\sum_{j=0}^{m_i-1}\left[\frac{(m_i\mu_i T_r)^j}{j!}\right]e^{-m_i\mu_i T_r}\right\}\right\}^L \tag{40}$$

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=$18/\mu$.

5.1 Effects of input parameters on α

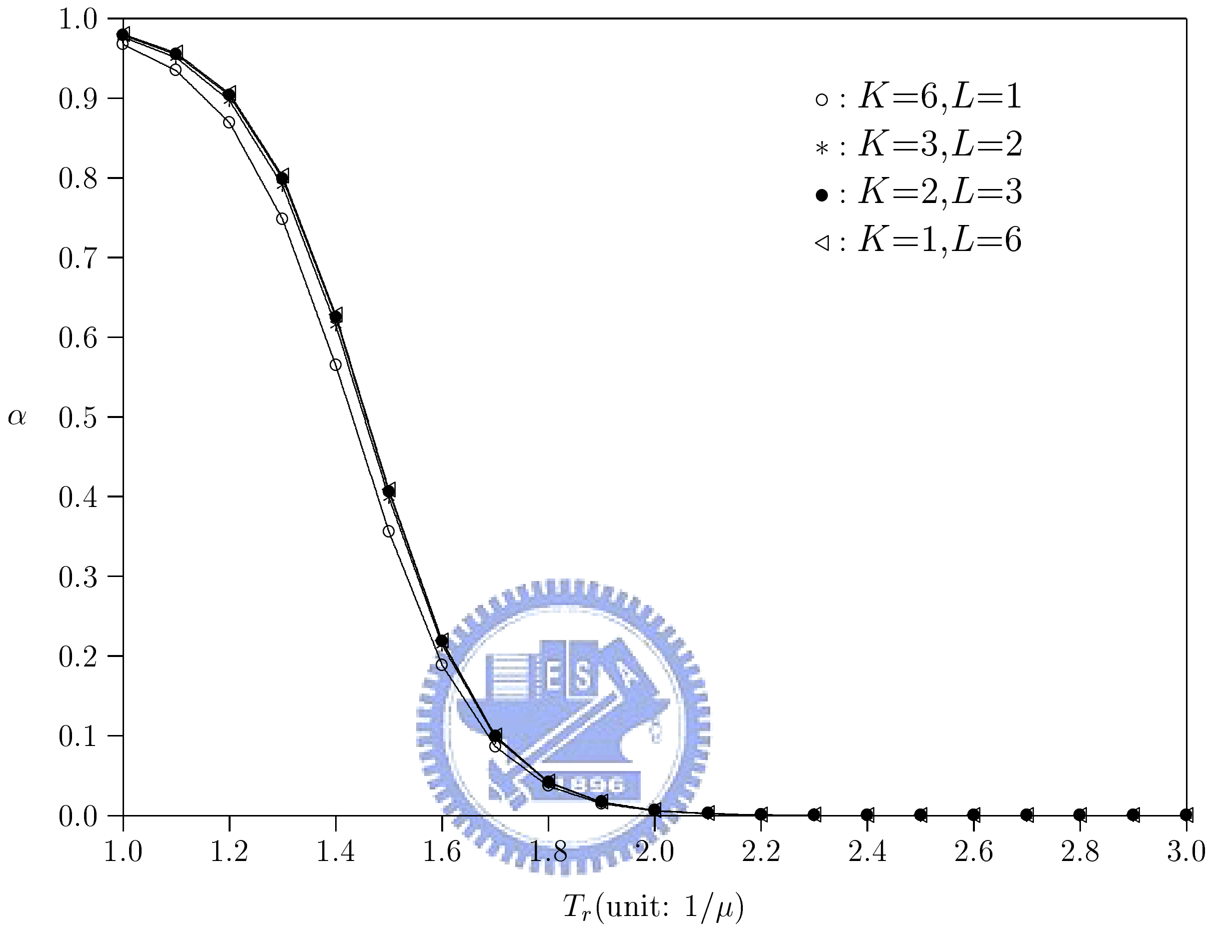


Figure 10: Effects of T_r and L on α **(**$\lambda_c = \mu/18$**,** $\lambda_f = 1\times10^{-5}\mu$**)**

Based on (9), Fig. 10 plots α against T_r and the (K, L) pair, where $\lambda_c = \mu/18$ and $\lambda_f = 1\times10^{-5}\mu$. It is trivial that α is a decreasing function of T_r. The non-trivial result is that Fig. 10 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.

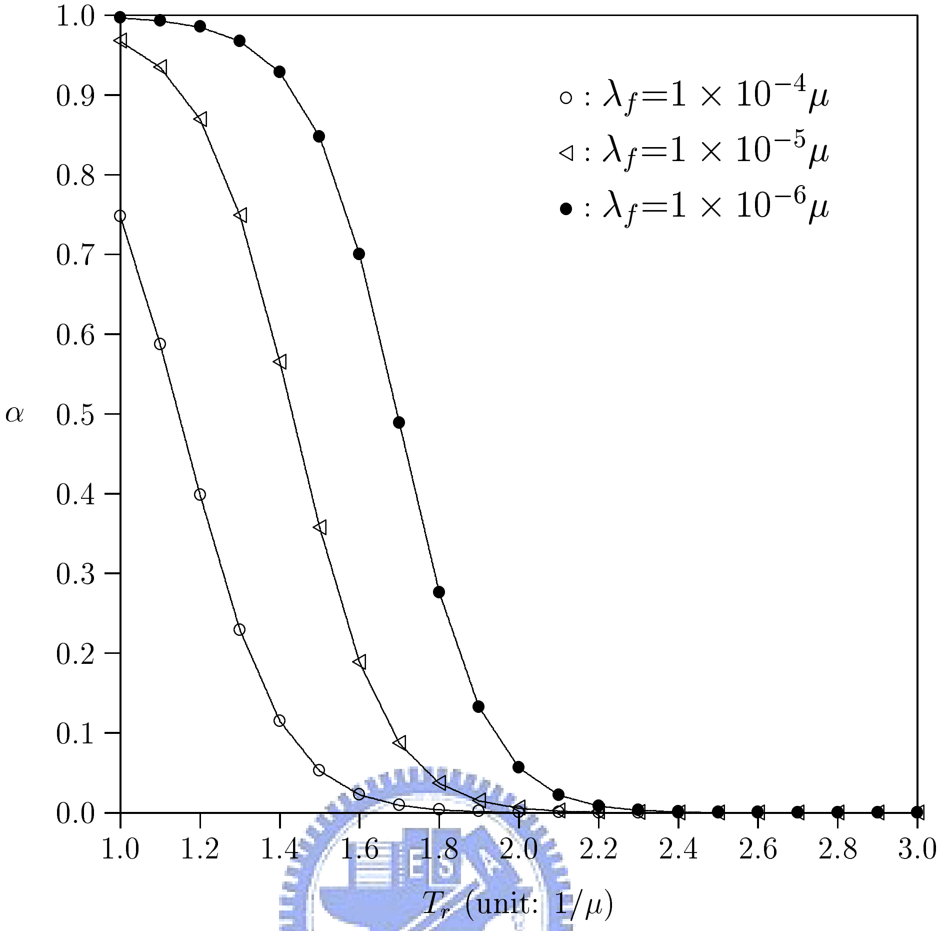


Figure 11: Effects of T_r and λ_f on α (*K*=6, *L*=1, $\lambda_c = \mu/18$)

Fig. 11 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 $\lambda_f = 1\times10^{-5}\mu$ to $\lambda_f = 1\times10^{-6}\mu$, α increases by 2.72 times. This effect becomes insignificant when T_r is large (e.g., $T_r > 2.2/\mu$).

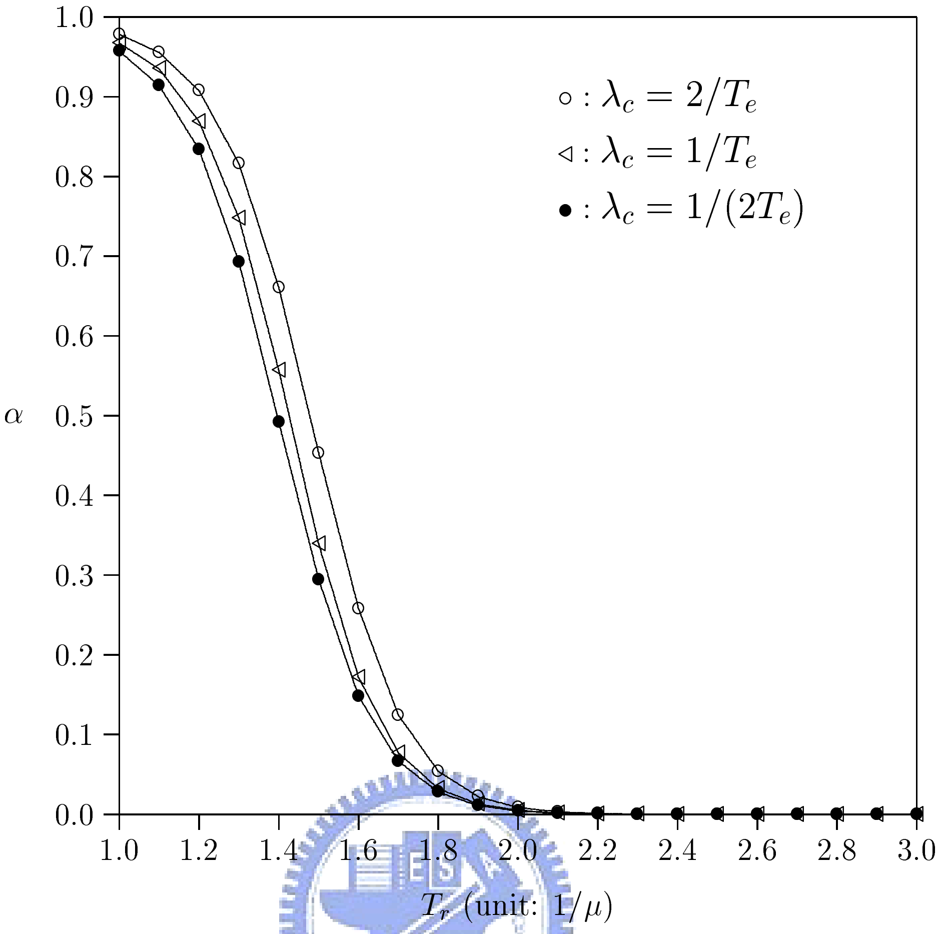


Figure 12: Effects of T_r and λ_c on α (K=6, L=1, $\lambda_f = 1\times10^{-5}\mu$)

Fig. 12 plots α as a function of T_r and λ_c, where K=6, L=1 and $\lambda_f = 1\times10^{-5}\mu$. 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., $T_r > 2/\mu$).

5.2 Effects of input parameters on $E[\tau_d]$

Based on (38), Fig. 13 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.

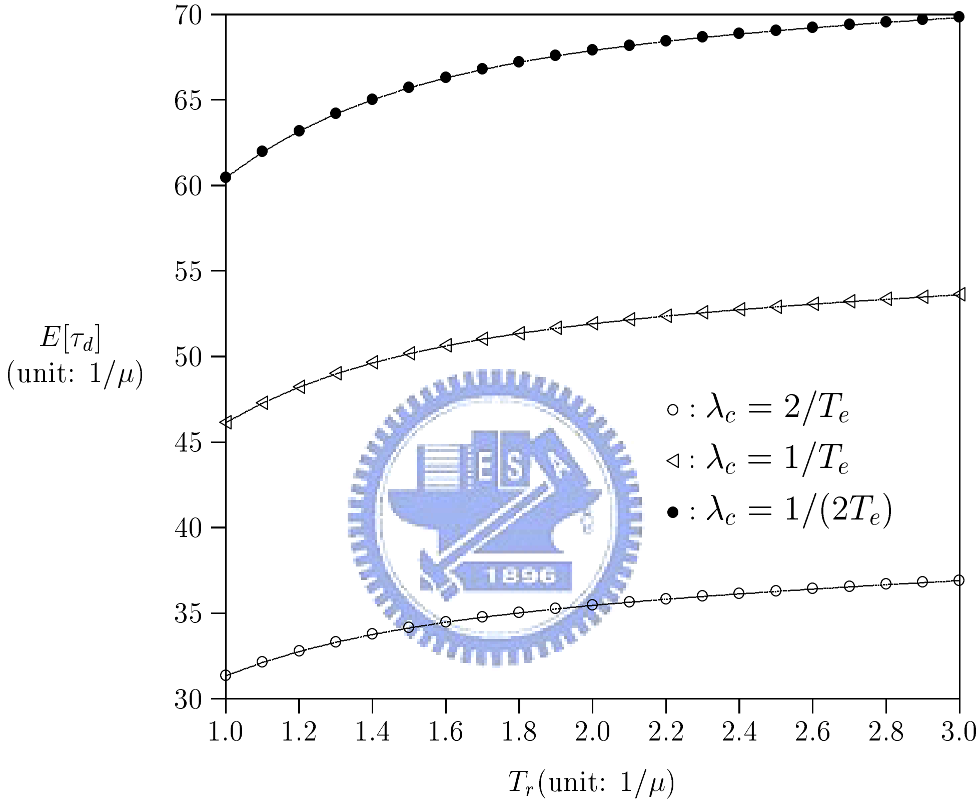
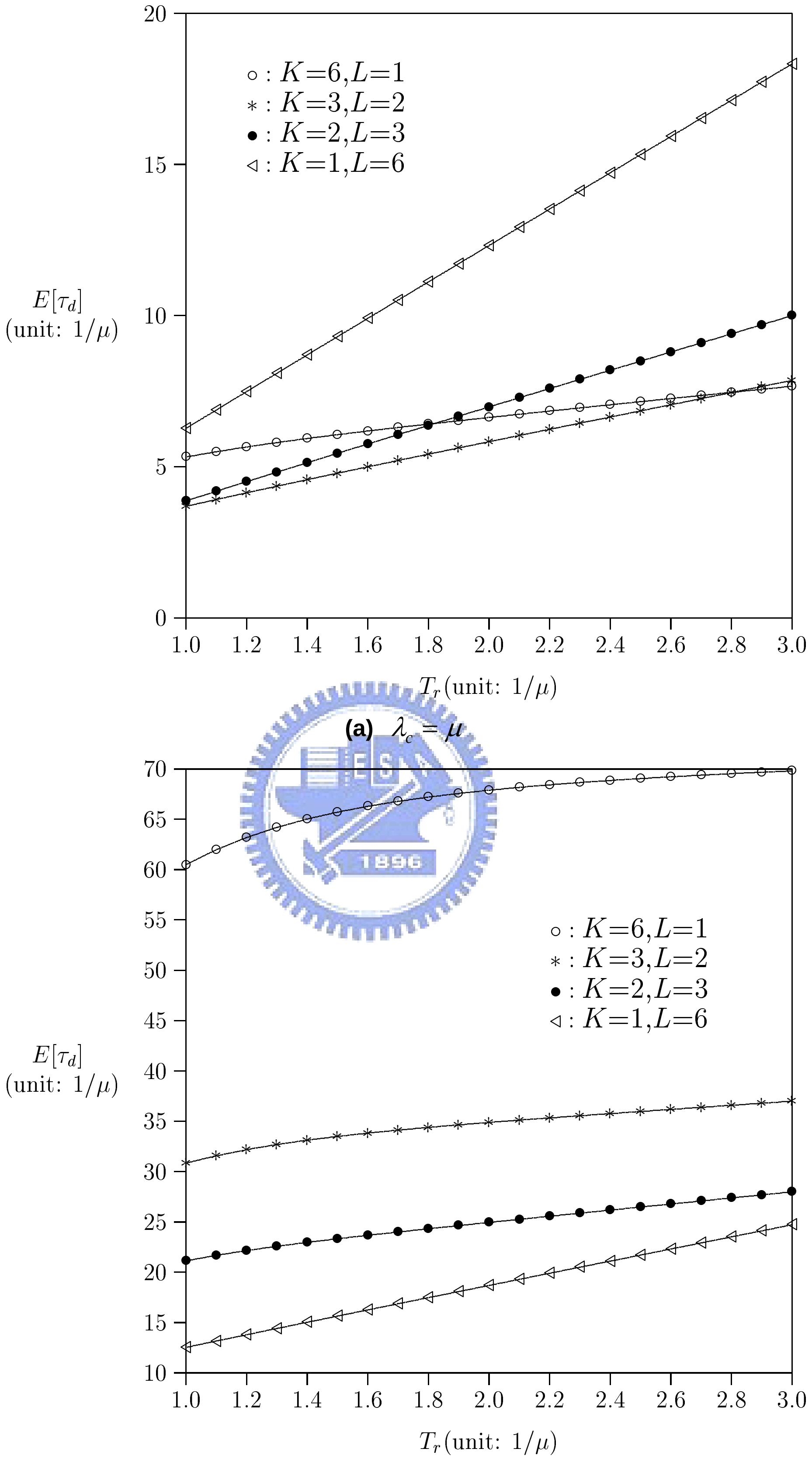


Figure 13: Effects of T_r and λ_c on $E[\tau_d]$ (K=6, L=1)



(a) $\lambda_c = \mu$

(b) $\lambda_c = \mu/36$

Figure 14 : Effects of T_r and L on $E[\tau_d]$

Based on (38), Figs. 14 (a) and (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 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. 14 (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. 14 (b).

Chapter 6
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. The analytic model was validated against simulation experiments. We make the following observations.

- When T_r is small, increasing T_r degrades 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.

Appendix A

Derivation for Pr[$N_{K\to\infty}(t_0)$=0]

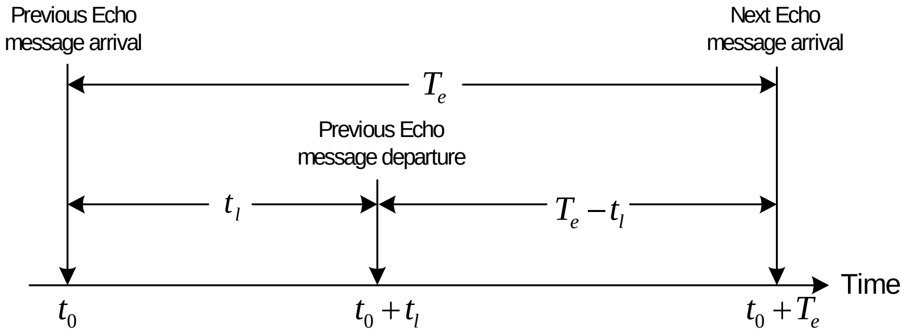


Figure 15: Timing Diagram for Deriving Pr[$N_{K\to\infty}(t_0)$=0]

This appendix derives Pr[$N_{K\to\infty}$(t_0)=0]. Fig. 15 shows the timing diagram between two consecutive arrivals of Echo messages at t_0 and t_0+T_e, respectively. We observe that $N_{K\to\infty}(t)$ is determined by the charging packet departures in [t_0 , t_0+T_e) and the initial value $N_{K\to\infty}$(t_0). The charging packet deliveries can be modeled by the M/G/∞ system, where the charging packet arrivals are a Poisson process with rate λ_c. The M/G/∞ model implies that the charging packet departures are also a Poisson process with the same rate λ_c. From the renewal property [15], the arrivals of Echo messages at fixed intervals can be treated as renewal points. Therefore in the steady state, the probabilities that $N_{K\to\infty}$(t_0)=0 and $N_{K\to\infty}$(t_0+T_e)=0 are identical. In Fig. 15, the delivery delay (includes retries) for the previous Echo message is t_l where $0 < t_l \le LT_r \le T_e$. In terms of t_l, Pr[$N_{K\to\infty}$(t_0)=0] can be derived in two cases:

Case I ($t_l = LT_r$ **).** The previous Echo message delivery is timed out (with probability p). In this case, $N_{K\to\infty}$(t_0+T_e)=0 if there are charging packet departures in [t_0+t_l , t_0+T_e) and the last one is a successful delivery (with probability $\left[1-e^{-\lambda_c(T_e-LT_r)}\right](1-p)$).

Case II ($0 < t_l < LT_r$ **).** The previous Echo message is successfully delivered (with probability $f_l(t_l)dt_l$). In this case, $N_{K\to\infty}$(t_0+T_e)=0 if there is no charging packet departure in period

[t_0+t_l , t_0+T_e) (with probability $e^{-\lambda_c(T_e-t_l)}$) or the last charging packet departure occurs in this period is a successful delivery (with probability $\left[1-e^{-\lambda_c(T_e-t_l)}\right](1-p)$).

From both Cases I and II, Pr[$N_{K\to\infty}$ (t_0)=0] is computed as

$$\begin{aligned}\Pr[N_{K\to\infty}(t_0)=0] &= p\left[1-e^{-\lambda_c(T_e-LT_r)}\right](1-p)+\int_{t_l=0}^{LT_r}\left\{e^{-\lambda_c(T_e-t_l)}+\left[1-e^{-\lambda_c(T_e-t_l)}\right](1-p)\right\}f_l(t_l)dt_l \\ &=(1-p)\left[1-pe^{-\lambda_c(T_e-LT_r)}\right]+pe^{-\lambda_cT_e}\int_{t_l=0}^{LT_r}e^{\lambda_ct_l}\ f_l(t_l)dt_l \qquad (41)\end{aligned}$$

From (43) derived in Appendix B, (41) can be expressed as

$$\Pr[N_{K\to\infty}(t_0)=0]=(1-p)\left[1-pe^{-\lambda_c(T_e-LT_r)}\right]+pe^{-\lambda_cT_e}\int_{t_l=0}^{LT_r}e^{\lambda_ct_l}\ [1-F_r(T_r)]^{\lfloor t_l/T_r\rfloor}f_r(t_l-\lfloor t_l/T_r\rfloor\ T_r)dt_l$$

(42)

Appendix B
Derivations for $f_l\,(t_l)$ and $F_l\,(t_l)$

This appendix derives the density function $f_l(t_l)$ and the distribution $F_L(t_l)$ of delivery delay t_l (including the first attempt and the subsequent retries) for a GTP’ message (i.e., a charging packet or an Echo message). If a message is successfully delivered, then $t_l < LT_r$, and $j = \lfloor t_l/T_r \rfloor$ is the number of re-transmissions (excluding the first attempt). In this case, the GSN awaits a period T_r for the i-th transmission with probability $1-F_r(T_r)$ (where $i \le j$), and the response time for the $(j+1)$-th transmission is $t_l - jT_r$ with probability $f_r(t_l - jT_r)dt_l$. Therefore, we have

$$f_l(t_l)dt_l = [1-F_r(T_r)]^j f_r(t_l - jT_r)dt_l \qquad \text{where } t_l < LT_r \text{ and } j = \lfloor t_l/T_r \rfloor \tag{43}$$

If the GTP’ message delivery fails, then $t_l = LT_r$. In this case, the GSN awaits a period T_r for each of the L transmissions (with probability $[1-F_r(T_r)]^L$), and the delay for the delivery is LT_r. Therefore,

$$\Pr[t_l = LT_r] = [1-F_r(T_r)]^L = p \tag{44}$$

From (43), the distribution function $F_L(t_l)$ for $0 \le t_l < LT_r$ is

$$\begin{aligned} F_L(t_l) &= 1-\Pr[t > t_l] \\ &= 1-\int_{t=t_l}^{\infty} [1-F_r(T_r)]^j f_r(t-jT_r)dt \quad \text{where } j = \lfloor t_l/T_r \rfloor \\ &= 1-[1-F_r(T_r)]^j \int_{\tau=t_l-jT_r}^{\infty} f_r(\tau)d\tau \\ &= 1-[1-F_r(T_r)]^{\lfloor t_l/T_r \rfloor}\left[1-F_r\left(t_l - \lfloor t_l/T_r \rfloor T_r\right)\right] \end{aligned} \tag{45}$$

Note that $F_L(LT_r^-) = 1-p$ and $F_L(LT_r) = 1$.

Appendix C

Derivation for $E[\tau_d|m=0]$ and Pr[m=0]

This appendix derives $E[\tau_d \mid m=0]$ and Pr[m =0]. Note that m =0 implies that $t_{a,n} \le t_f$ and $t_{d,n} > t_f$. Since $t_{a,n} = t_{d,n} - LT_r$, we have

$$t_f - LT_r < t_{a,n} \le t_f \tag{46}$$

As defined in Section 4.2, we denote a GTP' message as a cross message if it arrives before t_f and is timed out after t_f. Suppose that there are n_c cross charging packets and n_e cross Echo messages. Consider an arbitrary cross charging packet arriving at (t_f - LT_r)+x and departing at t_f+x , respectively, where $0 < x \le LT_r$. The density function $f_X(x)$ of x is derived as follows. In Fig. 16, an arbitrary GTP' charging packet arrives at (t_f - LT_r)+x . Since the true failure occurs at t_f, this packet delivery successes (i.e., the charging packet departs in period (t_f - LT_r+x , t_f]) with probability $F_L(LT_r - x)$ and fails (therefore becomes a cross charging packet) with probability $1 - F_L(LT_r - x)$.

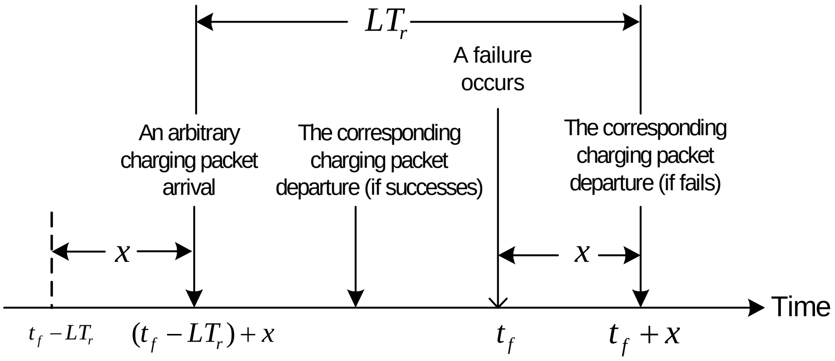

LT_r
A failure occurs
An arbitrary charging packet arrival
The corresponding charging packet departure (if successes)
The corresponding charging packet departure (if fails)
x
x
Time
$t_f - LT_r$
$(t_f - LT_r) + x$
t_f
$t_f + x$

Figure 16: Timing Diagram for Deriving $f_X(x)$

For an arbitrary cross packet, the packet must arrive in period (t_f - LT_r, t_f] and fail. Therefore, $f_X(x)$ can be expressed as

$$f_X(x) = \frac{1 - F_L(LT_r - x)}{\int_{x=0}^{LT_r} \left[1 - F_L(LT_r - x)\right] dx} = \frac{1 - F_L(LT_r - x)}{\int_{\tau=0}^{LT_r} \left[1 - F_L(\tau)\right] d\tau} \tag{47}$$

Suppose that for $1 \le i \le n_c$, the *i*-th cross charging packet arrives at time (t_f - LT_r)+ X_i. From the definition of order statistics [14], X_i has the density function

$$f_{X_i}(x_i) = \left[\frac{n_c!}{(i-1)!(n_c-i)!}\right] F_X(x_i)^{i-1} f_X(x_i)[1-F_X(x_i)]^{n_c-i} \tag{48}$$

where $0 < x_i \le LT_r$ and $F_X(x_i) = \int_{x=0}^{x_i} f_X(x)dx$.

For $1 < i \le n_c$ and $0 < x_{i-1} \le x_i \le LT_r$, let $\Pr[X_{i-1} < x_{i-1} < X_{i-1} + dx_{i-1},\ X_i < x_i < X_i + dx_i]$ $= f_{X_{i-1},X_i}(x_{i-1},x_i)dx_{i-1}dx_i$; that is, $f_{X_{i-1},X_i}(x_{i-1},x_i)$ is the joint density function for X_{i-1} and X_i. Then

$$f_{X_{i-1},X_i}(x_{i-1},x_i) = \left[\frac{n_c!}{(i-2)!(n_c-i)!}\right] F_X(x_{i-1})^{i-2} f_X(x_{i-1}) f_X(x_i)[1-F_X(x_i)]^{n_c-i} \tag{49}$$

Section 4.2 points out that n_e is either 0 or 1 and the first Echo message after t_f arrives at $t_f + \tau_0$. Therefore, the previous Echo message (i.e., the latest Echo message before t_f) arrives at $t_f + \tau_0 - T_e$. When n_e =1, $(t_f - LT_r) + \tau_0^{\ *} = t_f + \tau_0 - T_e$ is the arrival time of the previous cross Echo message. That is,

$$\tau_0^{\ *} = \tau_0 - T_e + LT_r \tag{50}$$

As mentioned in Section 4.2, $N_K(t_f) = K - n$. Let $E_{\tau_0,n,n_c,n_e}[\tau_d \mid m=0]$ be $E[\tau_d \mid m=0]$ for specific τ_0, *n*, n_c and n_e values. Under the condition that *m*=0, we have $n_c + n_e \ge n$. By considering whether there is a cross Echo message, we have

$$E_{\tau_0,n}[\tau_d \mid m=0] = \sum_{j=n}^{\infty} E_{\tau_0,n,n_c=j,0}[\tau_d \mid m=0]\big(1-\Pr[n_e=1 \mid \tau_0]\big)\Pr[n_c=j]$$

(51a)

$$+\sum_{j=n-1}^{\infty} E_{\tau_0,n,n_c=j,1}[\tau_d \mid m=0]\Pr[n_e=1 \mid \tau_0]\Pr[n_c=j] \qquad (51$$

b)

where $\Pr[n_e = 1 \mid \tau_0]$ and $\Pr[n_c = j]$ are obtained from (29) and (30), respectively.

In (51a), $E_{\tau_0,n,n_c,0}[\tau_d \mid m=0]$ is derived as follows. For n_e =0, the failure is detected at the departure time of the n-th cross charging packet. That is, $t_{d,n} = t_f + X_n$ and $\tau_d = X_n$. From (48), we have

$$E_{\tau_0,n,n_c,0}[\tau_d \mid m=0] = \int_{x=0}^{LT_r} x \; f_{X_n}(x)dx \tag{52}$$

In (51b), $E_{\tau_0,n,n_c,1}[\tau_d \mid m=0]$ is derived as follows. For n_e =1, there are n_c +1 cross messages. Based on the value of n_c, the following cases are considered:

Case I. For n_c =0. There is one cross message. It is clear that

$$E_{\tau_0,1,0,1}[\tau_d \mid m=0] = \tau_0^{\;*} \tag{53}$$

Case II. For n_c > 0, there are three possibilities:

Case II (a). For n=1, the failure is detected at the departure time of the first cross message, which can be the first cross charging packet (if $X_1 < \tau_0^{\;*}$; see (54a)) or the cross Echo message (if $X_1 \geq \tau_0^{\;*}$; see (54b)). From (48), we have

$$E_{\tau_0,1,n_c>0,1}[\tau_d \mid m=0\,]$$

$$= E_{\tau_0,1,n_c>0,1}[\tau_d \mid m=0 \;\text{ and }\; X_1 < \tau_0^{\;*}]\Pr[X_1 < \tau_0^{\;*}]$$

(54a)

$$+ E_{\tau_0,1,n_c>0,1}[\tau_d \mid m=0 \;\text{ and }\; X_1 \geq \tau_0^{\;*}]\Pr[X_1 \geq \tau_0^{\;*}]$$

(54b)

$$= \int_{x_1=0}^{\tau_0^{\;*}} x_1 \; f_{X_1}(x_1)dx_1 + \tau_0^{\;*} \int_{x_1=\tau_0^{\;*}}^{LT_r} f_{X_1}(x_1)dx_1 \tag{54}$$

Case II (b). For $n = n_c + 1$, the failure is detected at the departure time of the (n_c+1)-th cross message, which can be the cross Echo message (if $X_{n_c} < \tau_0^{\;*}$; see (55a)) or the n_c -th cross charging packet (if $X_{n_c} \geq \tau_0^{\;*}$; see (55b)). From (48), we have

$$E_{\tau_0,n=n_c+1,n_c>0,1}[\tau_d \mid m=0]$$

$$=E_{\tau_0,n=n_c+1,n_c>0,1}[\tau_d \mid m=0 \ \text{ and } \ X_{n_c}<\tau_0{}^*]\Pr[X_{n_c}<\tau_0{}^*]$$

(55a

)

$$+E_{\tau_0,n=n_c+1,n_c>0,1}[\tau_d \mid m=0 \ \text{ and } \ X_{n_c}\ge\tau_0{}^*]\Pr[X_{n_c}\ge\tau_0{}^*]$$

(55

b)

$$=\tau_0{}^*\int_{x_{n_c}=0}^{\tau_0{}^*} f_{X_{n_c}}(x_{n_c})dx_{n_c}+\int_{x_{n_c}=\tau_0{}^*}^{LT_r} x_{n_c}\ f_{X_{n_c}}(x_{n_c})dx_{n_c} \qquad (55)$$

Case II (c). For $1<n\le n_c$, the failure is detected at the departure time of the *n*-th cross message, which can be the *n*-th cross charging packet (if $X_n\le\tau_0{}^*$; see (56a)), the cross Echo message (if $X_{n-1}<\tau_0{}^*<X_n$; see (56b)) or the (*n*-1)-th cross charging packet (if $\tau_0{}^*\le X_{n-1}$; see (56c)). From (48) and (49), we have

$$E_{\tau_0,1<n\le n_c,n_c>0,1}[\tau_d \mid m=0]$$

$$=\ E_{\tau_0,1<n\le n_c,n_c>0,1}[\tau_d \mid m=0 \ \text{ and } \ X_n\le\tau_0{}^*]\Pr[X_n\le\tau_0{}^*]$$

(56a)

$$+E_{\tau_0,1<n\le n_c,n_c>0,1}[\tau_d \mid m=0 \ \text{ and } \ X_{n-1}<\tau_0{}^*<X_n]\Pr[X_{n-1}<\tau_0{}^*<X_n]$$

(56b)

$$+E_{\tau_0,1<n\le n_c,n_c>0,1}[\tau_d \mid m=0 \text{ and } \ \tau_0{}^*\le X_{n-1}]\Pr[\tau_0{}^*\le X_{n-1}]$$

(56c)

$$=\int_{x=0}^{\tau_0{}^*} x\ f_{X_n}(x)dx\ +\tau_0{}^*\int_{x_{n-1}=0}^{\tau_0{}^*}\int_{x_n=\tau_0{}^*}^{LT_r} f_{X_{n-1},X_n}(x_{n-1},x_n)dx_n dx_{n-1}$$
$$+\int_{x=\tau_0{}^*}^{LT_r} x\ f_{X_{n-1}}(x)dx \qquad (56)$$

Replacing $\tau_0{}^*$ by τ_0 using (50) and from (53)-(56), we have

$$E_{\tau_0,n,n_c,1}[\tau_d \mid m=0]$$
$$= \begin{cases} \tau_0 - T_e + LT_r & , n_c = 0 \\ \int_{x_1=0}^{\tau_0 - T_e + LT_r} x_1 \; f_{X_1}(x_1)dx_1 + \tau_0{}^* \int_{x_1=\tau_0-T_e+LT_r}^{LT_r} f_{X_1}(x_1)dx_1 \, , & , n_c > 0,\ n=1 \\ (\tau_0 - T_e + LT_r)\int_{x_{n_c}=0}^{\tau_0-T_e+LT_r} f_{X_{n_c}}(x_{n_c})dx_{n_c} + \int_{x_{n_c}=\tau_0-T_e+LT_r}^{LT_r} x_{n_c} \; f_{X_{n_c}}(x_{n_c})dx_{n_c} & , n_c > 0,\ n=n_c+1 \\ \int_{x=0}^{\tau_0-T_e+LT_r} x \; f_{X_n}(x)dx + \int_{x=\tau_0-T_e+LT_r}^{LT_r} x \; f_{X_{n-1}}(x)dx & , \; 1 < n \le n_c \\ \quad + (\tau_0 - T_e + LT_r)\int_{x_{n-1}=0}^{\tau_0-T_e+LT_r} \int_{x_n=\tau_0-T_e+LT_r}^{LT_r} f_{X_{n-1},X_n}(x_{n-1},x_n)dx_n dx_{n-1} & \end{cases} \tag{57}$$

Substituting (52) and (57) into (51a) and (51b), we obtain $E_{\tau_0,n}[\tau_d \mid m=0]$. Since τ_0 is uniformly distributed over (0, T_e], we have

$$E[\tau_d \mid m=0] = \sum_{n=1}^{K}\left(\frac{1}{T_e}\right)\int_{\tau_0=0}^{T_e} E_{\tau_0,n}[\tau_d \mid m=0]d\tau_0 \Pr[N_K(t_f)=K-n] \tag{58}$$

where $\Pr[N_K(t_f)=K-n]$ is obtained from (14).

For *m*=0, we have $n_c + n_e \ge n$. Therefore $\Pr[m=0]$ can be expressed as

$$\Pr[m=0] = \sum_{n=1}^{K}\left(\frac{1}{T_e}\right)\int_{\tau_0=0}^{T_e} \sum_{j'=n}^{\infty} \Pr[n_c + n_e = j' \mid \tau_0]d\tau_0 \Pr[N_K(t_f)=K-n] \tag{59}$$

where $\Pr[N_K(t_f)=K-n]$ and $\Pr[n_c + n_e = j' \mid \tau_0]$ are obtained from (14) and (31), respectively.

Appendix D
Notation

- α : the probability that a false failure is detected
- $f_f(t_f)$: the density function for the t_f distribution
- $f_l(t_l)$: the density function for the t_l distribution
- $f_r(t_r)$: the density function for the t_r distribution
- $f_X(x)$: the density function for the x distribution
- $f_{X_i}(x_i)$: the density function for the X_i distribution
- $f_{X_{i-1},X_i}(x_{i-1},x_i)$: the joint density function for $X_{i\text{-}1}$ and X_i
- $F_c(\tau_c)$: the distribution function of τ_c
- $F_L(t_l)$: the distribution function of t_l
- $F_m(\tau(m))$: the distribution function of $\tau(m)$
- $F_r(t_r)$: the distribution function of t_r
- K : the maximum number of consecutive failed deliveries that are attempted before the GSN considers a connection failure occurs
- L: the maximum number of attempts for a GTP' message that the GSN is allowed to send if it does not receive an acknowledgment
- λ_c : the arrival rate of the GTP' charging packets
- $1/\lambda_f$: the expected lifetime of a GTP' connection
- $1/\mu$: the expected round-trip transmission delay for a GTP' message attempt
- m : the number of arrivals for the failed GTP' message deliveries occurring after t_f
- m_c : the number of Echo message arrivals occurring during $\tau(m)$
- m_{e} : the number of charging packet arrivals occurring during $\tau(m)$
- n_c : the number of cross charging packets
- n_e : the number of cross Echo messages
- $N(t_f)$: the number of GTP' message deliveries during t_f

- $N_c(t_f)$: the number of charging packet arrivals (excluding retries) during t_f
- $N_e(t_f)$: the number of Echo message arrivals (excluding retries) during t_f
- N_K: the number of the consecutive failed GTP’ message deliveries
- $N_K(t)$: the N_K value at time t
- $N_{K\to\infty}(t)$: the N_K value at time t when $K \to \infty$
- p: the probability that a GTP’ message delivery is timed out
- t_0: the arrival time of the Echo message prior to t_f
- $t_{a,i}$: the arrival time correspond to the GTP’ message delivery with departure time $t_{d,i}$
- t_d: the time that a true failure is detected
- $t_{d,i}$: the departure time of the i-th failed GTP’ message delivery after t_f
- t_f: the time that a true failure occurs
- t_l : the delivery delay (including retries) for a GTP’ message delivery
- t_r: the round-trip transmission delay for a GTP’ message attempt
- T_e: the fixed interval between two consecutive Echo messages
- T_r: the maximum elapsed time the GSN is allowed to wait for the acknowledgement of a GTP’ message
- τ_0: the period between t_f and the arrival time of the next Echo message
- τ_0^*: the period between $t_f - LT_r$ and the arrival time of the cross Echo message
- τ_c: the interval between t_f and the arrival time of the m_c-th charging packet
- τ_d: the detection time for a true failure
- τ_e: the interval between t_f and the arrival time of the m_e-th Echo message
- $\tau(m)$: the period between t_f and the arrival time of the m-th GTP’ message
- $\theta(j)$: the probability that the false failure is detected at the j-th GTP’ message delivery
- $\bar{\theta}(j)$: the probability that no false failure is detected before (and including) the j-th GTP’ message delivery
- x: the period between $t_f - LT_r$ and the arrival time of an arbitrary cross charging packet
- X_i: the period between $t_f - LT_r$ and the arrival time of the i-th cross charging packet

References

[1] 3GPP. 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 1998), 3G TS 09.60 version 7.10.0 (2002-12), 2002.

[2] 3GPP. 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.

[3] 3GPP. 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.

[4] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; Telecommunication management; Charging management; Charging principles (Release 5), 3G TS 32.200 version 5.6.0 (2004-03), 2004.

[5] 3GPP. 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.

[6] Fang, Y., and Chlamtac, I., Teletraffic Analysis and Mobility Modeling for PCS Networks, *IEEE Transactions on Communications*, 47 (7): 1062-1072, 1999.

[7] Feng, V. W.-S., Wu, L.-Y., Lin, Y.-B., and Chen, W.E.WGSN: WLAN-based GPRS Environment Support Node with Push Mechanism. Accepted and to appear in *The Computer Journal*, 2003.

[8] Holma, H., and Toskala, A. (edited). *WCDMA for UMTS*. John Wiley & Sons, 2000.

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

[10] Lin, Y.-B., and Chen, Y.-K. Reducing Authentication Signaling Traffic in Third Generation Mobile Network. *IEEE Transactions on Wireless Communications*, 2(3): 493-501, 2003.

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

[12] Lin, Y.-B., Haung, Y.-R., Pang, A.-C., and Chlamtac, I. All-IP Approach for UMTS Third Generation Mobile Networks. *IEEE Network*, 16(5): 8-19 2002.

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

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

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