

國立交通大學

資訊工程學系

博士論文



行動通信網路之高效能
即時計費機制設計

A Design of Efficient Online Charging
Mechanisms for Mobile Telecommunications

研究生：蘇 淑 茵

指導教授：林 一 平 博士

中 華 民 國 九 十 七 年 一 月

本論文係接受
聯發科技獎學金贊助



This work was supported by
the Media Tek Fellowship

國立交通大學

資訊工程學系

博士論文



行動通信網路之高效能
即時計費機制設計

A Design of Efficient Online Charging
Mechanisms for Mobile Telecommunications

研究生：蘇 淑 茵

指導教授：林 一 平 博士

中 華 民 國 九 十 七 年 一 月

行動通信網路之高效能
即時計費機制設計

A Design of Efficient Online Charging
Mechanisms for Mobile Telecommunication

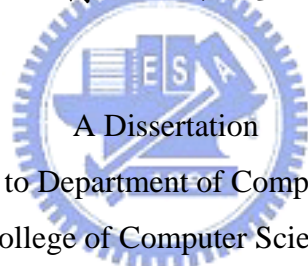
研究生：蘇淑茵

Student : Sok-Ian Sou

指導教授：林一平

Advisor : Yi-Bing Lin

國立交通大學
資訊工程系
博士論文



Submitted to Department of Computer Science

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

Jan 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年一月

行動通信網路之高效能 即時計費機制設計

研究生：蘇淑茵

指導教授：林一平博士

國立交通大學
資訊工程學系

摘要

隨著網際網路的蓬勃發展及通信技術的急速演進，利用網際網路提供高質素的多媒體服務已成為行動市場之主流趨勢。為了在行動通信系統上整合網際網路技術，第三代行動通訊規格組織（3GPP）制定了全球移動通信系統（UMTS）完全 IP（All-IP）架構的規格。其中，3GPP 第五版提出了一個處理網際網路多媒體服務的核心網路子系統（IMS）。透過 IMS 平台，行動業者能夠有效率地開發及整合包括語音、視訊與數據等組合的多媒體服務。然而，在 IMS 服務正式上市並獲取利潤前，行動業者必須先處理關於收費系統的各项議題。

若要成功地向用戶推廣 IMS 多媒體服務，行動電信業者必須能正確及合理地針對服務內容的價值收費。為了能夠有效率地開發及建置 IMS 服務，3GPP 第五版與第六版提出了一個具有開放性和通用性的即時計費系統（OCS）。透過即時計費的管理機制，業者可以更彈性地即時處理網路資源授權、客戶帳戶餘額以及線上批價等等。

行動服務的即時收費受許多因素影響，其中包括使用者的收費模式，服務類型，資源分配與服務品質等等。因此，即時計費系統的設計必須要達到兩項要求。第一，增加系統處理計費事項的準確性；第二，盡量減低控制訊息的交換。在這

篇論文的第一部份，我們首先提出一個能在 IMS 架構上整合預付式網際網路服務的伺服器。當預付式語音服務與訊息服務同時提供時，其中一項重要議題是要在不中斷進行中的語音服務下，提供即時訊息的傳送服務。在本論文中，我們提出一項控制即時訊息傳送的策略，並分析其對整個系統之效能影響。

在本論文的第二部份，我們研發出能在即時收費系統上提供高效能、低成本的餘額分配管理機制。我們探討兩個重要的即時收費程序：餘額保留機制與餘額重新授權機制。我們首先針對預付式使用者，解決了當餘額不足時之處理程序。在這個機制中，當使用者之可用餘額低於系統預設的臨界值時，即時計費系統會向使用者發出餘額過低的警示訊息，以便提醒使用者盡快補充其預付帳戶內之金額。業者在使用這個機制時，必須謹慎地設定這個臨界值參數。一方面要避免因餘額不足而被迫中斷使用中的服務的機率，另一方面要降低警示訊息之發送頻率。我們探討了這個臨界值的設置對即時收費系統之各種影響，並向業者提供適當之參數設定建議。為了進一步提高即時收費系統的效能，我們研究由服務品質改變而引起之餘額重新授權程序。本論文以一個動態改變的臨界值機制，有效地降低餘額重新授權程序中所交換的控制訊息數目。

針對以上三個研究題目，我們分別發展出數學模型及模擬實驗，以準確分析影響系統效能之各項指標。根據本論文的研究結果，我們為行動業者提供各項即時收費機制之參數建議。實驗結果證實了我們提出的方法能夠提高整個即時收費系統之效率。

關鍵字： 第三代行動通訊規格組織, IP 多媒體子系統, 行動通訊網路, 即時計費, 預付式服務

A Design of Efficient Online Charging Mechanisms for Mobile Telecommunications

Student: Sok-Ian Sou

Advisor: Dr. Yi-Bing Lin

Institute of Computer Science and Engineering
National Chiao Tung University

ABSTRACT

Introduction of the 3G mobile system has driven the Internet into new markets to support mobile users. To integrate IP with wireless technologies with the right business model, the *3rd Generation Partnership Project* (3GPP) proposed UMTS all-IP architecture that evolved from *Global System for Mobile Communications* (GSM) and *General Packet Radio Service* (GPRS). 3GPP Release 5 introduces the *IP Multimedia Core Network Subsystem* (IMS) on top of the PS service domain to provide multimedia services. Before the benefits of IMS and all-IP can be realized, a fundamental issue must be addressed is charging/billing, which is one of the most important activities in telecommunications.

In order to successfully promote IMS services, how to charge packet data service accurately and reasonably has become a major concern of operators. The deployment of the IP multimedia services requires effective charging management system and efficient service delivery mechanism. Therefore, 3GPP Releases 5 and 6 propose the convergent and flexible IP-based *Online Charging System* (OCS) to incorporate data applications with real-time control and management. Through online charging, the operator can ensure that credit limits are enforced and resources are authorized on a per-transaction basis.

In mobile data services, there are many factors that affect charging, for example, the service type, the amount of usage and the provisioned QoS. The design of online

charging system needs to reduce the signaling overhead and improve the system performance. In the first part of the dissertation, we first study how to develop an IMS application server to integrate existing IP-based prepaid services in the UMTS environment. When both voice and messaging are simultaneously offered, a potential problem is that the delivery of a prepaid message during a call may result in force-termination of that call due to credit depletion. To address this issue, we describe a strategy to determine if a prepaid message can be sent out during a call session.

In the second part of the dissertation, we investigate how to provide efficient credit quota management with lower signaling overhead and higher accuracy in the OCS. Specifically, we study the credit reservation and credit re-authorization mechanisms in the OCS, which are two most essential procedures in online charging management. First, we study the online credit reservation procedure for prepaid users in the UMTS online charging system. When the remaining amount of prepaid credit is below a threshold, the OCS should remind the user to refill the prepaid account. It is essential to choose an appropriate recharge threshold not only to avoid forced termination for the in-progress service sessions but also to decrease the recharge signaling overhead. Then we focus on the online credit re-authorization procedure of the OCS. The main objective is to reduce the traffic signaling overhead for credit re-authorization due to frequency QoS changes. We propose a threshold-based scheme that can significantly reduce the number of re-authorization message exchanges during a session.

We also develop analytic models and simulations experiments to evaluate the performance of the proposed schemes. Based on our study, the mobile operator can achieve high performance in the online charging management by configuring appropriate parameters.

Keywords: The 3rd Generation Partnership Project (3GPP), IP Multimedia Subsystem (IMS), mobile communication network, online charging, prepaid services

Acknowledgements

I would like to express my sincere gratitude to the following people who made this dissertation possible.

First of all, I am deeply indebted to my advisor Prof. Yi-Bing Lin for his help and full support throughout my research. In the course of my study, he has given me encouragement and guided me to the right research direction. I wish there were enough words for me to thank him. Without his supervision and perspicacious advice, I would not have completed this dissertation.

Next, I would like to gratefully and sincerely thank my committee members, Prof. Gin-Lian Chen, Prof. Chu-Sing Yang, Prof. Wanjiun Liao, Prof. Yao-Nan Lien, Prof. Ming-Feng Chang and Prof. Wen-Nung Tsai for their valuable and insightful comments. I would also like to express my thanks to Prof. Hui-Nien Hung, Prof. Yu-Chee Tseng, Prof. Rong-Jaye Chen, Prof. Quincy Wu, Dr. Jeu-Yih Jeng, Prof. Phone Lin, Prof. Ai-Chun Pang, Dr. Yan-Kai Chen, Prof. Shun-Ren Yang, Prof. Pei-Chun Lee and Dr. Lin-Yi Wu for their endless help. I would like to extend my heartfelt gratitude to all my teachers, friends and labmates met in NCTU. Special thanks to my best friends Chang, Karen, Cynthia and Miranda. I have been blessed with friendships that serves as a precious source of support, enjoyment and encouragement especially in times of difficulty and frustration.

Last but not the least, I wish to give my special thanks to my dear parents, my elder sister Bonny, my younger sister Sofia and my love Yinman for their patient love enabled me to complete this work. Their unlimited love is the best impetus and encouragement in my life.

This work was supported by the MediaTek Fellowship.

Sok-Ian (Ines) Sou
2008

Contents

Chinese Abstract	i
English Abstract	iii
Acknowledgements	v
Contents	v
List of Tables	ix
List of Figures	x
1 Introduction	1
1.1 Motivation	2
1.2 IP Multimedia Core Network Subsystem Architecture	3
1.3 IP-based Online Charging System (OCS)	5
1.4 IP-based Online Charging Protocol	7
1.4.1 Diameter Credit Control Message	8
1.4.2 Diameter Message Flow	10
1.5 Online Charging Scenarios	11
1.5.1 Immediate Event Charging	12
1.5.2 Event Charging with Unit Reservation	13
1.5.3 Session Charging with Unit Reservation	14
1.6 Dissertation Organization	16
2 Design of IMS Prepaid Application Server for SIP-based Services	18
2.1 Prepaid Application Server of SIP-based Services	19
2.2 Integrating Charging for Prepaid Calls and Instant Messaging	21
2.2.1 Prepaid IMS-to-PSTN Call Setup and Release	21



2.2.2	Prepaid Instant Messaging Delivery	24
2.3	Analytic Modeling for Prepaid Application Server	25
2.3.1	Derivation for the Remaining Credit Density Function	26
2.3.2	Derivation for the Unnecessary Force-termination Probability	29
2.3.3	Derivation for the Unnecessary Delay	29
2.4	Simulation Validation	30
2.5	Numerical Examples	33
2.6	Summary	37
2.7	Notation	37
3	Modeling Online Credit Reservation Procedure	39
3.1	Recharge Threshold-based Credit Reservation (RTCR)	40
3.2	Analytic Modeling for RTCR Mechanism	41
3.2.1	Derivation for the Number of RU Operations	41
3.3	Exact Analytic Model for Single-type Service	42
3.3.1	Derivation for the Force-termination Probability	43
3.3.2	Derivation for the Unused Credit Units	46
3.4	Approximate Analytic Model for Multiple-type Services	46
3.4.1	Derivation for the Force-termination Probability	46
3.4.2	Derivation for the Unused Credit Units	48
3.5	Simulation Validation	49
3.6	Numerical Examples	50
3.7	Summary	54
3.8	Notation	54
4	Reducing Credit Re-authorization Cost	56
4.1	Online Charging System for GPRS Sessions	56
4.1.1	Credit Re-authorization Procedure	57
4.1.2	Threshold-based Scheme for Credit Re-authorization	60
4.2	Analytic Modeling for the Basic Scheme	61
4.2.1	Derivation for the Number of ABMF message Exchanges	62
4.2.2	Derivation for the Inaccuracy of Credit Information	63

4.3	Analytic Modeling for the Threshold-based Scheme	63
4.3.1	Derivation for the Number of ABMF message Exchanges	64
4.3.2	Derivation for the Inaccuracy of Credit Information	68
4.4	Simulation Validation	73
4.5	Numerical Examples	74
4.6	Summary	76
4.7	Notation	77
5	Conclusions and Future Work	80
5.1	Concluding Remarks	80
5.2	Future Work	81
	Bibliography	83
	Publication List	89



List of Tables

2.1	Comparison of the Analytic and Simulation Results ($X_c > 0$, $T_m = 5$ CU, $X = 20T_m$, $1/\lambda_c = 50$ TU)	32
3.1	Comparison of the Analytic and Simulation Results ($n = 1$)	49
3.2	Comparison of the Analytic and Simulation Results ($n = 2$)	50
4.1	Comparison of the Analytic and Simulation Results ($\delta = 0$, $N = 2$ and $\hat{\alpha} = 2$)	73



List of Figures

1.1	The IMS Network Architecture.	4
1.2	The Online Charging System Architecture.	6
1.3	Diameter Message Flow for Online Session.	10
1.4	IMS Message Flow for Immediate Event Charging.	12
1.5	IMS Message Flow for Event Charging with Unit Reservation.	13
1.6	IMS Session with Online Charging.	14
1.7	IMS Message Flow for Session Charging with Unit Reservation.	15
2.1	IMS Environment for SIP-based Prepaid Services.	20
2.2	Message Flow for Prepaid IMS-to-PSTN Call Setup and Release.	22
2.3	Message Flow for Prepaid Messaging Delivery.	24
2.4	Timing Diagram for Prepaid Calls and Messaging Deliveries.	26
2.5	Simulation Flow Chart.	31
2.6	Effects of $E[t_c]$ on P_{UFT} and $E[t_d]$ ($X = 50E[t_c]$ and $\lambda_m = 5\lambda_c$)	34
2.7	Effects of V_c on P_{UFT} and $E[t_d]$ ($X = 25E[t_c]$ and $1/\lambda_m = 0.5E[t_c]$)	35
2.8	Effects of $1/\lambda_m$ on P_{UFT} and $E[t_d]$ ($X = 25E[t_c]$)	35
2.9	Effects of X on P_{UFT} and $E[t_d]$ ($1/\lambda_m = 0.5E[t_c]$)	36
3.1	Timing Diagram for the RTCR mechanism.	42
3.2	Timing Diagram for Deriving $\tilde{\theta}$	44
3.3	Effects of θ_i on $E[N_{r,i}]$	51
3.4	Effects of θ_i and C_{min} ($n = 2$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$)	52
3.5	Effects of $V_{h,i}$ ($n = 2$, $\theta_i = 2.5\mu_i$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$)	53
3.6	Effects of n ($C_{min} = 6c$, $\lambda_i = i\mu_1$ and $\mu_i = i\mu_1$)	53
4.1	Message Flow for Credit Re-authorization Procedure.	58
4.2	Timing Diagram for the Basic Scheme.	62
4.3	Timing Diagram for the Threshold-based Scheme.	65

4.4	Probability Transition Diagram ($N = 2$).	66
4.5	Timing Diagram for Deriving C_T	68
4.6	Effects of τ_g ($N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$)	74
4.7	Effects of δ ($N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$)	75
4.8	Effects of δ and p_0 ($N = 2$, $\alpha_2 = 2\alpha_1$ and $\tau_g = 1/\lambda$)	76
4.9	Effects of δ and N ($\alpha_i = i\alpha_1$, $p_0 = 0.01$ and $\tau_g = 1/\lambda$)	77



Chapter 1

Introduction

By providing ubiquitous connectivity for data communications, the Internet has become the most important vehicle for global information delivery. Introduction of the 3G mobile system has further driven the Internet into new markets to support mobile users. Specifically, the *Internet Protocol* (IP) has played a major role in UMTS in providing a wide range of connectionless services to mobile users [37].


Internet environment encourages global usage with flat-rate tariffs and low entry costs. A major problem of the “flat-rate” tariffs is that such business model cannot justify the expensive equipment/operation investments of mobile services. Mobile telecom operators have to move from a bit-pipe model to a revenue-generating services model. To integrate IP with wireless technologies with the “right” business model, the *3rd Generation Partnership Project* (3GPP) proposed UMTS all-IP architecture to enable Web-like services and new billing paradigm in the telephony world. This architecture has evolved from *Global System for Mobile Communications* (GSM) and *General Packet Radio Service* (GPRS) [1, 12]. 3GPP Release 5 introduces the *IP Multimedia Core Network Subsystem* (IMS) on top of the PS service domain to provides multimedia services [18]. This evolution requires new mechanisms to collect information about chargeable events and to impose flexible mobile billing schemes (such as time-based, volume-based, or content-based). By creating hybrid online/offline billing models, 3GPP Releases 5 and 6 propose the IP-based *Online Charging System* (OCS) [10] to allow both prepaid and postpaid subscribers to be charged in real-time. Through online charging, the operator can ensure that credit limits are enforced and resources are authorized on a per-transaction basis. From a subscriber’s perspective, knowing the charges in advance and having self-imposed credit limits can make the cost of services more transparent. In other words, real-time rating/charging functions help subscribers to control their budget, and telecom operators to reduce bad

debt.

Online charging requires a single set of management processes, such as service development, production support and pricing, which reduces operational costs and enhances flexibility on billing and product diversification. The OCS approach is estimated to reduce 25% of the product launch costs [28]. This real-time solution provides two-way communications between network nodes and the charging/billing system, which transfer information about rating, billing and accounting. When the OCS receives a service request, it queries other relevant components, then determines and returns a response to the network node. In contrast, in an offline environment, all usage records typically flow through the billing system in one direction after the service has been delivered.

This introductory chapter first describes the motivation of our work on the mobile online charging mechanism. We briefly introduce the UMTS/IMS architecture and its IP-based online charging system. Then we present the mobile charging protocol that used to control online IMS services in real-time. Finally, we discuss the performance analysis issues on the mobile online charging system and the organization of this dissertation.

1.1 Motivation



Traditional voice and basic data services such as *Short Message Service* (SMS) and *Multimedia Messaging Service* (MMS) deliveries are the principal sources of revenue for most mobile operators. The deployment of IMS will bring more competition for more specialized services and content. However, before the benefits of IMS and all-IP can be realized, a fundamental issue must be addressed is charging/billing, which is one of the most important activities in telecommunications. The IP-based services for GPRS and IMS specify more critical charging requirements that impose flexible mobile billing schemes (e.g., time-based, volume-based, content-based) [20, 29, 56].

With the wide application of packet data service, how to charge packet data service accurately and reasonably has become a common concern of operators. Creating a killer environment has now become a key point in terms of competition. The deployment of the IP multimedia services requires effective charging management system and efficient service delivery mechanism. Therefore, advanced mobile telecom requires a convergent and flexible online charging system to incorporate data applications with real-time control and management. Such convergence is essential to mitigate fraud and credit risks and provide more personalized advice to users about charges and credit limit controls.

In mobile data services, there are many factors that affect charging, for example, the content type, the radio access technology, the location information, the amount of usage, the provisioned QoS and so on [4, 5, 7, 9, 11, 16]. The design of online charging system needs to reduce the signaling overhead and improve the system performance. In this dissertation, we first study how to develop an IMS application server to integrate existing IP-based prepaid services in the UMTS environment. Then we investigate how to provide efficient credit quota management with lower signaling overhead and higher accuracy in the OCS. Specifically, we study the credit reservation and credit re-authorization mechanism in the OCS, which are two most essential procedures in online charging management.

1.2 IP Multimedia Core Network Subsystem Architecture

To effectively integrate mobile technology with the Internet, 3GPP Release 5 introduces the IMS architecture that effectively provides multimedia services [6, 8, 18, 47]. The IMS protocols allow the telecom operators to bring attractive new services to their customers. Such protocols are namely *Session Initiation Protocol* (SIP) for signaling and Diameter for *Authentication, Authorization and Accounting* (AAA). This section introduces the UMTS/IMS architecture.

As illustrated in Fig. 1.1, the IMS connects the *Gateway GPRS Support Node* (GGSN; Fig. 1.1 (a)) with the external *Packet Data Network* (PDN; Fig. 1.1 (b)) and the *Public Switched Telephone Network* (PSTN; Fig. 1.1 (i)). In this figure, the dashed lines represent signaling links and the solid lines represent data and signaling links. The IMS nodes are similar to those in a SIP-based *Voice over IP* (VoIP) network [27, 49]. Details are described as follows:

- The *Call Session Control Functions* (CSCFs; Fig. 1.1 (c), (d), and (e)) are SIP servers, which are in charge of call control and communications with the *Home Subscriber Server* (HSS; see Fig. 1.1 (m)) regarding location information. Specifically, IMS signaling is carried out by the *Proxy CSCF* (P-CSCF; Fig. 1.1 (c)), the *Interrogating CSCF* (I-CSCF; Fig. 1.1 (d)) and the *Serving CSCF* (S-CSCF; Fig. 1.1 (e)). When a *User Equipment* (UE) attaches to the GPRS/IMS network and performs *Packet Data Context* (PDP) context activation, a P-CSCF is assigned to the UE. The P-CSCF contains limited address translation functions to forward the requests

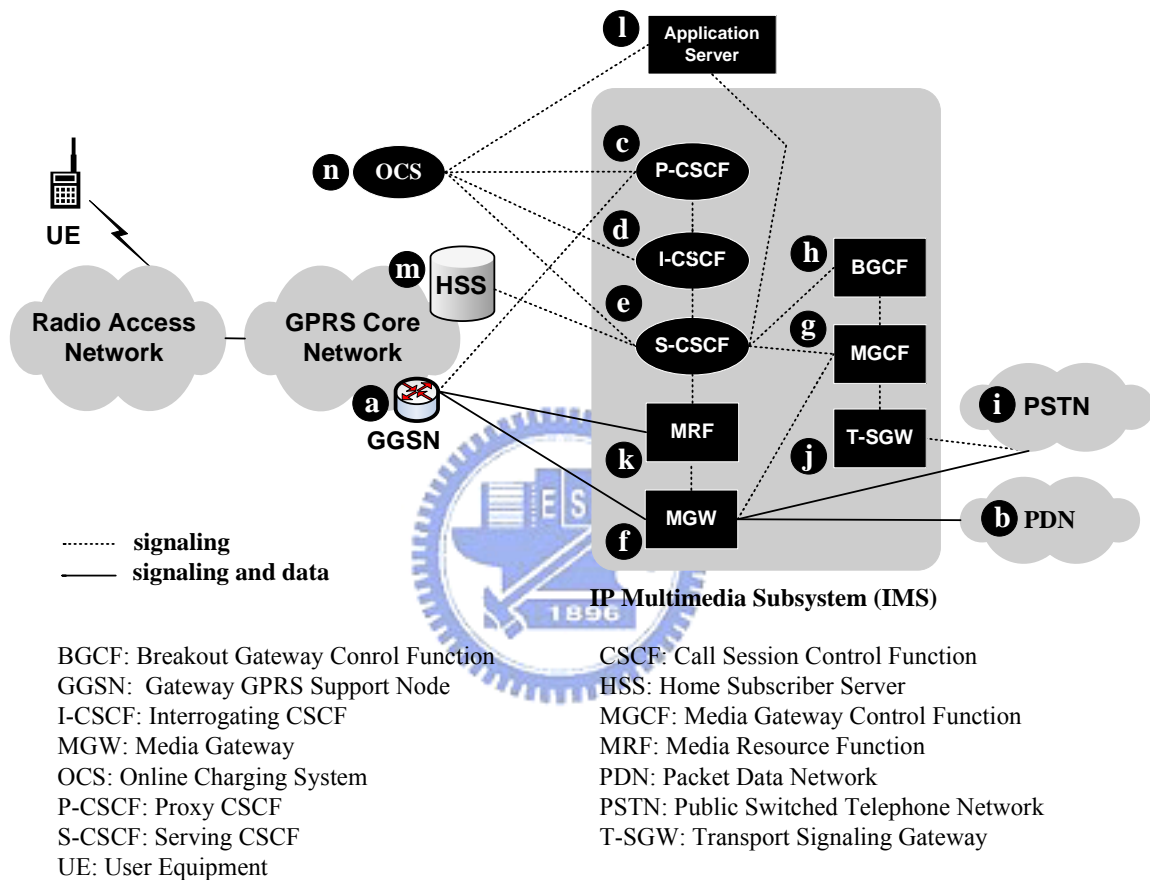


Figure 1.1: The IMS Network Architecture.

to the I-CSCF. The I-CSCF selects an S-CSCF to serve the UE during the IMS registration procedure. This S-CSCF supports the signaling for call setup and supplementary services control. All incoming calls are routed to the destination UE through the S-CSCF.

- The *Media Gateway* (MGW; Fig. 1.1 (f)) transports the IMS user data traffic. The MGW provides user data transport between the UMTS core network and the PSTN (including media conversion bearer control and payload processing).
- The *Media Gateway Control Function* (MGCF; Fig. 1.1 (g)) controls the media channels in an MGW.
- The *Breakout Gateway Control Function* (BGCF; Fig. 1.1 (h)) selects the network in which the PSTN (or circuit switched domain) breakout is to occur. If the BGCF determines that a breakout is to occur in the same network, it selects an MGCF that is responsible for interworking with the PSTN. If the breakout occurs in another IMS network, the BGCF forwards the SIP request to another BGCF or an MGCF in the selected IMS network.
- The *Transport Signaling Gateway* (T-SGW; Fig. 1.1 (j)) serves as the PSTN signaling termination point and provides the PSTN/legacy mobile network with IP transport level address mapping. Specifically, it maps call-related signaling between the MGCF and the PSTN.
- The *Media Resource Function* (MRF; Fig. 1.1 (k)) performs functions such as multiparty call, multimedia conferencing, and tone and announcement.

In Fig. 1.1, an *application server* (Fig. 1.1 (l)) provides value added IP multimedia services which reside either in the user's home IMS or in a third party location. In Chapter 2, we will design an IMS prepaid application server to integrate online prepaid call and messaging services. The OCS (Fig. 1.1 (n)) performs online charging and collects the billing information with the CSCFs. Details of the OCS functionalities are described in the next section.

1.3 IP-based Online Charging System (OCS)

This section introduces the UMTS online charging system. In online charging, network resource usage is granted by the OCS based on the price or the tariff of the requested

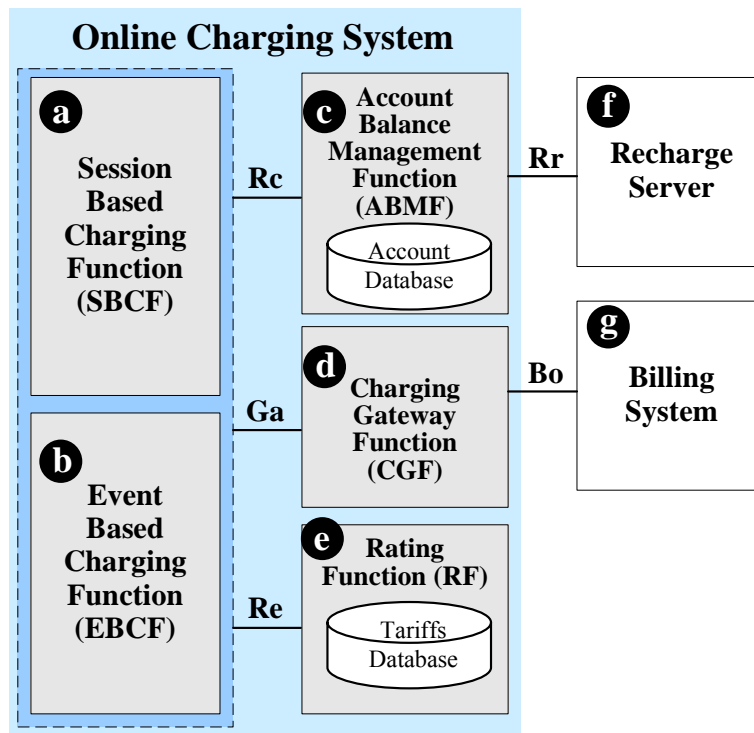


Figure 1.2: The Online Charging System Architecture.

service and the balance in the subscriber's account. Fig. 1.2 shows the OCS architecture defined in 3GPP 32.296 [10]. The OCS supports two types of *Online Charging Functions* (OCFs), namely the *Session Based Charging Function* (SBCF) and the *Event Based Charging Function* (EBCF). The SBCF (Fig. 1.2 (a)) is responsible for network bearer and session-based services such as voice calls, GPRS sessions or IMS sessions. The EBCF (Fig. 1.2 (b)) is responsible for event-based services, such as SMS/MMS delivery, and content (e.g., music or ring tones) downloading.

The *Account Balance Management Function* (ABMF; Fig. 1.2 (c)) maintains user balances and other account data. When a user's credit depletes, the ABMF connects the *Recharge Server* (Fig. 1.2 (f)) to trigger the recharge account function. The SBCF/EBCF interacts with the ABMF to query and update the user's account. The CDRs generated by the charging functions are transferred to the *Charging Gateway Function* (CGF; Fig. 1.2 (d)) in real-time [32]. The CGF acts as a gateway between the IMS/UMTS network and the *Billing System* (Fig. 1.2 (g)).

In this architecture, the *Rating Function* (RF; Fig. 1.2 (e)) determines the price/tariff of the requested network resource usage. The RF handles a wide variety of ratable in-

stances, including data volume, session/connection time, or service events. In addition, different charging rates can be adopted for different time segments (e.g., peak hours and off-peak hours). The OCF interacts with the RF to determine the price/tariff of the requested service through the Re interface. The details of the message exchanges are described as follows:

- The **Price Request/Response** message pair determines the price for an event-based service (e.g., SMS). This message exchange can be executed before the service delivery (e.g., for advice of charge and prepaid service) or after service delivery (e.g., for postpaid service). Based on the price information, the EBCF calculates the amount of credit granted to the network nodes.
- The **Tariff Request/Response** message pair determines the tariff information for a session-based service (e.g., voice calls). Based on the tariff information, the SBCF calculates the amount of credit granted to the network nodes.

1.4 IP-based Online Charging Protocol

This section presents the mobile charging protocol that used for credit control of online IMS services. The Diameter protocol was derived from *Remote Access Dial In User Service* (RADIUS) protocol with more flexibility, and is generally believed to be the next generation *Authentication, Authorization, and Accounting* (AAA) protocol [21, 45, 46]. Diameter is an extensible protocol enabling AAA within and across IP multimedia networks that relies on secure and reliable transports. Its modular architecture offers a flexible base protocol which allows application-specific extensions. Diameter has proven successful in overcoming the limitations of RADIUS. Development of new IP- and telephony-based services is gaining momentum, and every service requires charging. Therefore, rapid growth in the usage of Diameter-based charging can be expected. The 3GPP has chosen the Diameter protocol to enable IMS network AAA capabilities [18, 39, 42].

Like RADIUS, Diameter follows the client-server architecture, where a client and a server interact through the Diameter request and answer message exchange. Several Diameter applications defined by *Internet Engineering Task Force* (IETF) are utilized in IMS, including the *Diameter Credit Control* (DCC) application for IP-based online charging control [30]. All-IP mobile network utilizes the DCC application to communicate with the OCS [10, 15]. In IMS online charging, the IMS network node (e.g., CSCF) acts

as a DCC client and the OCS acts as a DCC server. In this section, we describe the Diameter credit control messages and the Diameter message flow.

1.4.1 Diameter Credit Control Message

The DCC messages are described as follows: The *Credit Control Request* (CCR) message is sent from the DCC client to the DCC server to request an amount of credit for an online service. The CCR message contains the following *Attribute Value Pairs* (AVPs) [15, 30]:

- The *Session-Id* AVP identifies the credit control session.
- The *Origin-Host* and the *Origin-Realm* AVPs contain the address and the realm of the DCC client.
- The *Destination-Host* and the *Destination-Realm* AVPs contain the address and the realm of the DCC server.
- The *Auth-Application-Id* AVP contains value 4 as defined in RFC 4006 [30]. This value indicates the Diameter Credit Control Application.
- The *Service-Context-Id* AVP contains the identifier allocated by the service provider or by the standardization body.
- The *CC-Request-Type* AVP indicates the credit control request type for the on-line service. It can be one of the following types: INITIAL_REQUEST (value 1) initiates a credit control session. UPDATE_REQUEST (value 2) contains update credit control information for an in-progress session. This request is sent when the credit units currently allocated for the session are completely consumed. TERMINATION_REQUEST (value 3) terminates an in-progress credit control session. EVENT_REQUEST (value 4) is used in one-time credit control for event-based service.
- The *CC-Request-Number* AVP contains the sequence number of the message.
- The *Subscription-Id* AVP contains the identification of the user, which can be a SIP *Uniform Resource Identifier* (URI), an *International Mobile Subscriber Identity* (IMSI) or a *Mobile Station ISDN Number* (MSISDN).

- The *Termination-Cause* AVP is presented if the *CC-Request-Type* is set to TERMINATION_REQUEST. This AVP provides the reason why the credit control session is terminated. For example, DIAMETER_LOGOUT (value 1) indicates that the user terminates the credit control session.
- The *Requested-Action* AVP is only presented in the EVENT_REQUEST message. This AVP specifies the action for the event, such as DIRECT_DEBITING, REFUND_ACCOUNT, CHECK_BALANCE or PRICE_ENQUIRY.
- The *Multiple-Services-Credit-Control* AVP contains the parameters for quota management, including the amount of request credit, the amount of used credit, the reporting reason, the identity of the used service, and the identifier of the rating group.
- The *User-Name* AVP contains the user name in the *Network Access Identifier* (NAI) format according to RFC 3588 [21].
- The *Event-Timestamp* AVP specifies the time when the accounting message is created.
- The *Service-Information* AVP contains the service specific parameters.

The *Credit Control Answer* (CCA) message is sent from the DCC server to the DCC client to grant an amount of credit for the online service. The CCA message contains the following AVPs:

- The *Result-Code* AVP contains the result of the credit control request.
- The *CC-Session Failover* AVP contains an indication to the DCC client whether or not a failover handling is used.
- The *Credit-Control-Failure-Handling* AVP determines the action in the Diameter client when the Diameter server is temporarily prevented, e.g., because of network failure. The action may be TERMINATE (with value 0), CONTINUE (with value 1) and RETRY_AND_TERMINATE (with value 2).
- The *Multiple-Services-Credit-Control* AVP contains the parameters for the quota management, including the amount of granted credit, the identifier of the requested service, the identifier for the rating group, the validity time for the usage of granted

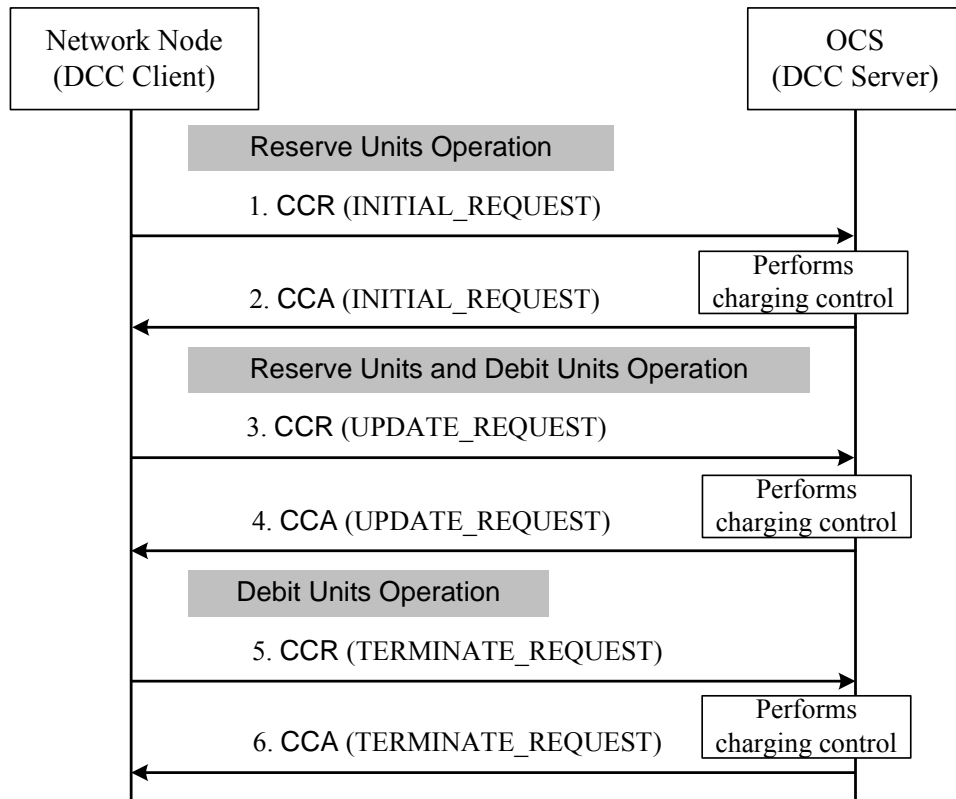


Figure 1.3: Diameter Message Flow for Online Session.

credit units (i.e., the time when the DCC client should perform another credit control request), and the events (such as QoS changes or SGSN changes) that will trigger credit re-authorization procedure.

- The *Cost-Information* AVP contains the cost information of the requested service.
- The *Session-Id*, the *Origin-Host*, the *Origin-Realm*, the *Auth-Application-Id*, the *CC-Request-Type*, the *CC-Request-Number*, the *Service-Information* AVPs are similar to those in the CCR message.

1.4.2 Diameter Message Flow

The Diameter message flow for session-based online charging includes three types of credit control operations: Reserve Units operation (Steps 1 and 2 in Fig. 1.3), Reserve Units and Debit Units operation (Steps 3 and 4 in Fig. 1.3) and Debit Units operation (Steps 5 and 6 in Fig. 1.3). The following operations are executed for session-based service.

Step 1. [Reserve Units (request)] To start the service with credit reservation, the net-

work node (e.g., GGSN or S-CSCF) sends the CCR message with *CC-Request-Type* “INITIAL_REQUEST” to the OCS. This message indicates the amount of requested credit.

Step 2. [Reserve Units (answer)] Upon receipt of the CCR message, the OCS determines the tariff of the requested service and then reserves an equivalent amount of credit. After the reservation is performed, the OCS acknowledges the network node with the CCA message including credit reservation information.

Step 3. [Reserve Units and Debit Units (request)] During the service session, the granted credit units may be depleted. If so, the network node sends a CCR message with *CC-Request-Type* “UPDATE_REQUEST” to the OCS. The network node reports the amount of used credit, and requests for additional credit units.

Step 4. [Reserve Units and Debit Units (answer)] When the OCS receives the CCR message, it debits the amount of consumed credit and reserves extra credit units for the service session. The OCS acknowledges the network node with the CCA message with the *Result-Code* “DIAMETER_SUCCESS” and the amount of the reserved credit. Note that the Reserve Units and Debit Units operation (i.e., Steps 3 and 4) may repeat many times before the service session is complete.

Step 5. [Debit Units (request)] When the service session is complete, the network node sends the CCR message with *CC-Request-Type* “TERMINATE_REQUEST” to the OCS. This action terminates the session and reports the amount of the consumed credit.

Step 6. [Debit Units (answer)] The OCS debits the consumed credit units and releases the unused reserved credit units. The OCS acknowledges the network node with the CCA message. This message may contain the total cost of the service.

1.5 Online Charging Scenarios

In this section, we elaborate more on how the IMS node and the OCS can manage online credit for the service delivery in real-time. 3GPP Release 6 defines three kinds of online charging: immediate event charging, event charging with unit reservation and session charging with unit reservation. The message flows for these online charging scenarios are explained in the following subsections.

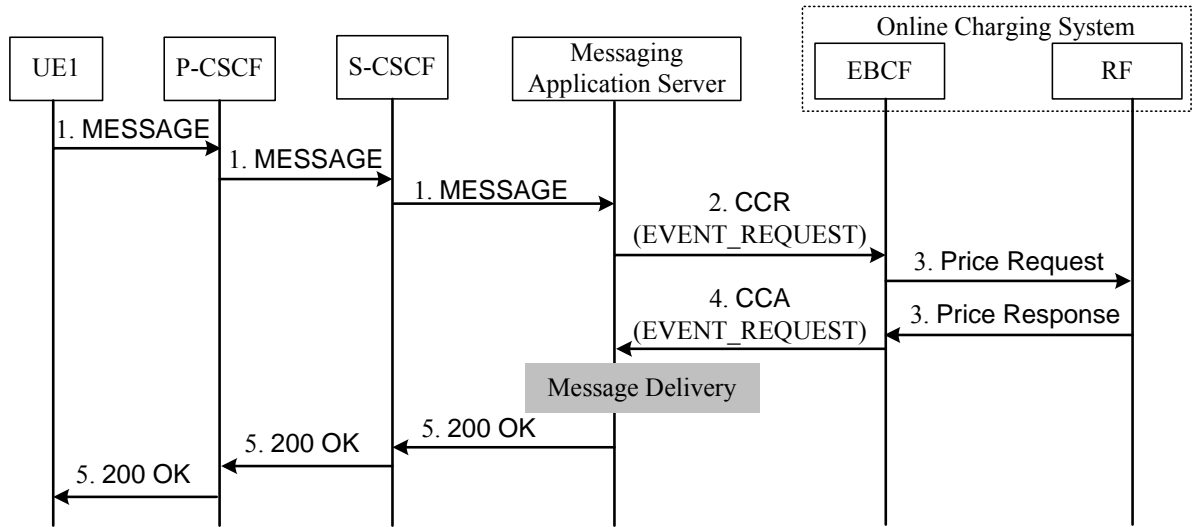


Figure 1.4: IMS Message Flow for Immediate Event Charging.

1.5.1 Immediate Event Charging

In immediate event charging, credit allocation to an IMS node is performed in a single operation, and the credit units are deducted immediately from the subscriber’s account. Assume that UE1 subscribes to the messaging service with online charging. The message flow for IMS event-based messaging service (offered by an IMS messaging application server) is described below (see Fig. 1.4).

Step 1. Subscriber UE1 sends a SIP MESSAGE request to the IMS messaging application server through the P-CSCF and the S-CSCF.

Step 2. The application server sends a CCR message with *CC-Request-Type* “EVENT_REQUEST” and *Requested-Action* “DIRECT_DEBITING” to the EBCF in the OCS.

Step 3. Upon receipt of the credit control request, the EBCF requests account information for the subscriber (such as billing plan and subscription profile) from the ABMF. Then the EBCF sends a Price Request message to the RF. The RF determines the price of the service based on the subscriber’s billing plan. Specifically, the RF calculates the price for the given service according to the service and subscriber information specified in the request. The calculated price and the billing information is returned to the EBCF through the Price Response message.

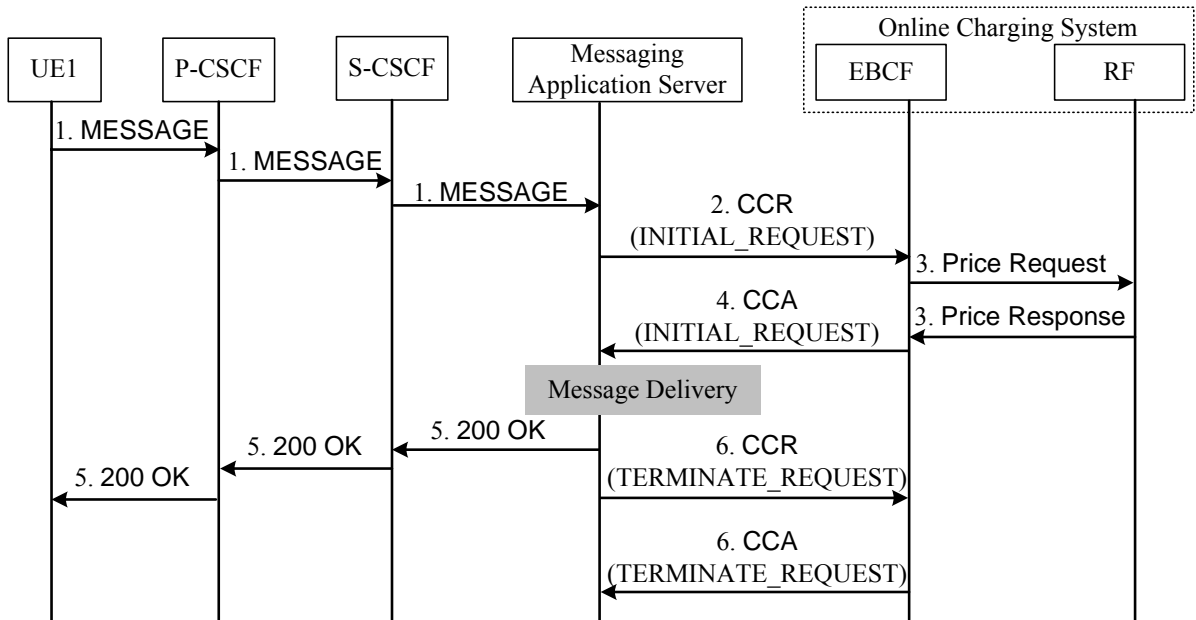


Figure 1.5: IMS Message Flow for Event Charging with Unit Reservation.

Step 4. The EBCF performs credit unit deduction through the ABMF. When the deduction is complete, the EBCF replies to the application server with the CCR message.

Step 5. After the application server has obtained the credit indicated in the CCR message, it delivers the messaging service to UE1. If the service is delivered successfully, the application server sends a SIP 200 OK message to UE1 through the S-CSCF and the P-CSCF. If the credit control fails, an appropriate SIP error message (e.g. 401 Unauthorized or 402 Payment Required) is sent to UE1.

1.5.2 Event Charging with Unit Reservation

Event charging with unit reservation conducts reserving and returning unused credit units for an event-based service. The message flow in Fig. 1.5 is described as follows:

Step 1. Subscriber UE1 sends a SIP MESSAGE request to the IMS messaging application server through the P-CSCF and the S-CSCF.

Step 2. The application server sends a CCR message with *CC-Request-Type* “INITIAL_REQUEST” to the EBCF of the OCS.

Step 3. Upon receipt of the CCR message, the EBCF requests account information for the subscriber from the ABMF. Then the EBCF sends a Price Request message to

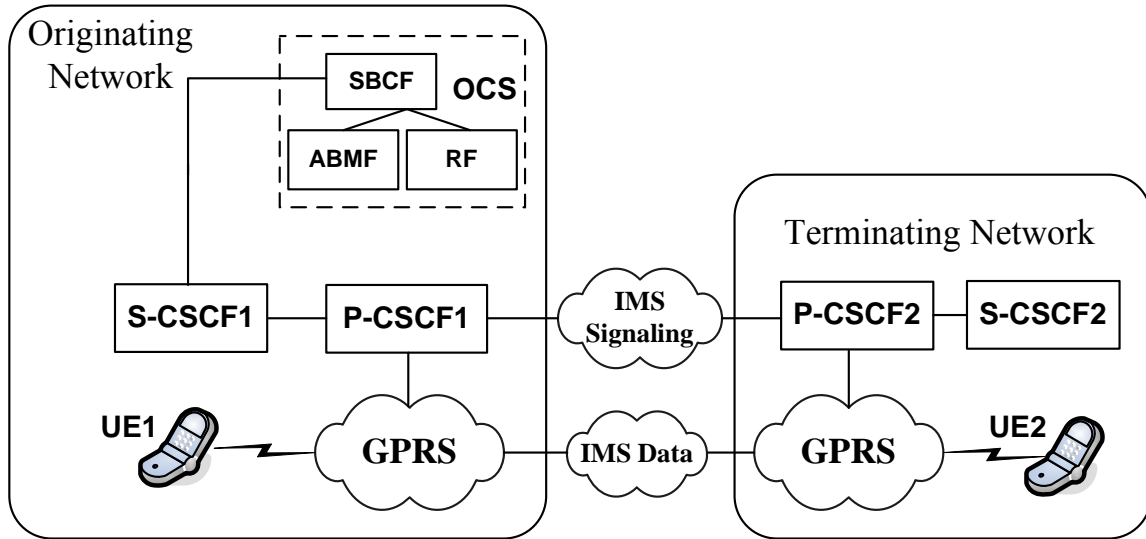


Figure 1.6: IMS Session with Online Charging.

the RF. This message includes the subscription identity and the service information. The RF then determines the price of the messaging service. Through the **Price Response** message, the calculated price and the billing information is returned to the EBCF.

Step 4. The EBCF performs credit unit reservation with the ABMF. Then it replies to the application server with the **CCR** message indicating the amount of granted credit.

Step 5. After the application server has obtained the credit, the messaging service is delivered to UE1, and then a SIP 200 OK message is sent to UE1 through the S-CSCF and the P-CSCF. If the credit control fails, an appropriate SIP error message (e.g., 401 Unauthorized or 402 Payment Required) is sent to UE1. Assume that the credit control succeeds, Step 6 is executed.

Step 6. After the message is successfully delivered, the application server sends the **CCR** message with *CC-Request-Type* "TERMINATE_REQUEST" to the EBCF. The EBCF debits the consumed credit from the subscriber's account. Then the OCS replies to the application server with the **CCA** message.

1.5.3 Session Charging with Unit Reservation

Session charging with unit reservation is performed in credit control for session-based services. Consider an online charging scenario in Fig. 1.6, where UE1 makes an IMS call

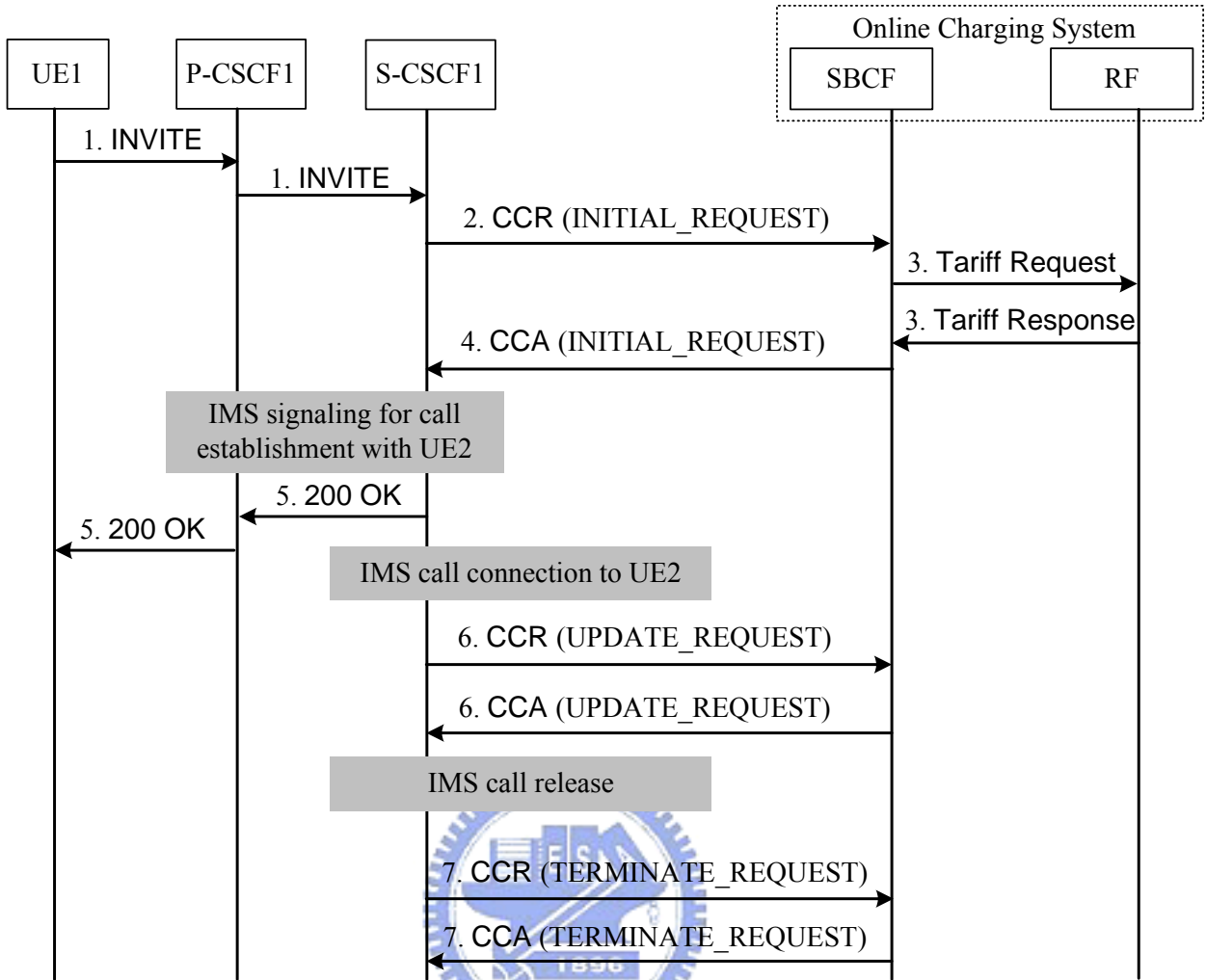


Figure 1.7: IMS Message Flow for Session Charging with Unit Reservation.

to UE2. We assume that the tariff information for this service is not changed during the session. In Chapter 4, we will consider the scenario where the tariff information is expired during a service session for reasons such as QoS change. The IMS message flow for session charging with unit reservation (see Fig. 1.7) is described as follows:

Step 1. Subscriber UE1 sends a SIP INVITE request to S-CSCF1 through P-CSCF1.

Step 2. S-CSCF1 sends a CCR message with *CC-Request-Type* “INITIAL_REQUEST” to the SBCF of the OCS. The *Service-Information AVP* contains the service parameters for the IMS sessions and the *Session Description Protocol (SDP)* parameters for the voice packets [6, 31].

Step 3. Upon receipt of the CCR message, the SBCF retrieves the account information

and the subscribed profile from the ABMF. Then the SBCF sends a **Tariff Request** message to the RF to determine the tariff of the IMS call. Based on the subscriber information, the RF replies the **Tariff Response** message to the SBCF. This message includes the billing plan and the tariff information for the IMS service.

Step 4. When the tariff information is received, the SBCF performs credit unit reservation with the ABMF. Then it replies to S-CSCF1 with the **CCA** message containing the granted credit (e.g., the number of minutes or bytes allowed). In this example, the request is successfully processed with *Result-Code* “DIAMETER_SUCCESS”.

Step 5. After S-CSCF1 has obtained the granted credit units, it continues the call setup to UE2 (through P-CSCF2 and S-CSCF2). When UE2 accepts the call, a SIP 200 OK message is sent to UE1 through S-CSCF1 and P-CSCF1. At this point, the IMS session starts.

Step 6. During the service session, S-CSCF1 supervises the network resource consumption by deducting the granted credit units. If the granted credit is depleted, the S-CSCF1 sends another **CCR** message with *CC-Request-Type* “UPDATE_REQUEST” to the SBCF. Through this message, S-CSCF1 also reports the amount of used credit, and requests additional credit for the remaining session. The SBCF deducts the consumed credit and reserves extra credit from the ABMF. Then the SBCF acknowledges S-CSCF1 with the **CCA** message including the amount of the next reserved credit. Note that this step may repeat several times before a service session is complete.

Step 7. When the service session is complete, S-CSCF1 sends another **CCR** message with *CC-Request-Type* “TERMINATE_REQUEST” to the SBCF. The SBCF debits the consumed credit units from the subscriber’s account with the ABMF. Then the SBCF sends the **CCA** message to S-CSCF1.

1.6 Dissertation Organization

Based on the above discussion, we present the performance analysis issues of the mobile online charging system and the dissertation organization. This dissertation contains five chapters in addition to this introductory chapter. Details for each chapter are described below.

In Chapter 2, we develop an IMS prepaid application server to handle both the prepaid calls and messaging services in UMTS/IMS environment. When both voice and messaging are simultaneously offered, a potential problem is that the delivery of a prepaid message during a call may result in force-termination of that call due to credit depletion. To address this issue, we describe a strategy to determine if a prepaid message can be sent out during a call session. We propose an analytic model to investigate the performance of this strategy.

In Chapter 3, we study the online credit reservation procedure for prepaid users in the UMTS online charging system. When the remaining amount of prepaid credit is below a threshold, the OCS should remind the user to refill the prepaid account. It is essential to choose an appropriate recharge threshold not only to avoid forced termination for the in-progress service sessions but also to decrease the recharge signaling overhead. An analytic model is developed to investigate the performance of this threshold-based recharge mechanism.

Chapter 4 focuses on the online credit re-authorization procedure of the OCS. The main objective is to reduce the traffic signaling overhead for credit re-authorization due to frequency QoS changes. In the OCS architecture described in Section 1.3, the ABMF and the SBCF may physically reside at different (and possibly remote) locations. Therefore, the message exchanges in the credit re-authorization may be expensive, and it is desirable to reduce the signaling overhead. We propose a threshold-based scheme that can significantly reduce the number of ABMF message exchanges during a session. An analytic model is developed to investigate the performance of this scheme.

We also conduct simulation experiments to validate the analytic models proposed in Chapters 2-4. Finally, Chapter 5 concludes this dissertation and gives the future directions of this work.

Chapter 2

Design of IMS Prepaid Application Server for SIP-based Services

The *Short Message Service* (SMS), which allows mobile users to send and receive simple text messages up to 140 bytes, is a mature wireless communication service. Most modern digital cellular phone systems offer SMS, which is considered a profitable value-added service [14, 41, 43]. In UMTS, *Multimedia Messaging Service* (MMS) is introduced to deliver messages that range in size from 30K bytes to 100K bytes [2, 13]. Formats that can be embedded within MMS include text (formatted with fonts, colors, and other style elements), images (in JPEG or GIF formats), audio (in MP3 or MIDI formats) and video (in MPEG format). With the growing demand of instant messaging services, RFC 3428 [22] proposes the MESSAGE method, a SIP extension that allows transfer of instant messages over the Internet. With this extension, Internet services (such as mail and instant messaging [44]) can be integrated with SMS/MMS through the IMS.

Billing mechanisms for messaging (including both SMS and MMS) and VoIP (especially for IMS calls toward the PSTN) are typically deployed for postpaid services. On the contrary, the prepaid billing mechanisms for combining these services are seldom studied in the literature. This chapter proposes a prepaid application server approach that can simultaneously process prepaid charging for SMS/MMS message deliveries and IMS calls. When both voice and messaging are simultaneously offered in the prepaid application server, a potential problem is that the delivery of a message during a call may result in force-termination of that call due to credit depletion. To address this issue, we describe a strategy to determine if a prepaid message can be sent out during a call session. An analytic model is derived to investigate the performance of this strategy. Based on our study [53], we provide guidelines to select appropriate input parameters for the prepaid

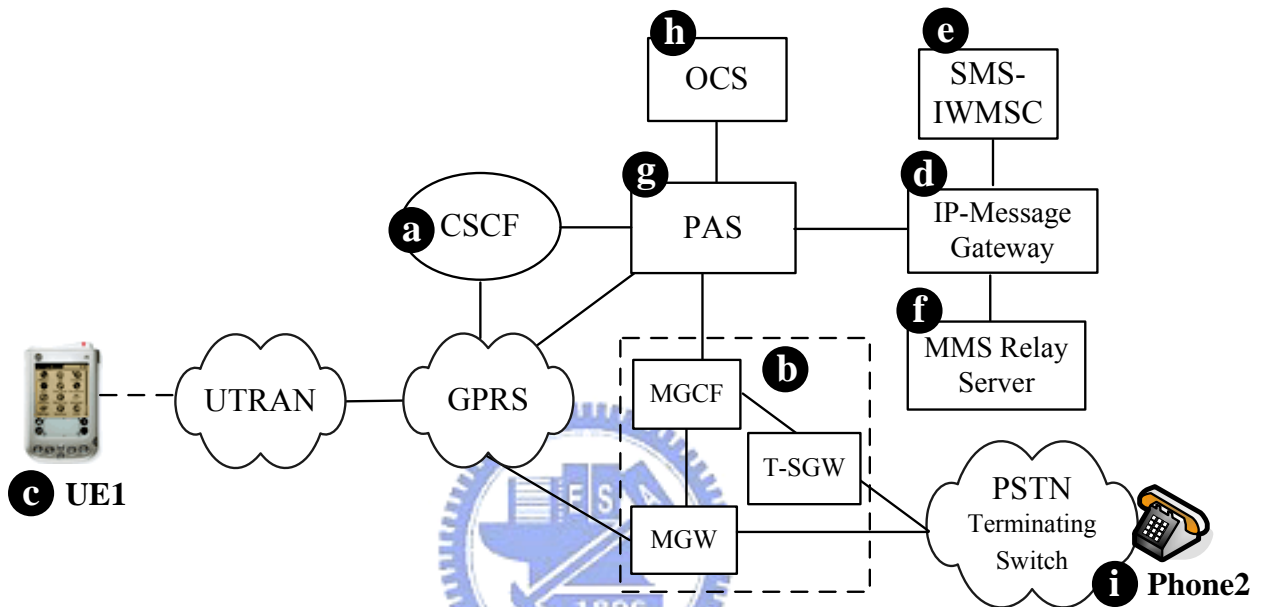
application server. The notation used in this chapter is listed in Section 2.7.

2.1 Prepaid Application Server of SIP-based Services

In the IMS architecture described in Section 1.2, the CSCF is basically a SIP server responsible for call control. Other IMS nodes for VoIP include the MGCF, the MGW and the T-SGW. These IMS nodes interact with each other through SIP [47, 18] signaling. SIP-based prepaid services in the Internet have been recently investigated (see [54, 55] and the references therein). In 3GPP Release 6, the *Online Charging System* (OCS) supports prepaid services for IMS. However, how to manage and allocates credit for the SIP-based prepaid services is not mentioned in the 3GPP specifications. Therefore a network entity is required to provide these functionalities without modifying the existing network entities such as the OCS and the CSCF. This section proposes a *Prepaid Application Server* (PAS) to manage credit allocation for both the prepaid SIP calls and messaging services in IMS. In Chapters 3 and 4, we will also consider credit reservation in various scenarios.

Fig. 2.1 illustrates the PAS that interacts with several network nodes in the IMS network: The CSCF (Fig. 2.1 (a)) utilizes SIP to provide control signaling for IP-based multimedia services. The MGCF, the MGW and the T-SGW (Fig. 2.1 (b)) interwork the IMS with the PSTN. The *User Equipment* (UE; Fig. 2.1 (c)) includes the SIP user agent to support IMS calls and SIP-based instant messaging services. The IP-Message Gateway (Fig. 2.1 (d)) translates the SIP-based instant messages into short/multimedia messages. The *Short Message Service - Interworking Mobile Switching Center* (SMS-IWMSC; see Fig. 2.1 (e)) sends/receives SMS to/from the IP-Message Gateway using standard SS7 *Mobile Application Part* (MAP) signaling [14], while the MMS Relay/Server (see Fig. 2.1 (f)) sends/receives MMS to/from the IP-Message Gateway. By exchanging Diameter *Credit Control Request* (CCR) and *Credit Control Answer* (CCA) messages, the PAS (Fig. 2.1 (g)) interacts with the OCS (Fig. Fig. 2.1 (h)) to reserve the amount of prepaid credit and process the online charging information. The PAS sets up prepaid IMS-to-PSTN calls through the MGCF, the MGW and the T-SGW, and supports prepaid SMS/MMS services through interaction with the IP-Message Gateway.

The prepaid call function in PAS works as follows: By utilizing the *Back-to-Back User Agent* (B2BUA) technique [55], the prepaid charging mechanism is inserted into a SIP connection session by breaking it into two sub-sessions. In this way, the PAS can monitor and terminate the call session when the prepaid credit is depleted. To set up an IMS-to-



CSCF: Call Session Control Function
 IWMSC: Interworking Mobile Switching Center
 MGCF: Media Gateway Control Function
 OCS: Online Charging System
 PSTN: Public Switched Telephone Network
 T-SGW: Transport Signaling Gateway
 UTRAN: UMTS Terrestrial Radio Access Network

GPRS: General Packet Radio Service
 MGW: Media Gateway
 MMS: Multimedia Messaging Service
 PAS: Prepaid Application Server
 SMS: Short Message Service
 UE: User Equipment

Figure 2.1: IMS Environment for SIP-based Prepaid Services.

PSTN call for an IMS user UE1, the signaling message is first routed from UE1 (Fig. 2.1 (c)) to the CSCF (Fig. 2.1 (a)). The CSCF identifies the charging type of the service (i.e., prepaid or postpaid). For a postpaid SIP request, the call is set up by standard IMS procedures. For a prepaid SIP request, the CSCF forwards the request to the PAS (Fig. 2.1 (g)) for further authorization and call session control. The details are given in the next section.

2.2 Integrating Charging for Prepaid Calls and Instant Messaging

This section first describes prepaid IMS-to-PSTN call setup and release. Then we describe instant message delivery through the prepaid mechanism. Finally, we describe the PAS charging policy for integrating IMS-to-PSTN calls and instant messages.

2.2.1 Prepaid IMS-to-PSTN Call Setup and Release

Consider the scenario where an IMS user (UE1; Fig. 2.1 (c)) makes a VoIP call to a PSTN user (Phone2 in Fig. 2.1 (i)). In this case, the PAS (i.e., the prepaid B2BUA) breaks the SIP session between UE1 and the MGCF into one subsession (subsession 1) between UE1 and the PAS (Fig. 2.1 (g)) and another subsession (subsession 2) between the PAS and the MGCF (Fig. 2.1 (b)).

In Fig. 2.1, the SIP INVITE request from UE1 (a prepaid user) is first routed to the PAS through the CSCF (path (1)→(2)→(3)→(4)) to establish subsession 1. Then the PAS interacts with the OCS to reserve the subscriber's credit (path (5)). When the user's authorized time (i.e., the prepaid credit) is granted, the PAS generates a new INVITE message for subsession 2, and then sends it to the MGCF through path (6). The MGCF instructs the T-SGW to deliver the call setup request to Phone2 via path (7). When the called party answers, the answer message is sent to the PAS through path (7)→(6). Then the PAS responds SIP 200 OK to UE1 through path (4)→(3)→(2)→(1). Finally, the PAS starts an authorized session timer with the value (the available prepaid credit) granted from the OCS. UE1 then sends the ACK message (for subsession 1) to the PAS through path (1)→(2)→(8); note that the ACK message need not be routed through the CSCF. The PAS sends the ACK message (for subsession 2) to the MGCF through path (6). At this point, the MGW opens a *Real-time Transport Protocol* (RTP) [50] connection so that UE1 can communicate with Phone2 through the MGW. The media path for the prepaid

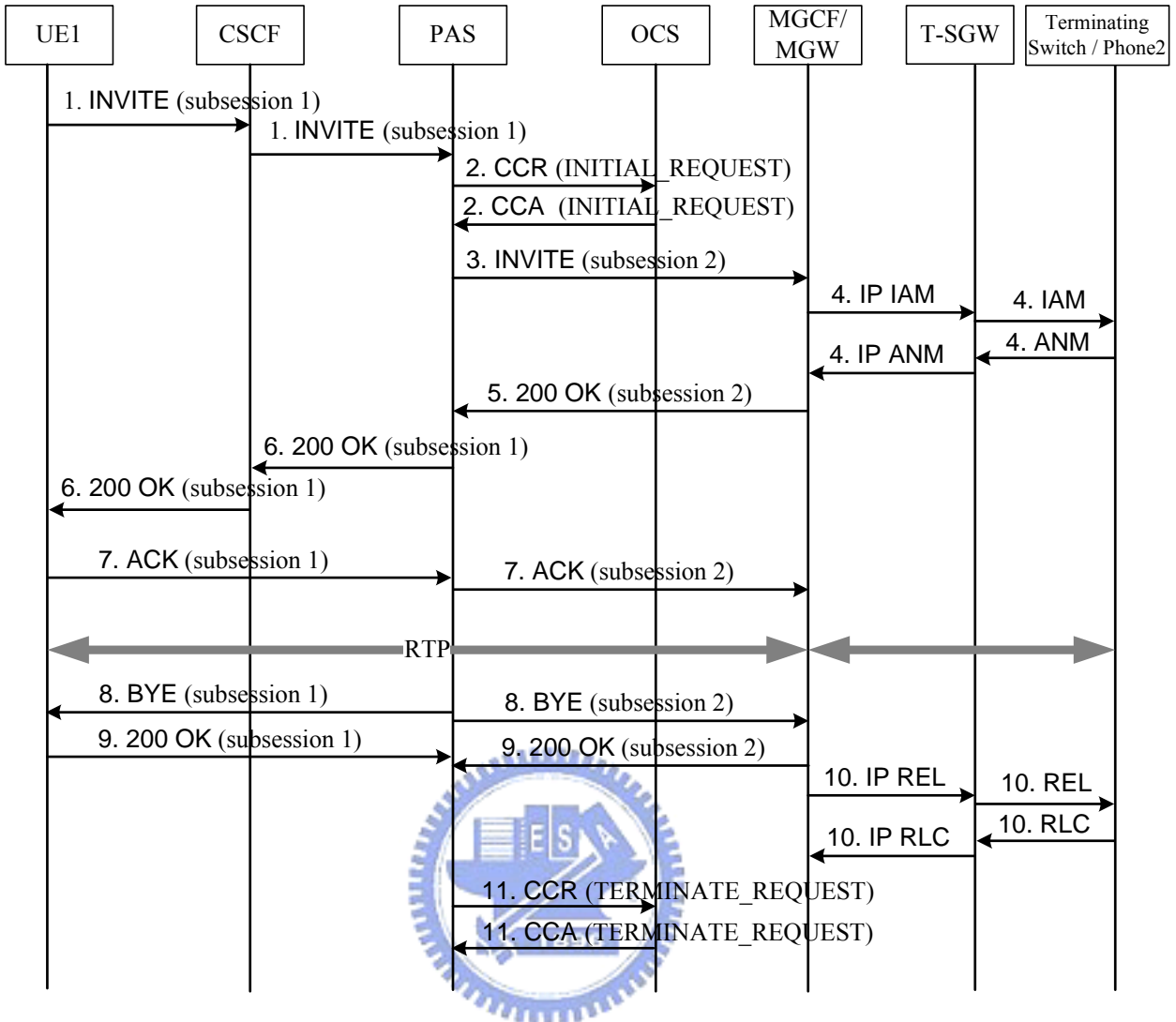


Figure 2.2: Message Flow for Prepaid IMS-to-PSTN Call Setup and Release.

call is (1)→(2)→(9)→(10). If the prepaid credit is exhausted (i.e., the authorized session timer expires) before the conversation is complete, the call is forced to terminate by the PAS. When the authorized session timer expires, the PAS sends the BYE messages to both UE1 (path (8)→(2)→(1)) and Phone2 through the MGCF (path (6)). Specifically, the MGCF instructs the MGW to release the RTP connection, and the subsequent voice packets delivered between UE1 and Phone2 are not allowed to pass through the MGW. The MGCF instructs the T-SGW to ask Phone2 to release the call via path (7). Finally, the PAS triggers the OCS to debit the user account through path (5). The message flow for the prepaid call setup and force-termination is described below (see Fig. 2.2).

Step 1. The SIP INVITE request from UE1 is first routed to the PAS through the CSCF.

- Step 2.** The PAS sends a CCR message with *CC-Request-Type* “INITIAL_REQUEST” to the OCS to reserve the subscriber’s credit. The OCS replies to the PAS with the CCA message containing the credit information (i.e., the user’s authorized time).
- Step 3.** When the user’s authorized time quota is granted, the PAS generates a new INVITE message for subsession 2, and then forwards it to the MGCF.
- Step 4.** The MGCF instructs the T-SGW to deliver the SS7 ISUP **Initial Address Message** (IAM; the SS7 call setup message [36]) to the terminating switch of Phone2. When the called party answers, the SS7 ISUP **Answer Message** (ANM) is sent to the T-SGW. Then the T-SGW sends an IP-based ANM message to the MGCF.
- Steps 5 and 6.** The MGCF sends a final response 200 OK to the PAS; the PAS sends a final response 200 OK to UE1.
- Step 7.** UE1 sends the ACK message (for subsession 1) to the PAS; the PAS sends the ACK message (for subsession 2) to the MGCF. At this point, the MGW opens a RTP connection so that UE1 and Phone 2 can deliver voice packets through the MGW. Finally, the PAS starts an authorized session timer with the value (the available prepaid credit) granted from the OCS.
- Step 8.** If the prepaid credit is exhausted before the conversation is complete (i.e., the authorized session timer expires), the call is forced to terminate by the PAS. The PAS will send the BYE messages to both UE1 and the MGCF. Then the MGCF instructs the MGW to release the RTP connection, and the subsequent voice packets delivered between UE1 and Phone2 are not allowed to pass through the MGW.
- Step 9.** UE1 and the MGCF reply to the PAS with 200 OK messages.
- Step 10.** The MGCF instructs the T-SGW to send the SS7 ISUP **Release Message** (REL) to the terminating switch of Phone2. The terminating switch replies to the T-SGW with the SS7 ISUP **Release Complete** (RLC). Then the T-SGW sends an IP-based RLC message to the MGCF.
- Step 11.** Finally, the PAS sends a CCR message to the OCS. The *CC-Request-Type* of the message is “TERMINATE_REQUEST”. The OCS debits the user account, and then replies the PAS with a CCA message.

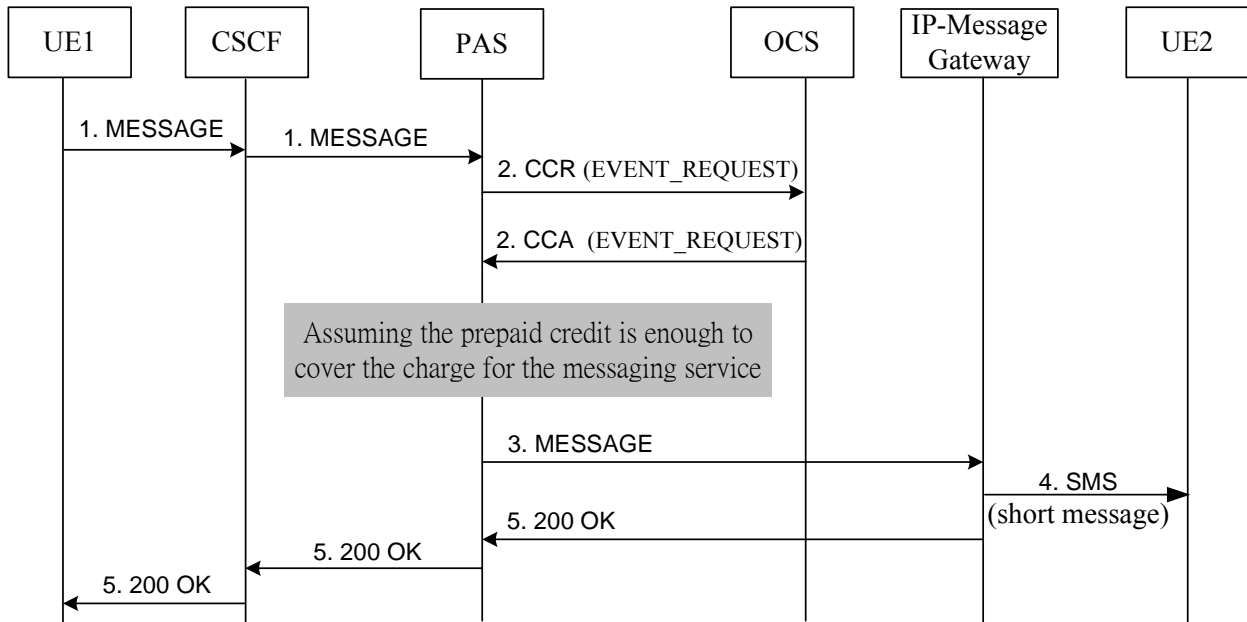


Figure 2.3: Message Flow for Prepaid Messaging Delivery.

2.2.2 Prepaid Instant Messaging Delivery

The message flow for prepaid instant messaging service is shown in Fig. 2.3, and is described in the following steps:

Step 1. A prepaid user UE1 sends the SIP MESSAGE request to the PAS through the CSCF (including P-CSCF and S-CSCF). The instant message to be delivered is attached in this SIP request.

Step 2. The PAS sends a CCR message to the OCS. The *CC-Request-Type* is “EVENT_REQUEST”. The OCS debits the subscriber’s prepaid account, and then replies to the PAS with the CCA message containing the cost information (i.e., the amount of credit charged for this service).

Step 3. If the amount of prepaid credit in UE1’s account is large enough to cover the charge for the messaging service, the PAS forwards the MESSAGE request to the IP-Message Gateway.

Step 4. The IP-Message Gateway retrieves the instant message content from the MESSAGE request and delivers it to UE2 through the SMS/MMS network.

Step 5. The IP-Message Gateway sends the SIP 200 OK message to UE1 through the PAS and the CSCF.

2.3 Analytic Modeling for Prepaid Application Server

The PAS for pure voice services or pure message services can be easily implemented as described in Section 2.2. However, when both voice and messaging are simultaneously offered, an important issue must be resolved for the PAS. That is, during a prepaid IMS-to-PSTN call, the user may attempt to send messages. Deduction of prepaid credit for sending out this message may result in insufficient credit left for the in-progress prepaid call. (Note that a message can be a short text or a huge multimedia data file which causes large deduction.) In this case, the IMS-to-PSTN call is forced to terminate. Therefore, a strategy is required to determine if the prepaid message can be sent out without causing force-termination of an on-going call. To avoid unnecessary force-termination, we set a threshold amount X_T of credit units to protect the in-progress IMS-to-PSTN call. Consider the timing diagram in Fig. 2.4. A prepaid call arrives at t_1 and completes at t_3 . Without loss of generality, we assume that each time unit (TU) of the call is charged for 1 credit unit (CU). A prepaid authorized timer for this call starts at t_1 and expires at t_4 . That is, the amount of the prepaid credit is $t_4 - t_1$. Upon receipt of the prepaid message request at t_2 , where $t_1 < t_2 < t_3$, the PAS estimates if the remaining credit $x^* = t_4 - t_2$ suffices to support both the remaining call and the message service. Assume that the prepaid message service is charged for a fixed amount of T_m credit units. The strategy used in our PAS is described as follows:

- If the remaining prepaid credit $x^* \geq X_T + T_m$, the message is sent immediately and the amount of remaining prepaid credits is reduced to $x^* - T_m$.
- If not, the message is stored and will be processed after the IMS-to-PSTN call is completed.

If X_T is set too large, the PAS may reserve too many prepaid credit units, and therefore message deliveries are unnecessarily delayed. If X_T is set too small, the PAS may reserve too few prepaid credit units for the remaining prepaid call, and therefore results in unnecessary force-terminated (UFT) calls. We present an analytic model to derive the UFT (unnecessary force-termination) probability P_{UFT} and the expected unnecessary delay $E[t_d]$ with the following assumptions:

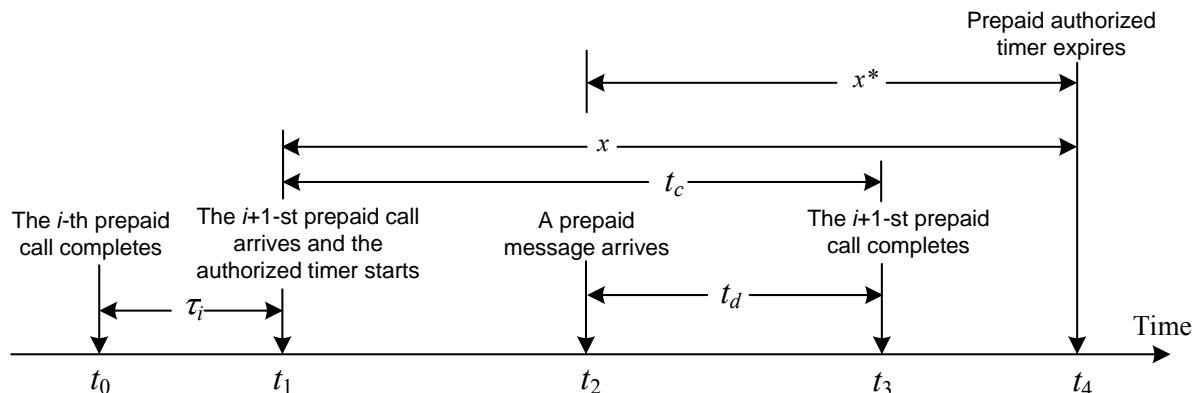


Figure 2.4: Timing Diagram for Prepaid Calls and Messaging Deliveries.

1. Let X be the amount of the initial prepaid credit.
2. The prepaid message arrivals are a Poisson stream with rate λ_m . Let $N(t)$ be the number of prepaid messages occurring in period t . The probability mass function of $N(t)$ is

$$\Pr[N(t) = n] = \left[\frac{(\lambda_m t)^n}{n!} \right] e^{-\lambda_m t} \quad (2.1)$$

3. Let τ_i be the interval between when the i -th prepaid call completes and when the $i + 1$ -st prepaid call starts ($i > 0$). In Fig. 2.4, $\tau_i = t_1 - t_0$. We define τ_0 as the interval between when the prepaid account is first activated and when the first prepaid call arrives. Let τ_i be Exponentially distributed with the mean $1/\lambda_c$.
4. The prepaid call holding time t_c (in Fig. 2.4, $t_c = t_3 - t_1$) is Exponentially distributed with the density function

$$f_c(t_c) = \mu e^{-\mu t_c} \quad (2.2)$$

2.3.1 Derivation for the Remaining Credit Density Function

Consider the timing diagram in Fig. 2.4, where a prepaid call arrives at t_1 and the remaining credit left is $x = t_4 - t_1$. To derive P_{UFT} and $E[t_d]$, we first derive the density function $f_x(x)$ of x . Let N_c be the number of prepaid calls completed during $[0, t_1]$. Let X_c be the amount of credit units that are consumed by the prepaid calls arrived before t_1 . Since we assume that 1 TU of the call is charged for 1 CU, the amount of the consumed prepaid credit (i.e., the accumulated call holding times) for these N_c prepaid calls is X_c .

For $N_c > 0$, X_c has the Erlang distribution with the mean N_c/μ and the shape parameter N_c . The conditional probability mass function of N_c given that $X_c = x_c$ can be expressed as

$$\Pr[N_c = n_c | X_c = x_c] = \frac{\left[\frac{(\mu x_c)^{n_c-1}}{(n_c-1)!} \right] \mu e^{-\mu x_c}}{\sum_{i=1}^{\infty} \left[\frac{(\mu x_c)^{i-1}}{(i-1)!} \right] \mu e^{-\mu x_c}} = \frac{e^{-\mu x_c} (\mu x_c)^{n_c-1}}{(n_c-1)!} \quad (2.3)$$

Let $T_i = \sum_{i=0}^{N_c} \tau_i$. Since τ_i has the Exponential distribution with the mean $1/\lambda_c$, T_i has the Erlang distribution with the mean $(N_c+1)/\lambda_c$ and the shape parameter N_c+1 . Therefore the density function $f_T(T_i)$ of T_i is

$$f_T(T_i) = \left[\frac{\lambda_c (\lambda_c T_i)^{N_c}}{N_c!} \right] e^{-\lambda_c T_i} \quad (2.4)$$

Let N_m be the number of prepaid messages arrived before t_1 . In other words, there are N_m prepaid messages occurring in period $X_c + T_i$. Then N_m is a Poisson random variable with the mean $\lambda_m(X_c + T_i)$. From (2.1), the probability mass function of N_m can be expressed as

$$\Pr[N_m = n_m] = \Pr[N(X_c + T_i) = n_m] = \left\{ \frac{[\lambda_m(X_c + T_i)]^{n_m}}{n_m!} \right\} e^{-\lambda_m(X_c + T_i)} \quad (2.5)$$

Note that $N(t)$ (see Eq. (2.1)) represents the number of prepaid messages occurring in any arbitrary period t . Therefore $N_m = N(X_c + T_i)$ (see Eq. (2.5)) represents the number of prepaid messages occurring before the prepaid call arriving at t_1 . From (2.3)-(2.5), the conditional probability mass function of N_m given that $X_c = x_c$ is derived as

$$\begin{aligned} & \Pr[N_m = n_m | X_c = x_c] \\ &= \sum_{n_c=1}^{\infty} \Pr[N_c = n_c | X_c = x_c] \int_{t_i=0}^{\infty} \Pr[N_m = n_m | X_c = x_c, N_c = n_c, T_i = t_i] f_T(t_i) dt_i \\ &= \sum_{n_c=1}^{\infty} \frac{e^{-\mu x_c} (\mu x_c)^{n_c-1}}{(n_c-1)!} \int_{t_i=0}^{\infty} \left\{ \frac{[\lambda_m(x_c + t_i)]^{n_m} e^{-\lambda_m(x_c + t_i)}}{n_m!} \right\} \left[\frac{(\lambda_c t_i)^{n_c} \lambda_c e^{-\lambda_c t_i}}{n_c!} \right] dt_i \\ &= \left[\frac{\lambda_c \lambda_m^{n_m} e^{-(\mu + \lambda_m)x_c}}{\mu} \right] \int_{t_i=0}^{\infty} e^{-(\lambda_c + \lambda_m)t_i} \sum_{k=0}^{n_m} \frac{t_i^k x_c^{n_m-k-1}}{k!(n_m-k)!} \sum_{n_c=1}^{\infty} \frac{(\mu \lambda_c x_c t_i)^{n_c}}{(n_c-1)! n_c!} dt_i \\ &= \left[\frac{\lambda_c \lambda_m^{n_m} e^{-(\mu + \lambda_m)x_c}}{\mu} \right] \sum_{k=0}^{n_m} \sum_{n_c=1}^{\infty} \left[\frac{(n_c + k)!}{(\lambda_c + \lambda_m)^{n_c+k+1}} \right] \left[\frac{x_c^{n_m-k-1}}{k!(n_m-k)!} \right] \left\{ \frac{(\mu \lambda_c x_c)^{n_c}}{(n_c-1)! n_c!} \right\} \\ &= \left[\frac{\lambda_c \lambda_m^{n_m} e^{-(\mu + \lambda_m)x_c}}{\mu} \right] \sum_{k=0}^{n_m} \left[\frac{x_c^{n_m-k-1}}{(\lambda_c + \lambda_m)^{k+1} k!(n_m-k)!} \right] \sum_{n_c=1}^{\infty} \left(\frac{\mu \lambda_c x_c}{\lambda_c + \lambda_m} \right)^{n_c} \\ & \times \left[\frac{(n_c + k) \dots (n_c + 1)}{(n_c - 1)!} \right] \end{aligned} \quad (2.6)$$

Let $Z = \mu\lambda_c x_c / (\lambda_c + \lambda_m)$. Since $\frac{d}{dZ} (Z^{n_c+k})^{(k)} = (n_c+k)\dots(n_c+1)Z^{n_c}$. Then (2.6) can be re-written as

$$\begin{aligned} & \Pr[N_m = n_m | X_c = x_c] \\ &= \left[\frac{\lambda_c \lambda_m^{n_m} e^{-(\mu+\lambda_m)x_c}}{\mu} \right] \sum_{k=0}^{n_m} \left[\frac{x_c^{n_m-k-1}}{(\lambda_c + \lambda_m)^{k+1} k! (n_m - k)!} \right] \sum_{n_c=1}^{\infty} \left[\frac{d}{dZ} \frac{(Z^{n_c+k})^{(k)}}{(n_c - 1)!} \right] \\ &= \left[\frac{\lambda_c \lambda_m^{n_m} e^{-(\mu+\lambda_m)x_c}}{\mu} \right] \sum_{k=0}^{n_m} \left[\frac{x_c^{n_m-k-1}}{(\lambda_c + \lambda_m)^{k+1} k! (n_m - k)!} \right] \frac{d}{dZ} (Z^{k+1} e^Z)^{(k)} \end{aligned} \quad (2.7)$$

According to the Leibniz's rule for higher derivatives of products, we can expand Eq. (2.7) as

$$\begin{aligned} & \Pr[N_m = n_m | X_c = x_c] \\ &= \left[\frac{\lambda_c \lambda_m^{n_m} e^{-(\mu+\lambda_m)x_c}}{\mu} \right] \sum_{k=0}^{n_m} \left[\frac{x_c^{n_m-k-1}}{(\lambda_c + \lambda_m)^{k+1} k! (n_m - k)!} \right] \sum_{l=0}^k \binom{k}{l} \frac{d}{dZ} (Z^{k+1})^{(l)} \frac{d}{dZ} (e^Z)^{(k-l)} \\ &= \left(\frac{\lambda_c \lambda_m^{n_m}}{\mu} \right) \sum_{k=0}^{n_m} \left[\frac{(k+1)!}{(\lambda_c + \lambda_m)^{k+1} (n_m - k)!} \right] \sum_{l=0}^k \left[\frac{(\frac{\mu\lambda_c}{\lambda_c + \lambda_m})^{k-l}}{l!(k-l)!(k-l+1)!} \right] \\ & \quad \times e^{-(\mu+\lambda_m - \frac{\mu\lambda_c}{\lambda_c + \lambda_m})x_c} x_c^{n_m-l} \end{aligned} \quad (2.8)$$

When $T_m = 0$, it is clear that $X_c > 0$ is uniformly distributed over $(0, X]$ with the density $1/X$. When $T_m > 0$ and $x > 0$, the number of prepaid messages that arrived before t_1 is at most $\lfloor \frac{X-X_c}{T_m} \rfloor$. From (2.8),

$$\begin{aligned} & \Pr[N_m = n_m | X_c = x_c, T_m > 0, x > 0] \\ &= \frac{\frac{1}{X} \Pr[N_m = n_m | X_c = x_c]}{\Pr[X > X_c + N_m T_m > 0]} \\ &= \frac{\frac{1}{X} \Pr[N_m = n_m | X_c = x_c]}{\frac{1}{X} \int_{t=0}^X \sum_{n=0}^{\lfloor \frac{X-t}{T_m} \rfloor} \Pr[N_m = n | X_c = t] dt} \\ &= \frac{1}{C} \Pr[N_m = n_m | X_c = x_c] \end{aligned} \quad (2.9)$$

where

$$\begin{aligned} C &= \int_{t=0}^X \sum_{n=0}^{\lfloor \frac{X-t}{T_m} \rfloor} \Pr[N_m = n | X_c = t] dt \\ &= \sum_{n=0}^{\lfloor \frac{X}{T_m} \rfloor} \int_{t=0}^{X-nT_m} \Pr[N_m = n | X_c = t] dt \\ &= \frac{\lambda_c}{\mu} \sum_{n=0}^{\lfloor \frac{X}{T_m} \rfloor} \lambda_m^n \left[\sum_{k=0}^n \frac{(k+1)!}{(\lambda_c + \lambda_m)^{k+1} (n-k)!} \right] \end{aligned}$$

$$\begin{aligned}
& \times \sum_{l=0}^k \left[\frac{\left(\frac{\mu\lambda_c}{\lambda_c+\lambda_m}\right)^{k-l} (n-l)!}{l!(k-l)!(k-l+1)! \left(\mu + \lambda_m - \frac{\mu\lambda_c}{\lambda_c+\lambda_m}\right)^{n-l+1}} \right] \\
& \times \left\{ 1 - e^{-(\mu+\lambda_m - \frac{\mu\lambda_c}{\lambda_c+\lambda_m})(X-iT_m)} \sum_{j=0}^{n-i} \frac{[(\mu + \lambda_m - \frac{\mu\lambda_c}{\lambda_c+\lambda_m})(X - iT_m)]^j}{j!} \right\} \quad (2.10)
\end{aligned}$$

In Fig. 2.4, suppose that there are N_m prepaid messages that are delivered before t_1 and assume that $X_c = X - x - N_m T_m$. When a prepaid call arrives at t_1 , it is clear that the remaining prepaid credit left is x . Therefore, from (2.9), the density function $f_x(x)$ of x can be derived as

$$f_x(x)dx = \sum_{n_m=0}^{\lfloor \frac{X-x}{T_m} \rfloor} \Pr[N_m = n_m | X_c = X - x - n_m T_m, T_m > 0, x > 0] \quad (2.11)$$

where $\Pr[N_m = n_m | X_c = x_c, T_m > 0, x > 0]$ can be obtained from (2.8), (2.9) and (2.10).

2.3.2 Derivation for the Unnecessary Force-termination Probability

Based on (2.11), we derive the output measure P_{UFT} for $X_T = 0$ (that is, a prepaid message is immediately served if $x^* > T_m$). In this case, a call is unnecessary force-terminated if $N(t_c) > \lfloor \frac{x-t_c}{T_m} \rfloor$. Based on (2.1), (2.2) and (2.11), if $X_c > 0$, the UFT probability P_{UFT} can be derived as

$$\begin{aligned}
P_{UFT} &= \int_{x=0}^X \int_{t_c=0}^x \Pr \left[N(t_c) > \left\lfloor \frac{x-t_c}{T_m} \right\rfloor \right] f_c(t_c) f_x(x) dt_c dx \\
&= \int_{x=0}^X \int_{t_c=0}^x \left[1 - \sum_{i=0}^{\lfloor \frac{x-t_c}{T_m} \rfloor} \frac{(\lambda_m t_c)^i e^{-\lambda_m t_c}}{i!} \right] \mu e^{-\mu t_c} dt_c f_x(x) dx \\
&= 1 - \int_{x=0}^X \left\{ e^{-\mu x} + \mu \sum_{i=0}^{\lfloor \frac{x}{T_m} \rfloor} \left\{ 1 - e^{-(\mu+\lambda_m)(x-iT_m)} \sum_{j=0}^i \frac{[(\mu + \lambda_m)(x - iT_m)]^j}{j!} \right\} \right\} \\
&\quad \times f_x(x) dx \quad (2.12)
\end{aligned}$$

where $f_x(x)$ is expressed in (2.11).

2.3.3 Derivation for the Unnecessary Delay

The output measure $E[t_d]$ for $X_T \geq X$ is derived as follows: A prepaid message that arrives during an in-progress call must be stored and will be processed after the prepaid

call is completed. Since the prepaid message arrivals are random observation points of the prepaid call holding interval [48], the delay for the message can be considered as the residual life of the prepaid call. Consider the situation that a prepaid call with the call holding time t_c arrives at t_1 , where $t_c = t_3 - t_1 \leq x - T_m$. If a prepaid message arrives at t_2 during interval $[t_1, t_3]$, the message is sent at t_3 with the delay $t_d = t_3 - t_2$. According to the residual life theorem [33] and for $X_c > 0$, the expected unnecessary delay $E[t_d]$ for the SMS delivery can be expressed as

$$\begin{aligned}
E[t_d] &= \frac{E[t_c^2 | x > t_c + T_m]}{2E[t_c | x > t_c + T_m]} \\
&= \frac{\int_{x=T_m}^X \int_{t_c=0}^{x-T_m} t_c^2 f_c(t_c) f_x(x) dt_c dx}{\int_{x=T_m}^X \int_{t_c=0}^{x-T_m} t_c f_c(t_c) f_x(x) dt_c dx} \\
&= \frac{X - T_m - \int_{x=T_m}^X e^{-\mu x} [1 + \mu x + \frac{1}{2}(\mu x)^2] f_x(x) dx}{\mu [X - T_m - \int_{x=T_m}^X e^{-\mu x} (1 + \mu x) f_x(x) dx]} \tag{2.13}
\end{aligned}$$

where $f_x(x)$ is expressed in (2.11).

2.4 Simulation Validation

We also developed a simulation model for the PAS. The simulation experiments are validated against the analytic model developed in this section. Base on (2.12) and (2.13), Table 2.1 lists the analytic and the simulation results for the unnecessary force-termination probability P_{UFT} and the expected unnecessary delay $E[t_d]$. Table 2.1 indicates that for all cases considered in our study, the errors are within 0.6%. Therefore, the analytic and the simulation results are consistent.

In this discrete event simulation model for the prepaid application server, an event e consists of two fields: **e.ts** is the timestamp of the event and **e.type** is the event type. There are three event types: **CALL_ARRIVAL** (a new prepaid call arrival), **CALL_COMPLETION** (a prepaid call completion) and **MSG_ARRIVAL** (a prepaid message arrival). These events are inserted into an event list, and are deleted/processed from the event list in the non-decreasing timestamp order. The output measures of the simulation are C (the number of call arrivals), N (the number of unnecessary force-terminated calls), M (the number of message deliveries which arrive during the in-progress calls), and T_d (the accumulated delay for these M messages). These output measures are used to compute P_{UFT} and $E[t_d]$ as follows:

$$P_{UFT} = N/C \quad \text{and} \quad E[t_d] = T_d/M$$

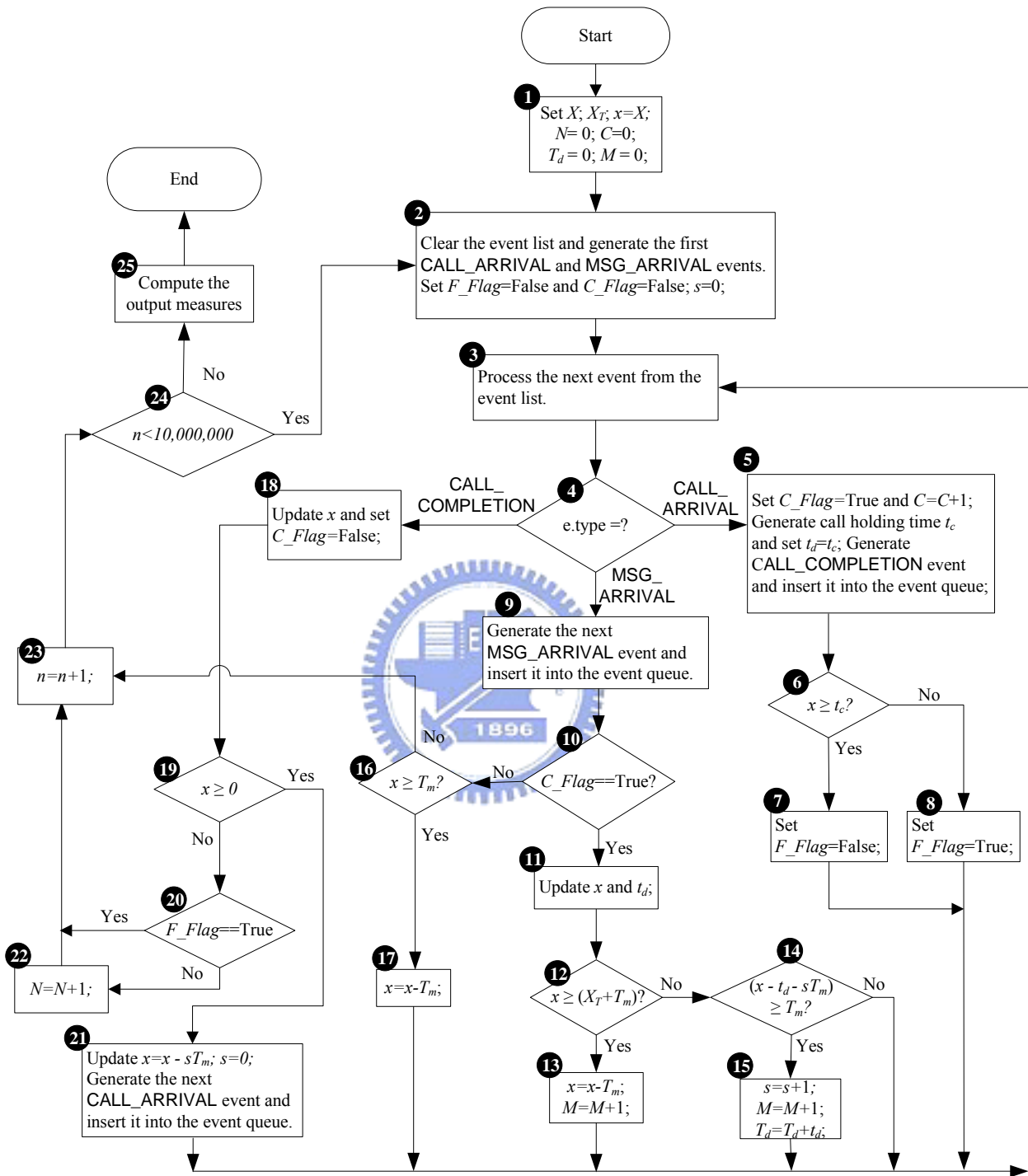


Figure 2.5: Simulation Flow Chart.

Table 2.1: Comparison of the Analytic and Simulation Results ($X_c > 0$, $T_m = 5$ CU, $X = 20T_m$, $1/\lambda_c = 50$ TU)

$1/\lambda_m$ (Unit: TU)	1	5	10	50	100
Simulation	43.086%	13.172%	6.409%	1.079%	0.520%
Analytic	42.960%	13.196%	6.419%	1.078%	0.520%
Error	-0.29%	0.18%	0.16%	-0.09%	-0.00%

(a) $P_{UFT} (X_T = 0, E[t_c] = \frac{1}{5\lambda_c})$

$E[t_c]$ (Unit: TU)	1	5	10	50	100
Simulation (Unit: TU)	0.990	4.678	8.520	18.128	20.056
Analytic (Unit: TU)	0.985	4.681	8.534	18.172	20.050
Error	-0.51%	0.06%	0.16%	0.24%	-0.03%

(b) $E[t_d] (X_T \geq X, \lambda_m = \lambda_c)$

To simulate the behavior of the prepaid account, several variables are maintained in the simulation. When a call is in-progress, C_Flag is set to “True”. If the remaining credit units x is larger than the call holding time t_c , then F_Flag is set to “False”. The period t_d represents the residual life of the call and s indicates the number of messages which are stored and are delivered when the call is completed.

The flowchart of the simulation is shown in Fig. 2.5. Step 1 initializes the input parameters (X, X_T, x, N, C, T_d and M) of a prepaid account. Step 2 generates the first **CALL_ARRIVAL** and **MSG_ARRIVAL** events, and inserts these events into the event list. Initially, F_Flag and C_Flag are set to “False” and s is set to 0. In Steps 3 and 4, the next event \mathbf{e} in the event list is processed based on its type.

If $\mathbf{e.type} == \mathbf{CALL_ARRIVAL}$ event at Step 4, then at step 5, C_Flag is set to “True” and C is incremented by one. Then the call holding time t_c for the prepaid call is generated. The residual life t_d is set to t_c . Step 5 also generates a **CALL_COMPLETION** event $\mathbf{e1}$, where $\mathbf{e1.ts} = \mathbf{e.ts} + t_c$. Event $\mathbf{e1}$ is inserted into the event list. At Step 6, if the remaining CUs x is larger than t_c , then F_Flag is set to “False” (Step7). Otherwise, F_Flag is set to “True” (Step 8).

If $\mathbf{e.type} == \mathbf{MSG_ARRIVAL}$ at Step 4, the next **MSG_ARRIVAL** event is generated and inserted into the event list at Step 9. If a prepaid call is in-progress (i.e., $C_Flag = \mathbf{True}$ at Step 10), it first updates the values of the remaining CUs x and the

residual call time t_d . At Step 12, if the remaining CUs $x \geq (X_T + T_m)$, then at Step 13 the message is sent immediately and x is reduced to $x - T_m$. The number M of message deliveries is incremented by one. If $x < (X_T + T_m)$ at Step 12, then Step 14 checks if the remaining credits x suffices to support the message delivery when the call is completed (i.e., $(x - t_d - sT_m) \geq T_m$). If so, at Step 15 the message is stored and will be processed when the call is completed. Both s and M are incremented by one. The accumulated delivery delay T_d is increased by t_d . If no prepaid call is in-progress (i.e., $C_Flag = \text{“False”}$) at Step 10, the simulation proceeds to Step 16. If the remaining credits x is larger than T_m , the message is sent and x is reduced to $x = x - T_m$ (Step 17). If $x < T_m$ at Step 16, all CUs are consumed and n is incremented by one at Step 23, and the simulation proceeds to Step 24.

If **e.type == CALL_COMPLETION** at Step 4, then the simulation proceeds to Step 18. This step first updates the remaining CUs x and sets C_Flag to “False”. At Step 19, if the remaining CUs are not depleted (i.e., $x \geq 0$), the stored messages can be delivered and x is updated as $x - sT_m$. The number s of stored messages is set to 0, the next **CALL_ARRIVAL** event is generated and inserted into the event list. If the remaining CUs is insufficient (i.e., $x < 0$ at Step 19) and the the call need not be force-terminated (i.e., $F_Flag = \text{“False”}$ at Step 20), then the number N of unnecessary false-terminated calls is incremented by one (Step 22), and the simulation proceeds to Step 23. If $F_Flag = \text{“True”}$ at Step 20, the simulation also proceeds to Step 23. At Step 23, all CUs are consumed, n is incremented by one and the simulation proceeds to Step 24.

If 10,000,000 prepaid accounts have been processed at Step 24, the simulation terminates and the output measures are calculated.

2.5 Numerical Examples

This section uses the simulation experiments to investigate the performance of the PAS. The input parameter threshold X_T and the output measure $E[t_d]$ are normalized by the mean of the call holding time $E[t_c] = 1/\mu$. For the purposes of demonstration, we assume that the prepaid message service is charged for $T_m = 5$ CUs and the expected inter-call arrival time $1/\lambda_c = 50$ TUs.

Effects of the expected call holding time $E[t_c]$. Fig. 2.6 plots P_{UFT} and $E[t_d]$ against the threshold X_T and $E[t_c]$, where $X = 50E[t_c]$ and $\lambda_m = 5\lambda_c$. This figure shows

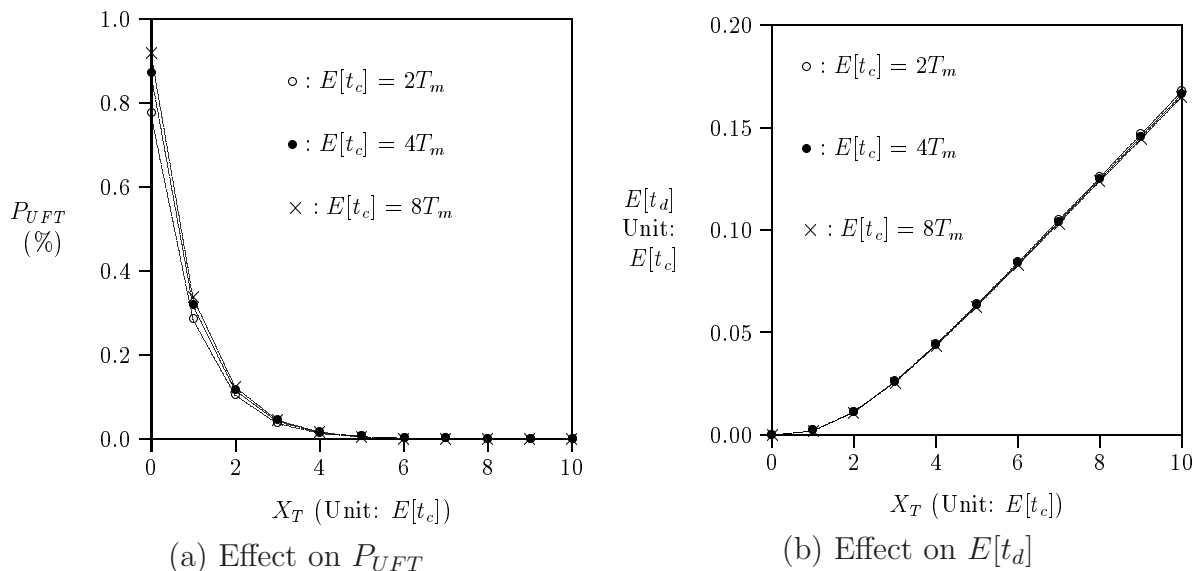


Figure 2.6: Effects of $E[t_c]$ on P_{UFT} and $E[t_d]$ ($X = 50E[t_c]$ and $\lambda_m = 5\lambda_c$)

that $E[t_c]$ has insignificantly effects on P_{UFT} and $E[t_d]$ when the proportion of X and $E[t_c]$ is fixed.

In the remainder of this section, the expected inter-message arrival time $1/\lambda_m$ and the initial prepaid credit X are normalized by $E[t_c] = 4T_m$.

Effects of the variance V_c of the call holding time t_c . Fig. 2.7 plots P_{UFT} and $E[t_d]$ against X_T and the variance V_c of the call holding time t_c with Gamma distribution, where $X = 25E[t_c]$ and $1/\lambda_m = 0.5E[t_c]$. This figure shows that both P_{UFT} and $E[t_d]$ increase as V_c increases. This phenomenon is explained as follows: As the variance V_c of t_c increases, more long and short t_c periods are observed. The prepaid message (i.e., the random observer) is more likely to fall in the long t_c periods than the short t_c periods, and larger residual call holding times t_d are expected for larger variance V_c . Therefore, the performance of both P_{UFT} and $E[t_d]$ degrade as V_c increases.

Effects of the expected inter-message arrival time $1/\lambda_m$. Fig. 2.8 plots P_{UFT} and $E[t_d]$ against X_T and λ_m , where $X = 25E[t_c]$. Fig. 2.8 (a) shows that P_{UFT} increases as λ_m increases. When λ_m increases, more message deliveries are likely to occur during an in-progress call. Therefore the UFT probability P_{UFT} increases. For $X_T = 2E[t_c]$, when $1/\lambda_m$ decreases from $50E[t_c]$ to $5E[t_c]$ and from $5E[t_c]$ to $0.5E[t_c]$, P_{UFT} increases by 9.13 and 9.01 times, respectively. This effect becomes insignificant

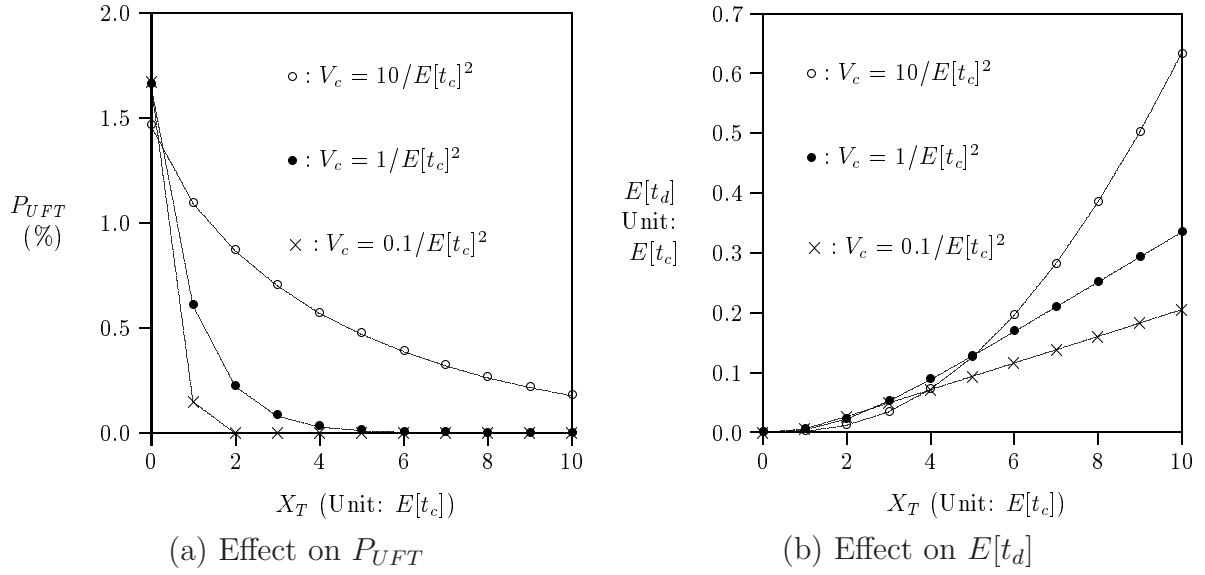


Figure 2.7: Effects of V_c on P_{UFT} and $E[t_d]$ ($X = 25E[t_c]$ and $1/\lambda_m = 0.5E[t_c]$)

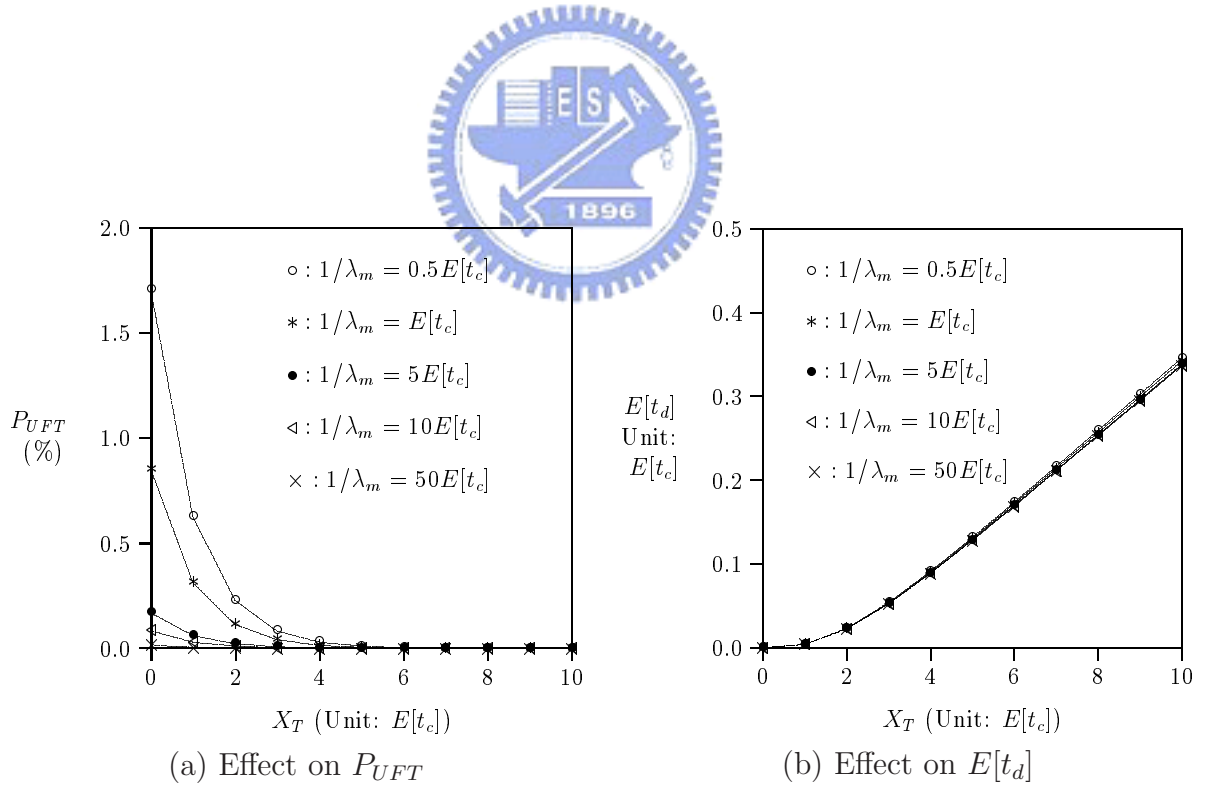


Figure 2.8: Effects of $1/\lambda_m$ on P_{UFT} and $E[t_d]$ ($X = 25E[t_c]$)

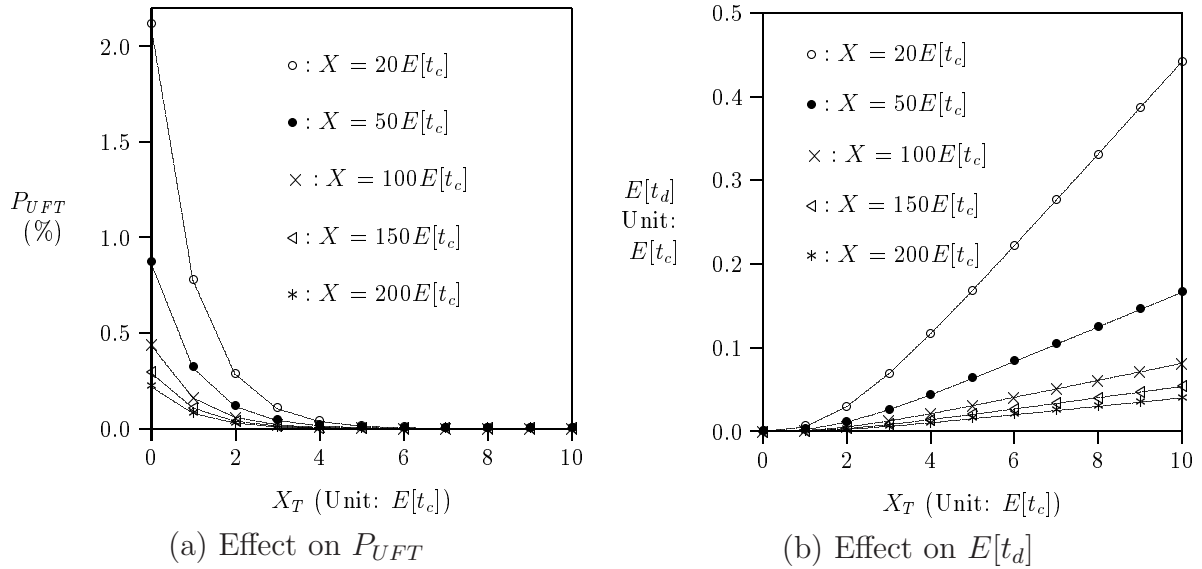


Figure 2.9: Effects of X on P_{UFT} and $E[t_d]$ ($1/\lambda_m = 0.5E[t_c]$)

when X_T is large (e.g., $X_T \geq 6E[t_c]$). P_{UFT} is not significantly affected by λ_m when $1/\lambda_m > 10E[t_c]$. Fig. 2.8 (b) shows that $E[t_d]$ is insignificantly affected by λ_m . This phenomenon can be explained as follows: Since the prepaid messages are random observation points of the prepaid call holding interval, the delivery delays are not significantly affected by the message arrival rate λ_m .

Effects of the initial prepaid credit X . Fig. 2.9 plots P_{UFT} and $E[t_d]$ against X_T and X , where $1/\lambda_m = 0.5E[t_c]$. This figure shows that both P_{UFT} and $E[t_d]$ decrease as X increases. When the initial prepaid credit X increases, there are more CUs left for a prepaid call, and it is more likely that there are enough CUs for both the remaining call and the message (i.e., the amount of the prepaid credits left is larger than $X_T + T_m \geq t_c + T_m$). Therefore, the performance of both P_{UFT} and $E[t_d]$ improve as X increases. For $X_T = 3E[t_c]$, when X increases from $20E[t_c]$ to $50E[t_c]$, P_{UFT} decreases by 58.64% and $E[t_d]$ decreases by 36.71%. When X increases from $50E[t_c]$ to $100E[t_c]$, P_{UFT} decreases by 49.96% and $E[t_d]$ decreases by 50.87%. When $X > 150E[t_c]$ (i.e., the initial prepaid credit can support more than 150 voice calls), increasing X only has insignificant impacts on P_{UFT} and $E[t_d]$.

Effects of the threshold X_T . From Figs. 2.6-2.9, when X_T is small, increasing X_T reduces P_{UFT} significantly. When $X_T \geq 5E[t_c]$, effects of X_T on P_{UFT} becomes insignificant. On the other hand, increasing X_T always increases $E[t_d]$. Therefore

in these scenarios, it is appropriate to choose $X_T = 5E[t_c]$ in the PAS.

2.6 Summary

This chapter proposed a SIP-based prepaid application server to handle both the prepaid IMS-to-PSTN calls and messaging services in UMTS. When both voice and messaging are simultaneously offered, a strategy is required to determine if a prepaid message can be sent out during an in-progress call without force-terminating this call. To avoid unnecessary force-termination, a threshold amount X_T of prepaid credit is set to protect the in-progress IMS-to-PSTN call. This paper provided guidelines to select an appropriate X_T . The output measures are the UFT (unnecessary force-termination) probability P_{UFT} and the expected unnecessary delay $E[t_d]$. We investigated how these two output measures are affected by input parameters including the expected call holding time $E[t_c]$, the variance V_c of the call holding time, the message arrivals rate λ_m , X_T and the initial prepaid credit X . We make the following observations:

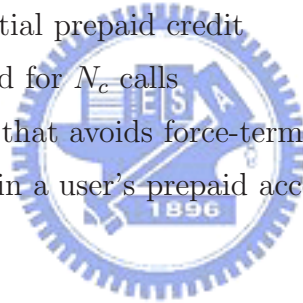
- When the proportion of X and $E[t_c]$ is fixed, P_{UFT} and $E[t_d]$ are not affected by the change of $E[t_c]$.
- The performance of both P_{UFT} and $E[t_d]$ degrade as V_c increases.
- P_{UFT} increases as λ_m increases. This effect becomes insignificant when X_T is large (e.g., $X_T \geq 6E[t_c]$ in our examples). On the other hand, $E[t_d]$ is insignificantly affected by λ_m .
- Both P_{UFT} and $E[t_d]$ decrease as X increases.
- When X_T is small, increasing X_T reduces P_{UFT} significantly. When X_T is large (e.g. $X_T > 5E[t_c]$ in our examples), increasing X_T only has insignificant impact on P_{UFT} . On the other hand, increasing X_T always increases $E[t_d]$.

Based on the above discussion, the network operator can select the appropriate X_T values for various traffic conditions based on our study.

2.7 Notation

The notation used in this chapter is listed below.

- $1/\lambda_c$: the expected interval between when the previous prepaid call completes and when the next prepaid call arrives
- λ_m : the arrival rate of the instant messages
- N_c : the number of prepaid call completed
- N_m : the number of prepaid message arrived
- $N(t)$: the number of prepaid messages occurring in period t
- P_{UFT} : the UFT (unnecessary force-termination) probability of an in-progress call in the prepaid application server
- τ_i : the interval between when the i -th prepaid call completes and when the $i + 1$ -st prepaid call starts
- t_c : the prepaid call holding time
- t_d : the extra delay between when an instant message service request arrives and when the instant message is delivered
- T_m : the charge for a prepaid instant message delivery
- V_c : the variance of the call holding time t_c
- X : the amount of the initial prepaid credit
- X_c : the credit units spend for N_c calls
- X_T : the credit threshold that avoids force-termination of prepaid calls
- x^* : the remaining credit in a user's prepaid account



Chapter 3

Modeling Online Credit Reservation Procedure

Prepaid telecommunications service requires a user to make an advanced payment before enjoying the service. Usage of prepaid service does not require deposit and monthly bill. Instead the usage fee is directly deducted from the user's prepaid account. Four billing technologies have been used in mobile prepaid service: hot billing approach [24], service node approach [23, 25], *Intelligent Network* (IN) approach [35] and handset-based approach [36]. By creating hybrid online/offline billing models, 3GPP Releases 5 and 6 propose the IP-based *Online Charging System* (OCS) [10] to allow both prepaid and postpaid subscribers to be charged in real-time.

As described in Section 1.3, online charging provides real-time charging information to the billing system. This chapter investigates credit reservation for the OCS. We assume that a mobile user may access n types of session-based IMS services. After an online service user has purchased some credit units, she is allowed to enjoy multiple sessions simultaneously. Each service type has its own traffic characteristics and communications parameters. For example, the average call holding time for a VoIP call session is 1-3 minutes, and the average session holding time for the interactive mobile gaming sessions may range from 10 to 30 minutes [18, 26]. Note that the service sessions can be charged according to time (duration) or transmitted packet volumes. For a session of type i , each time the OCS grants θ_i credit units to the session. When these credit units are consumed, the OCS grants extra credit units to the session through the Diameter credit reservation procedure. When the balance of the user account at the OCS is below a recharge threshold C_{min} , the OCS does not allow the user to initiate new sessions, and reminds the user to refill the prepaid account by sending a recharge message. This chapter shows how to

select an appropriate recharge threshold C_{min} . Note that the usage of C_{min} is different from that of X_T in Chapter 2. The notation used in this chapter is listed in Section 3.8.

3.1 Recharge Threshold-based Credit Reservation (RTCR)

The Diameter credit reservation procedure exercised between an IMS *Application Server* (AS) and the OCS is described in Fig. 1.3. In the Reserve Units /Debit Units operations (i.e., Steps 1 or 3 of this procedure), the IMS AS sends the CCR messages to the OCS to request extra credit. In both Steps 2 and 4 of the procedure, the OCS reserves extra credit units for the IMS AS. Specifically, the OCS needs to determine how to allocate the credit units when the remaining credit left in the prepaid account is too small, and when to send the recharge message. For the discussion purpose, we refer an *RU* operation as a Reserve Units operation (Step 1) or a Reserve Units and Debit operation (Step 3). The credit allocation issue can be addressed by a simple mechanism called *Recharge Threshold-based Credit Reservation* (RTCR). In this mechanism, when the amount of the remaining prepaid credit in the OCS is less than a recharge threshold C_{min} , the OCS reminds the user to refill the prepaid account by sending a recharge message. Also, the OCS will reject new service requests, and only allow the existing service sessions to consume the remaining credit. We say that the RTCR execution “ends” if all in-progress sessions are complete or force-terminated after the recharge threshold C_{min} is reached.

In RTCR, if C_{min} is set too small, then the amount of the remaining credit may not be large enough to support all in-progress service sessions, and some of them will be forced to terminate. Clearly, force-termination degrades user satisfaction. For example, when all credit units are consumed during an interactive multimedia game, the user is forced to terminate the game.

On the other hand, if C_{min} is set too large, the OCS will inappropriately reject new session requests that should be accommodated and can be completed before the user’s credit is actually depleted. The user will be frequently asked to refill the prepaid account while she still has enough credit. Therefore, C_{min} should be appropriately selected to balance unnecessary force-termination against frequent recharging. In the next section, an analytic model is proposed to study the impact of parameter C_{min} on the RTCR mechanism.

3.2 Analytic Modeling for RTCR Mechanism

This section proposes an analytic model to investigate the RTCR mechanism. Assume that there are n types of session-based IMS services. For $1 \leq i \leq n$, the sessions for type- i service can be activated from time to time. In Fig. 3.1, the current type- i service session starts at t_0 and completes at t_4 , and the next type- i service session starts at t_5 . Let the service session holding time be $t_{h,i} = t_4 - t_0$ and the inter-arrival time be $t_{a,i} = t_5 - t_4$. For VoIP call session services, $t_{h,i}$ represents the call holding time. For mobile data downloading service, $t_{h,i}$ represents the file transmission time. Assume that $t_{h,i}$ and $t_{a,i}$ are exponentially distributed with rates μ_i and λ_i , respectively (the exponential assumptions will be relaxed in the simulation experiments). We assume that each time unit of the type- i service session is charged for α_i credit units. Without loss of generality, let $\alpha_i = 1$ (i.e., the time unit is equal to the credit unit). Let C denote the amount of the initial prepaid credit for a mobile user. Let θ_i be the amount of credit that the OCS grants in each RU operation for a type- i session. It is essential to select appropriate C_{min} and θ_i values to “optimize” the performance of the RTCR mechanism in terms of the following output measures:

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

3.2.1 Derivation for the Number of RU Operations

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

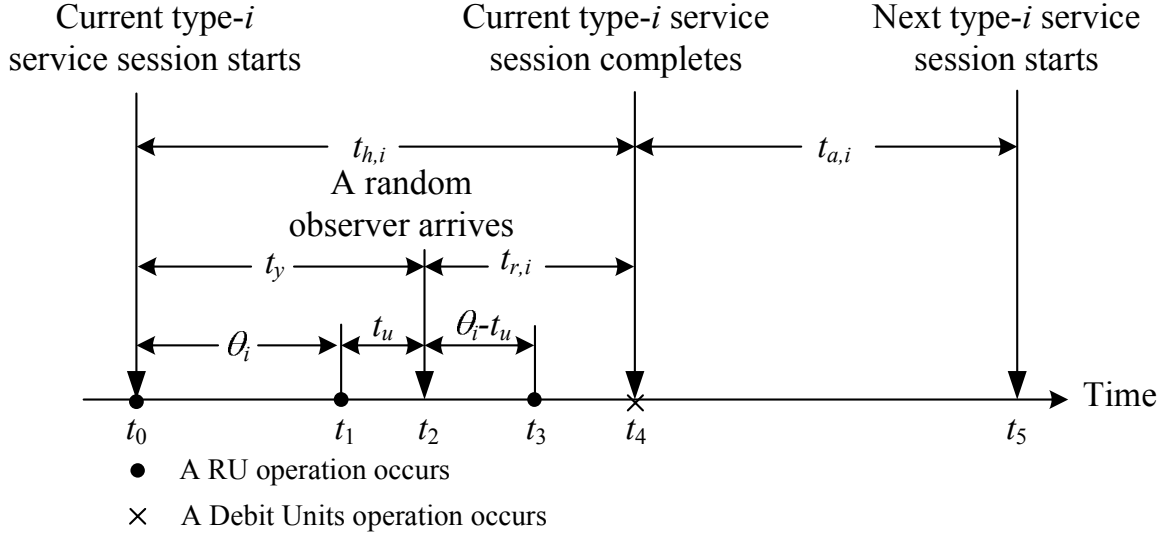


Figure 3.1: Timing Diagram for the RTCR mechanism.

units from the OCS. Then

$$\begin{aligned}
 E[N_{r,i}] &= 1 + \Pr[t_{h,i} > \theta_i] + \Pr[t_{h,i} > 2\theta_i] + \Pr[t_{h,i} > 3\theta_i] + \dots \\
 &= 1 + \sum_{j=1}^{\infty} \Pr[t_{h,i} > j\theta_i]
 \end{aligned} \tag{3.1}$$

Eq. (3.1) says that the first RU operation (i.e., the Reserve Units operation) is always executed, and for $j > 0$, the j -th RU operation (i.e., the Reserve Units and Debits Units operation) will be executed with probability $\Pr[t_{h,i} > j\theta_i]$. Since $t_{h,i}$ is exponentially distributed with the density function

$$f_{h,i}(t_{h,i}) = \mu_i e^{-\mu_i t_{h,i}} \tag{3.2}$$

From (3.2), Eq. (3.1) is derived as

$$\begin{aligned}
 E[N_{r,i}] &= 1 + \sum_{j=1}^{\infty} \int_{t_{h,i}=j\theta_i}^{\infty} f_{h,i}(t_{h,i}) dt_{h,i} \\
 &= 1 + \sum_{j=1}^{\infty} e^{-\mu_i j\theta_i} = \frac{1}{1 - e^{-\mu_i \theta_i}}
 \end{aligned} \tag{3.3}$$

3.3 Exact Analytic Model for Single-type Service

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

3.3.1 Derivation for the Force-termination Probability

Consider the timing diagram in Fig. 3.1 where the RU operations in the current service session occur at t_0 , t_1 and t_3 , respectively. Assume that the OCS grants θ_1 units to the AS in every RU operation and $\theta_1 \leq C_{min}$. Then $t_3 - t_1 = t_1 - t_0 = \theta_1$. Suppose that a random observer arrives at t_2 . Let $t_y = t_2 - t_0$ and $t_{r,1} = t_4 - t_2$ be the elapsed holding time (i.e., the age) and the residual holding time of the service session, respectively. Let $t_u = t_2 - t_1$ denote the consumed credit (time) units in the AS with density function $f_u(t_u)$. Then $0 \leq t_u \leq \theta_1$ and the unused credit units left in the AS is $\theta_1 - t_u$.

For an in-progress service session at a particular time point t , the exact unused credit units for the user is $C_x(t) = C_r(t) + (\theta_1 - t_u)$, which is the sum of the remaining credit units $C_r(t)$ in the OCS and the unused credit units $(\theta_1 - t_u)$ in the AS. Note that at times t_0 , t_1 and t_3 , no unused credit units are left in the AS and therefore $C_x(t_0) = C_r(t_0)$, $C_x(t_1) = C_r(t_1)$ and $C_x(t_3) = C_r(t_3)$. Define “critical” time t^* as the time when $C_x(t^*) = C_{min} + \theta_1$. The first RU operation occurs after t^* is referred to as the “critical” RU operation. Immediately after the critical RU operation is performed, the OCS will send the recharge message, and newly incoming session requests will be rejected. The critical RU operation may occur during a session execution (Case I; see Fig. 3.2 (a)) or when a new session arrives (Case II; see Fig. 3.2 (b)). In Fig. 3.2, let $\tilde{\theta} = \theta_1 - (\min(t_7, t_8) - t^*)$. Note that when the critical RU operation arrives at t_7 , $C_x(t_7) = C_r(t_7) = C_{min} + \tilde{\theta}$. Furthermore, in Case II, no credit units are consumed during $[t_8, t_7]$, and $C_x(t_7) = C_x(t_8)$. The $\tilde{\theta}$ value will be derived later. After the OCS has granted θ_1 units to this request, the remaining credit left in the prepaid account becomes $C_r(t_7^+) = C_{min} - (\theta_1 - \tilde{\theta}) < C_{min}$. At this point, the OCS will send a recharge message to the prepaid user, and the maximum service time that the remaining credit can support is $C_{min} + \tilde{\theta}$. Therefore P_f is computed as

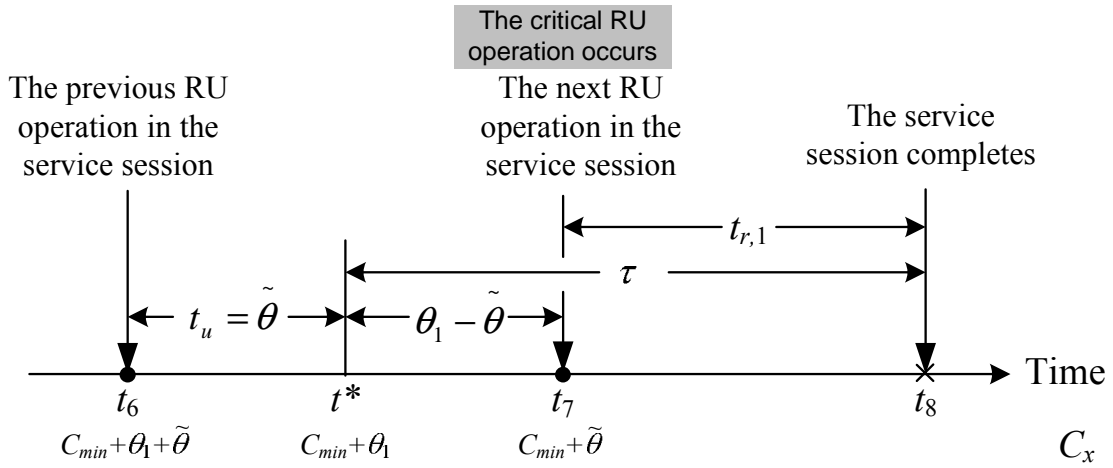
$$P_f = \Pr[t_{r,1} > C_{min} + \tilde{\theta}] \quad (3.4)$$

Since the session holding time $t_{h,1}$ is exponentially distributed with rate μ_1 , and from the residual time theorem [48], $t_{r,1}$ has the same density function as $t_{h,1}$; i.e.,

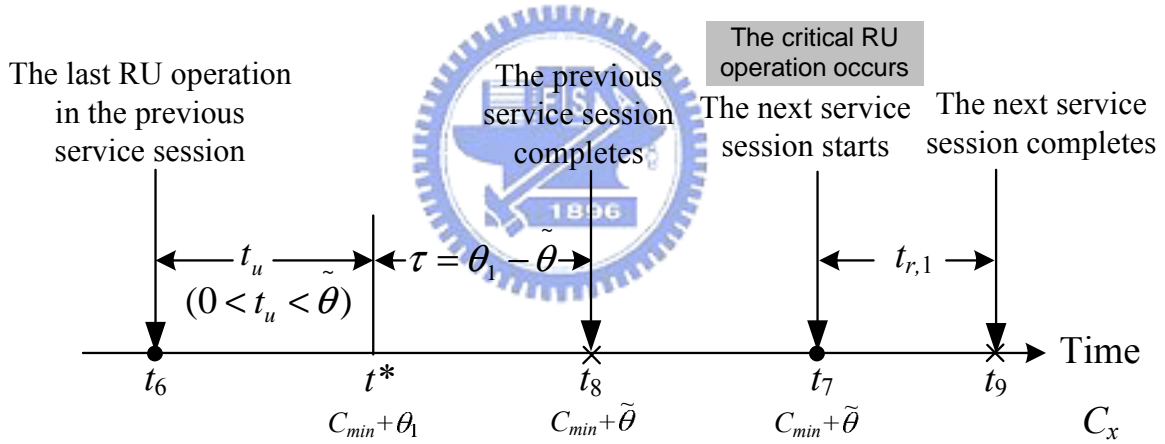
$$f_{r,1}(t_{r,1}) = \mu_1 e^{-\mu_1 t_{r,1}} \quad (3.5)$$

In (3.4), $\tilde{\theta}$ is derived as follows:

Case I. In Fig. 3.2 (a), the critical RU operation is issued by an in-progress session at t_7 . In this case, the previous RU operation is issued at $t_6 = t_7 - \theta_1$ where



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



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

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

Figure 3.2: Timing Diagram for Deriving $\tilde{\theta}$.

$C_r(t_6) = C_x(t_6) = C_{min} + \theta_1 + \tilde{\theta}$, and $t_6 < t^* < t_7$. At the critical time t^* , the consumed time units t_u must be $\tilde{\theta}$ (with density $f_u(\tilde{\theta})$) and the residual holding time $\tau = t_8 - t^*$ for the session must be longer than $\theta_1 - \tilde{\theta}$ (with probability $\int_{\tau=\theta_1-\tilde{\theta}}^{\infty} f_{r,1}(\tau)d\tau$). Therefore, the density of $\tilde{\theta}$ for Case I is expressed as

$$f_{\tilde{\theta},I}(\tilde{\theta}) = f_u(\tilde{\theta}) \int_{\tau=\theta_1-\tilde{\theta}}^{\infty} f_{r,1}(\tau)d\tau \quad (3.6)$$

Case II. In Fig. 3.2 (b), the critical RU operation is issued by a new session at t_7 . The last RU operation in the previous service session occurs at t_6 and the previous service session completes at t_8 , and $C_x(t_8) = C_{min} + \tilde{\theta}$. Note that $t^* > t_6 \geq t_8 - \theta_1$. Therefore, $C_{min} + \theta_1 + \tilde{\theta} \geq C_x(t_6) > C_{min} + \theta_1$. At t^* , the AS consumes t_u credit units, where $0 < t_u < \tilde{\theta}$ (with probability $\int_{t_u=0}^{\tilde{\theta}} f_u(t_u)dt_u$), and the residual holding time $\tau = t_8 - t^*$ for the session is $\theta_1 - \tilde{\theta}$ (with density $f_{r,1}(\theta_1 - \tilde{\theta})$). Therefore, the density of $\tilde{\theta}$ for Case II is expressed as

$$f_{\tilde{\theta},II}(\tilde{\theta}) = \int_{t_u=0}^{\tilde{\theta}} f_u(t_u)dt_u f_{r,1}(\theta_1 - \tilde{\theta}) \quad (3.7)$$

Combining (3.6) and (3.7), we have

$$f_{\tilde{\theta}}(\tilde{\theta}) = f_u(\tilde{\theta}) \int_{\tau=\theta_1-\tilde{\theta}}^{\infty} f_{r,1}(\tau)d\tau + \int_{t_u=0}^{\tilde{\theta}} f_u(t_u)dt_u f_{r,1}(\theta_1 - \tilde{\theta}) \quad (3.8)$$

In (3.8), $f_u(t_u)$ is derived as follows: Let t_y be the elapsed holding time of the in-progress session with density function $f_y(t_y)$. According to the reverse residual time theorem [48], t_y has the same distribution as the residual of $t_{h,1}$. That is, $f_y(t_y) = f_{r,1}(t_y) = \mu_1 e^{-\mu_1 t_y}$. Since $t_u = t_y - \left\lfloor \frac{t_y}{\theta_1} \right\rfloor \theta_1$, the density function for t_u (for $0 \leq t_u \leq \theta_1$) is

$$f_u(t_u) = \sum_{j=0}^{\infty} f_y(t_u + j\theta_1) = \frac{\mu_1 e^{-\mu_1 t_u}}{1 - e^{-\mu_1 \theta_1}} \quad (3.9)$$

Substituting (3.5) and (3.9) into (3.8), we have

$$\begin{aligned} f_{\tilde{\theta}}(\tilde{\theta}) &= \left(\frac{\mu_1 e^{-\mu_1 \tilde{\theta}}}{1 - e^{-\mu_1 \theta_1}} \right) \int_{\tau=\theta_1-\tilde{\theta}}^{\infty} \mu_1 e^{-\mu_1 \tau} d\tau + \int_{t_u=0}^{\tilde{\theta}} \left(\frac{\mu_1 e^{-\mu_1 t_u}}{1 - e^{-\mu_1 \theta_1}} \right) dt_u \mu_1 e^{-\mu_1 (\theta_1 - \tilde{\theta})} \\ &= \frac{\mu_1 e^{\mu_1 \tilde{\theta}}}{e^{\mu_1 \theta_1} - 1} \end{aligned} \quad (3.10)$$

From (3.4), (3.5) and (3.10), P_f is derived as

$$\begin{aligned}
P_f &= \int_{\tilde{\theta}=0}^{\theta_1} \int_{t_{r,1}=C_{min}+\tilde{\theta}}^{\infty} f_{r,1}(t_{r,1}) f_{\tilde{\theta}}(\tilde{\theta}) dt_{r,1} d\tilde{\theta} \\
&= \int_{\tilde{\theta}=0}^{\theta_1} \int_{t_{r,1}=C_{min}+\tilde{\theta}}^{\infty} \mu_1 e^{-\mu_1 t_{r,1}} \left(\frac{\mu_1 e^{\mu_1 \tilde{\theta}}}{e^{\mu_1 \theta_1} - 1} \right) dt_{r,1} d\tilde{\theta} \\
&= \frac{\mu_1 \theta_1 e^{-\mu_1 C_{min}}}{e^{\mu_1 \theta_1} - 1}
\end{aligned} \tag{3.11}$$

3.3.2 Derivation for the Unused Credit Units

It is clear that $E[C_d] = E[\lim_{t \rightarrow \infty} C_r(t)]$ assuming that the prepaid account is not recharged at the end of the RTCR execution. In Fig. 3.2, the critical RU operation occurs at t_7 when $C_x(t_7) = C_{min} + \tilde{\theta}$. If $C_{min} + \tilde{\theta} > t_{r,1}$, then $C_d = C_{min} + \tilde{\theta} - t_{r,1}$. Otherwise, $C_d = 0$. Therefore, $E[C_d]$ is expressed as

$$E[C_d] = E[\max\{C_{min} + \tilde{\theta} - t_{r,1}, 0\}] = E[C_{min} + \tilde{\theta} - t_{r,1}; C_{min} + \tilde{\theta} > t_{r,1}] \tag{3.12}$$

From (3.5) and (3.10), Eq. (3.12) is derived as

$$\begin{aligned}
E[C_d] &= \int_{\tilde{\theta}=0}^{\theta_1} \int_{t_{r,1}=0}^{C_{min}+\tilde{\theta}} (C_{min} + \tilde{\theta} - t_{r,1}) f_{r,1}(t_{r,1}) f_{\tilde{\theta}}(\tilde{\theta}) dt_{r,1} d\tilde{\theta} \\
&= C_{min} + \frac{\theta_1 (e^{\mu_1 \theta_1} + e^{-\mu_1 C_{min}})}{e^{\mu_1 \theta_1} - 1} - \frac{2}{\mu_1}
\end{aligned} \tag{3.13}$$

3.4 Approximate Analytic Model for Multiple-type Services

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

3.4.1 Derivation for the Force-termination Probability

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

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

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

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

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

$$P_f = P_A \Pr[t_{r,1} > C_{min}] + P_B \Pr[t_{r,2} > C_{min}] + P_C \Pr[t_{r,1} + t_{r,2} > C_{min}] \quad (3.14)$$

In (3.14) we assume that $C_x(t_7) = C_{min}$, where t_7 is the time that the critical RU operation occurs. Note that $C_{min} \leq C_x(t_7) < C_{min} + \sum_i \theta_i$. Therefore, (3.14) incurs error when θ_i is large. Substituting (3.5) into (3.14) to yield

$$\begin{aligned} P_f &= P_A \int_{t_{r,1}=C_{min}}^{\infty} f_{r,1}(t_{r,1}) dt_{r,1} + P_B \int_{t_{r,2}=C_{min}}^{\infty} f_{r,2}(t_{r,2}) dt_{r,2} \\ &\quad + P_C \int_{t=C_{min}}^{\infty} \int_{t_{r,1}=0}^t f_{r,2}(t-t_{r,1}) f_{r,1}(t_{r,1}) dt_{r,1} dt \\ &= P_A e^{-\mu_1 C_{min}} + P_B e^{-\mu_2 C_{min}} + P_C \left(\frac{\mu_1 e^{-\mu_2 C_{min}} - \mu_2 e^{-\mu_1 C_{min}}}{\mu_1 - \mu_2} \right) \end{aligned} \quad (3.15)$$

In (3.15), the probabilities P_A , P_B and P_C are derived as follows: We first compute the probability p_i that a type- i service session is active at a random observation point. In Fig.3.1, a random observation point occurs at t_2 in the renewal period $[t_0, t_5]$. From the alternating renewal theory [48], p_i is expressed as

$$p_i = \frac{E[t_{h,i}]}{E[t_{h,i}] + E[t_{a,i}]} = \frac{\lambda_i}{\mu_i + \lambda_i} \quad (3.16)$$

At t_2 , there is only one active type-1 service session with probability $p_1(1-p_2)$, there is only one active type-2 service session with probability $p_2(1-p_1)$, and both type-1 and type-2 service sessions are active with probability $p_1 p_2$. The critical RU operation is issued by a type-1 and a type-2 service sessions in Cases A and B, respectively. In Case C, the critical RU operation may be issued by a type-1 or a type-2 services. Since the sending of recharge message can be modeled as a random observer for sufficiently small θ_i value in a service session, from (3.16), the ratio $P_A : P_B : P_C$ can be computed as

$$P_A : P_B : P_C \approx p_1(1-p_2) : p_2(1-p_1) : 2p_1 p_2 = \lambda_1 \mu_2 : \mu_1 \lambda_2 : 2\lambda_1 \lambda_2 \quad (3.17)$$

Since $P_A + P_B + P_C = 1$ and from (3.17), we have

$$P_A \approx \frac{\lambda_1 \mu_2}{\lambda_1 \mu_2 + \mu_1 \lambda_2 + 2\lambda_1 \lambda_2}, \quad P_B \approx \frac{\mu_1 \lambda_2}{\lambda_1 \mu_2 + \mu_1 \lambda_2 + 2\lambda_1 \lambda_2}, \quad P_C \approx \frac{2\lambda_1 \lambda_2}{\lambda_1 \mu_2 + \mu_1 \lambda_2 + 2\lambda_1 \lambda_2} \quad (3.18)$$

Substitute (3.18) into (3.15) to yield

$$P_f = \left[\lambda_1 \mu_2 e^{-\mu_1 C_{min}} + \mu_1 \lambda_2 e^{-\mu_2 C_{min}} + 2\lambda_1 \lambda_2 \left(\frac{\mu_1 e^{-\mu_2 C_{min}} - \mu_2 e^{-\mu_1 C_{min}}}{\mu_1 - \mu_2} \right) \right] \times \left(\frac{1}{\lambda_1 \mu_2 + \mu_1 \lambda_2 + 2\lambda_1 \lambda_2} \right) \quad (3.19)$$

For $n > 2$, Eq. (3.14) can be extended by including all active session combinations (there are $2^n - 1$ combinations). Then P_f can be computed following the same derivations for (3.15)-(3.19).

3.4.2 Derivation for the Unused Credit Units

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

$$\begin{aligned} E[C_d] &= P_A E[\max\{C_{min} - t_{r,1}, 0\}] + P_B E[\max\{C_{min} - t_{r,2}, 0\}] \\ &\quad + P_C E[\max\{C_{min} - t_{r,1} - t_{r,2}, 0\}] \\ &= P_A E[C_{min} - t_{r,1}; C_{min} > t_{r,1}] + P_B E[C_{min} - t_{r,2}; C_{min} > t_{r,2}] \\ &\quad + P_C E[C_{min} - t_{r,1} - t_{r,2}; C_{min} > t_{r,1} + t_{r,2}] \end{aligned} \quad (3.20)$$

Substitute (3.5) and (3.18) into (3.20) to yield

$$\begin{aligned} E[C_d] &= P_A \int_{t_{r,1}=0}^{C_{min}} (C_{min} - t_{r,1}) f_{r,1}(t_{r,1}) dt_{r,1} + P_B \int_{t_{r,2}=0}^{C_{min}} (C_{min} - t_{r,2}) f_{r,2}(t_{r,2}) dt_{r,2} \\ &\quad + P_C \int_{t=0}^{C_{min}} (C_{min} - t) \int_{t_{r,1}=0}^t f_{r,2}(t - t_{r,1}) f_{r,1}(t_{r,1}) dt_{r,1} dt \\ &= C_{min} - \left(\frac{1}{\lambda_1 \mu_2 + \mu_1 \lambda_2 + 2\lambda_1 \lambda_2} \right) \left\{ \frac{\lambda_1 \mu_2 (1 - e^{-\mu_1 C_{min}})}{\mu_1} + \frac{\mu_1 \lambda_2 (1 - e^{-\mu_2 C_{min}})}{\mu_2} \right. \\ &\quad \left. + \frac{2\lambda_1 \lambda_2 [\mu_1^2 (1 - e^{-\mu_2 C_{min}}) - \mu_2^2 (1 - e^{-\mu_1 C_{min}})]}{\mu_1 \mu_2 (\mu_1 - \mu_2)} \right\} \end{aligned} \quad (3.21)$$

For $n > 2$, $E[C_d]$ can be computed through the same derivations for (3.20)-(3.21) by considering $2^n - 1$ active session combinations.

Table 3.1: Comparison of the Analytic and Simulation Results ($n = 1$)

C_{min} (Unit: c)	1	2	3	4	5
Simulation	21.4081%	7.8867%	2.9089%	1.0616%	0.3926%
Analytic	21.4097%	7.8762%	2.8975%	1.0659%	0.3921%
Error	0.01%	-0.13%	-0.39%	0.41%	-0.11%

(a) P_f ($\theta_1 = 1/\mu_1$)

C_{min} (Unit: c)	1	2	3	4	5
Simulation	0.8004	1.6481	2.5971	3.5880	4.5820
Analytic	0.7961	1.6607	2.6110	3.5926	4.5859
Error	-0.55%	0.76%	0.53%	0.13%	0.08%

(b) $E[C_d]$ (unit: c) ($\theta_1 = 1/\mu_1$)

3.5 Simulation Validation

The analytic model developed in this section is validated against the simulation experiments. The discrepancies between analytic analysis (specifically, Eqs. (3.11), (3.13), (3.19) and (3.21)) and simulation are within 2% in Tables 3.1 and 3.2. The simulation model follows the discrete event approach described in Chapter 2, and the details are omitted. The input parameter θ_i is normalized by the mean $1/\mu_i$ of the service session holding time. The input parameter C_{min} and output measure $E[C_d]$ are normalized by the expected credit units c consumed in a session, where c is derived as follows:

Consider $n = 2$ and let n_1 and n_2 be the numbers of session completions for type-1 and type-2 services in an observation period, respectively. In (3.16), p_i represents the fraction of time that the type- i service is active in an observed period, and the expected service session holding time for a type- i service is $1/\mu_i$, the ratio $n_1 : n_2$ can be computed as

$$n_1 : n_2 = p_1\mu_1 : p_2\mu_2 = \frac{\lambda_1\mu_1}{\mu_1 + \lambda_1} : \frac{\lambda_2\mu_2}{\mu_2 + \lambda_2} \quad (3.22)$$

From (3.22), c can be computed as

$$\begin{aligned} c &= \frac{n_1/\mu_1 + n_2/\mu_2}{n_1 + n_2} \\ &= \frac{\lambda_1(\mu_2 + \lambda_2) + \lambda_2(\mu_1 + \lambda_1)}{\lambda_1\mu_1(\mu_2 + \lambda_2) + \lambda_2\mu_2(\mu_1 + \lambda_1)} \end{aligned} \quad (3.23)$$

Table 3.2: Comparison of the Analytic and Simulation Results ($n = 2$)

C_{min} (Unit: c)	1	2	3	4	5
Simulation	57.3683%	31.1314%	16.3689%	8.5267%	4.4320%
Analytic	57.5872%	31.2126%	16.4590%	8.5647%	4.4274%
Error	0.38%	0.26%	0.55%	0.44%	-0.10%

(a) P_f ($\theta_i = 0.01/\mu_i$, $\lambda_1 = \mu_1$ and $\mu_2 = \lambda_2 = 2\mu_1$)

C_{min} (Unit: c)	1	2	3	4	5
Simulation	0.2281	0.7970	1.5672	2.4478	3.3837
Analytic	0.2257	0.7937	1.5628	2.4419	3.3792
Error	-1.06%	-0.41%	-0.28%	-0.24%	-0.13%

(b) $E[C_d]$ (unit: c) ($\theta_i = 0.01/\mu_i$, $\lambda_1 = \mu_1$ and $\mu_2 = \lambda_2 = 2\mu_1$)

3.6 Numerical Examples

This section uses numerical examples to investigate the performance of the RTCR mechanism. For the examples in Figs. 3.4 and 3.5, $n = 2$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$. In Fig. 3.6, $1 \leq n \leq 4$ and $\lambda_i = \mu_i = i\mu_1$. Similar results are observed for other parameter values and are not presented. The effects of the input parameters are described below.

Effects of θ_i on $E[N_{r,i}]$. Fig. 3.3 plots the expected number $E[N_{r,i}]$ of RU operations executed in a type- i service session against the granted credit θ_i . Note that $E[N_{r,i}]$ is not affected by C_{min} and n . This figure shows a trivial result that $E[N_{r,i}]$ decreases as θ_i increases. A non-trivial observation is that when $\theta_i \geq 2.5/\mu_i$, $E[N_{r,i}] \approx 1$. It implies that selecting θ_i value larger than $2.5/\mu_i$ will not improve the $E[N_{r,i}]$ performance.

Effects of C_{min} . Fig. 3.4 plots the force-termination probability P_f and the expected credit $E[C_d]$ against θ_i and C_{min} , where $n = 2$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$. Fig. 3.4 (a) shows that P_f decreases as C_{min} increases. When the critical RU operation occurs, more unused credit units are available in the prepaid account when C_{min} increases. Therefore, the possibility of force-termination reduces. For $\theta_i = 1/\mu_i$, when C_{min} increases from $2c$ to $4c$, P_f decreases from 18.78% to 4.99%. Fig. 3.4 (b) shows that $E[C_d]$ increases as C_{min} increases. It is apparent that when the critical RU operation occurs, the exact unused credit units $C_x(t^*)$ for the user increases as

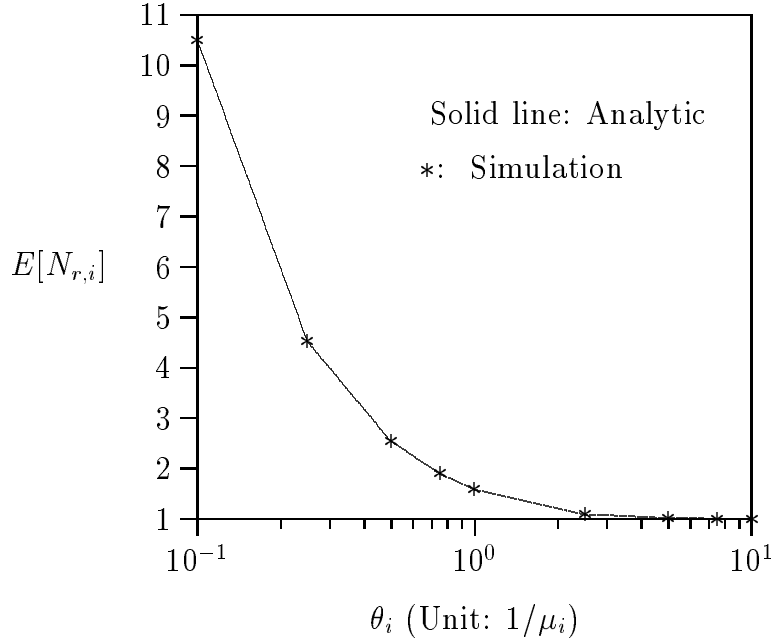


Figure 3.3: Effects of θ_i on $E[N_{r,i}]$.

C_{min} increases. That is, the amount of consumed credit reduces, and the expected credit $E[C_d]$ increases. For $\theta_i = 1/\mu_i$, when C_{min} increases from $2c$ to $4c$, $E[C_d]$ increases from $1.60c$ to $3.41c$. In this scenario, we expect that $3.41 - 1.60 = 1.81$ more sessions are complete when $C_{min} = 2c$ than when $C_{min} = 4c$.

Effects of θ_i . Fig. 3.4 (a) shows that P_f is a decreasing function of θ_i . When θ_i increases, more credit units are granted to the AS. Therefore, the possibility of force-termination reduces. For $C_{min} = 6c$, when θ_i increases from $1/\mu_i$ to $2.5/\mu_i$, P_f decreases from 1.32% to 0.37%. This effect becomes insignificant when θ_i is large (e.g., $\theta_i \geq 5/\mu_i$). Fig. 3.4 (b) shows that $E[C_d]$ is an increasing function of θ_i . For $C_{min} = 6c$, when θ_i increases from $1/\mu_i$ to $2.5/\mu_i$, the $E[C_d]$ increases from $5.37c$ to $7.64c$. Fig. 3.4 (b) also quantitatively indicates how the θ_i and C_{min} values affect $E[C_d]$. When $\theta_i \leq 1/\mu_i$, $E[C_d] \approx C_{min}$. On the other hand, $E[C_d] \gg C_{min}$ as θ_i increases. For example, when $C_{min} = 6c$ and $\theta_i = 10/\mu_i$, $E[C_d] = 23.63c \gg 6c$.

Effects of $V_{h,i}$. Fig. 3.5 plots P_f and $E[C_d]$ against C_{min} and the variance $V_{h,i}$ of the Gamma service session holding time $t_{h,i}$, where $n = 2$, $\theta_i = 2.5\mu_i$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$. Fig. 3.5 (a) shows that P_f increases as $V_{h,i}$ increases. This phenomenon is explained as follows: As $V_{h,i}$ increases, more long and short $t_{h,i}$ periods are observed.

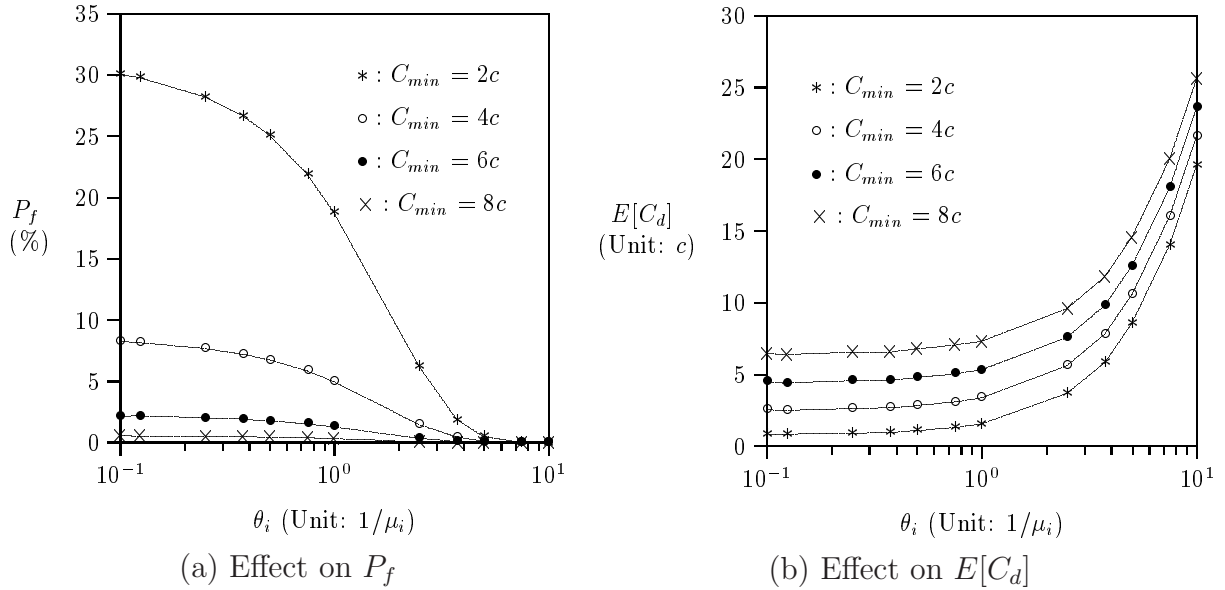
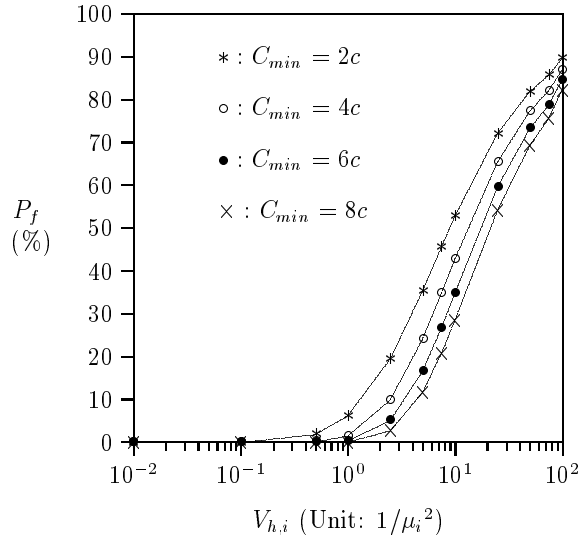


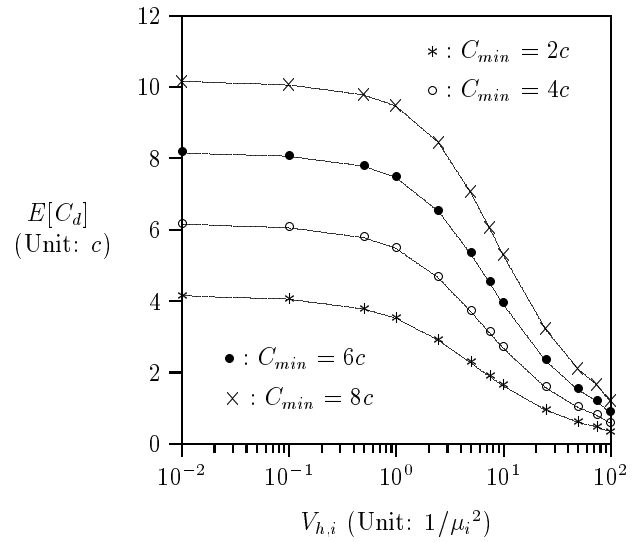
Figure 3.4: Effects of θ_i and C_{min} ($n = 2$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$)

The recharge message is more likely to be sent in the long $t_{h,i}$ periods than the short $t_{h,i}$ periods, and larger residual service session holding time $t_{r,i}$ are expected. Therefore, P_f increases as $V_{h,i}$ increases. Fig. 3.5 (b) shows that $E[C_d]$ decreases as $V_{h,i}$ increases. As $V_{h,i}$ increases, the recharge message is likely to be sent in the long $t_{h,i}$ periods. Then the possibility that $t_{r,i} \geq C_{min}$ (i.e., $C_d = 0$) increases. Therefore $E[C_d]$ decreases as $V_{h,i}$ increases.

Effects of n . Fig. 3.6 plots P_f and $E[C_d]$ against θ_i and the number n of the service session types. Fig. 3.6 (a) shows that P_f is an increasing function of n . As n increases, the number of in-progress service sessions increases when the recharge message is sent, and therefore P_f increases. In Fig. 3.6 (b), when θ_i is small (e.g. $\theta_i \leq 1/\mu_i$), $E[C_d]$ decreases as n increases. On the other hand, when θ_i is large (e.g. $\theta_i \geq 2.5/\mu_i$), $E[C_d]$ increases as n increases. When the recharge message is sent, the number of simultaneous in-progress service sessions increases as n increases. There are two conflicting effects as n increases. First, more credit units will be consumed by these sessions and $E[C_d]$ decreases. Second, the “net” unused credit units which have granted to the AS increases, and $E[C_d]$ increases. When θ_i is small, the first effect will dominate. On the other hand, the second effect will dominate when θ_i is large.

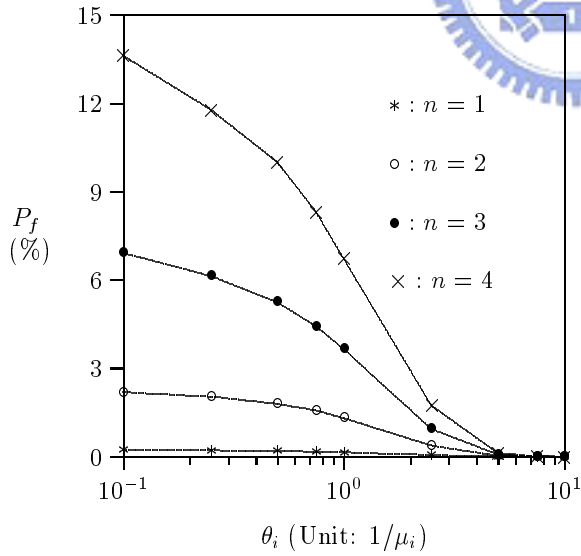
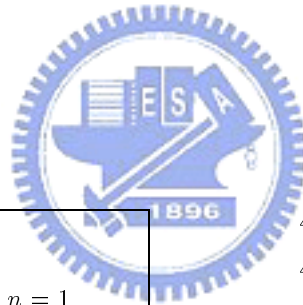


(a) Effect on P_f

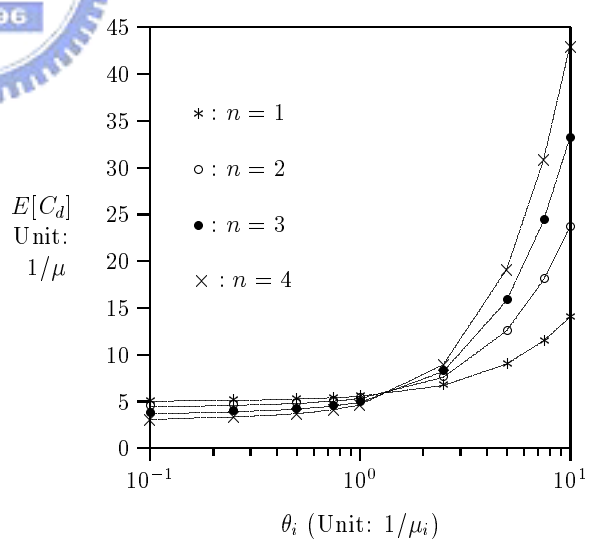


(b) Effect on $E[C_d]$

Figure 3.5: Effects of $V_{h,i}$ ($n = 2$, $\theta_i = 2.5\mu_i$, $\lambda_1 = \mu_1$ and $\lambda_2 = \mu_2 = 2\mu_1$)



(a) Effect on P_f



(b) Effect on $E[C_d]$

Figure 3.6: Effects of n ($C_{min} = 6c$, $\lambda_i = i\mu_1$ and $\mu_i = i\mu_1$)

3.7 Summary

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

- $E[N_{r,i}]$ decreases as the granted credit θ_i increases. When $\theta_i \geq 2.5/\mu_i$, increasing θ_i will not improve the $E[N_{r,i}]$ performance.
- P_f decreases as the recharge threshold C_{min} or θ_i increases. This effect becomes insignificant when θ_i is large (e.g., $\theta_i \geq 5/\mu_i$). $E[C_d]$ increases as C_{min} or θ_i increases.
- P_f increases as the variance $V_{h,i}$ of the service session holding time increases, and $E[C_d]$ decreases as $V_{h,i}$ increases.
- P_f increases as the number n of the service session types increases. When θ_i is small (e.g. $\theta_i \leq 1/\mu_i$), $E[C_d]$ decreases as n increases. On the other hand, when θ_i is large (e.g. $\theta_i \geq 2.5/\mu_i$), $E[C_d]$ increases as n increases.

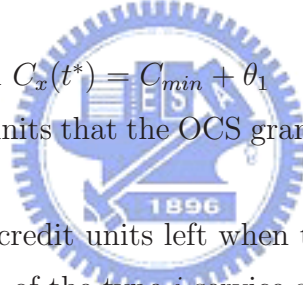
Based on the above discussion, a mobile operator can select the appropriate C_{min} and θ_i values for various traffic conditions.

3.8 Notation

The notation used in this chapter is listed below.

- C : the amount of the initial prepaid credit for a mobile user
- C_{min} : the recharge threshold in RTCR
- C_r : the amount of the remaining prepaid credit in the OCS
- $C_r(t)$: the remaining credit units in the OCS for the user at a particular time point t
- $C_x(t)$: the exact unused credit units for the user at a particular time point t

- $E[C_d]$: the expected amount of unused credit units in the user account at the end of RTCR execution (before recharging)
- $E[N_{r,i}]$: the expected number of the RU operations executed during a type- i session
- λ_i : the rate of $t_{a,i}$; i.e., $\lambda_i = 1/E[t_{a,i}]$
- μ_i : the rate of $t_{h,i}$; i.e., $\mu_i = 1/E[t_{h,i}]$
- n : the number of types of session-based IMS services
- P_A : the probability that the critical RU operation occurs when there is one active type-1 service session
- P_B : the probability that the critical RU operation occurs when there is one active type-2 service session
- P_C : the probability that the critical RU operation occurs when both type-1 and type-2 service sessions are active
- P_f : the probability that an in-progress session is forced to terminate (for all service type- i)
- p_i : the probability that a type- i service session is active at a random observation point
- t^* : the critical time when $C_x(t^*) = C_{min} + \theta_1$
- θ_i : the amount of credit units that the OCS grants in each RU operation for a type- i session
- $\tilde{\theta} + C_{min}$: the remaining credit units left when the critical RU operation arrives
- $t_{a,i}$: the inter-arrival time of the type- i service session
- $t_{h,i}$: the holding time of the type- i service session
- $t_{r,i}$: the residual holding time of the type- i service session
- t_u : the consumed credit units in the AS when a random observer arrives
- t_y : the elapsed holding time of the service session when a random observer arrives



Chapter 4

Reducing Credit Re-authorization Cost

The GPRS transport network consists of *GPRS Support Nodes* (GSNs) such as *Serving GSNs* (SGSNs) and *Gateway GSNs* (GGSNs). To start an online GPRS session, the GGSN sends a credit control request to the OCS. The OCS determines the rating and allocates the granted credit units (based on the charge for time units or data volume units) to the GGSN. During a GPRS session, a number of mid-session events, such as *Quality of Service* (QoS) change of GPRS session or change of SGSN, could dynamically affect the rating of the in-progress service. Therefore, the GGSN needs to re-authorize the granted credit units when such events occur. If the change of QoS occurs frequently, the signaling load incurred by the re-authorization procedure increases heavily. This chapter proposes a cost reduction method for credit re-authorization procedure in the OCS. Then we conduct both simulations and analytic models to investigate the performance of the proposed scheme.

4.1 Online Charging System for GPRS Sessions

As described in Section 1.3, the *Session Based Charging Function* (SBCF; see Fig. 1.2 (a) and Fig. 4.1 (a)) in the OCS is responsible for online charging of network bearer and user sessions [10]. The SBCF interacts with the *Account Balance Management Function* (ABMF; see Fig. 1.2 (c) and Fig. 4.1 (b)) to query and update the user's account through the Rc interface. The ABMF keeps the subscriber's account data and manages the account balance. At the time of writing this dissertation, the message exchanges in the Rc interface are not defined. In this chapter, we use `ABMF Request` and `ABMF Response` to

represent the message exchanges between the SBCF and the ABMF. The SBCF interacts with the *Rating Function* (RF; see Fig. 1.2 (e) and Fig. 4.1 (c)) to determine the price and tariff of the requested service through the Re interface. By using Diameter **Tariff Request/Response** and **Price Request/Response** messages described in Section 1.3, the RF handles a wide variety of rateable instances, such as data volume and session connection time. In online GPRS sessions, the Diameter credit control protocol is used for communications between the GGSN and the OCS. The OCS credit control is achieved by exchanging the Diameter *Credit Control Request* (CCR) and the *Credit Control Answer* (CCA) messages described in Section 1.4.

In a telecom network, the ABMF and the SBCF may physically reside at different (and possibly remote) locations. Therefore, the message exchanges in the Rc interface may be expensive, and it is desirable to reduce the credit re-authorization message cost.

4.1.1 Credit Re-authorization Procedure

Consider a scenario where a mobile user is viewing a streaming video through the GPRS network [17]. The streaming services are charged according to the amount of time units and the related QoS provisioned (i.e., the bandwidth allocated to the bearer session). Due to user mobility between different UMTS coverage areas, and depending on the work load of the radio network, the QoS of the streaming session may change from time to time. In terms of bandwidth allocated, we assume that there are N QoS classes for the GPRS bearer session. For $1 \leq i \leq N$, let α_i be the number of credit units charged for every time unit in a class i session, and the OCS granted $\alpha_i \tau_g$ credit units (which can last for τ_g time units) to the GGSN in each credit reservation. Note that the bandwidth allocated to a class i session is typically proportioned to the charge (i.e., α_i). Whenever the QoS of the GPRS bearer changes, the credit re-authorization procedure illustrated in Fig. 4.1 must be executed with the following steps:

Step 1. To start the streaming session with online charging, the GGSN sends the INITIAL_REQUEST CCR message to the OCS. This message indicates the QoS parameter (e.g., QoS class i) for the session.

Step 2. When the OCS receives the credit control request, the SBCF sends the **Tariff Request** message, including the QoS parameter in the *Service-Information* field, to the Rating Function. The Rating Function replies with the **Tariff Response** message to indicate the applicable tariff α_i for this session [10].

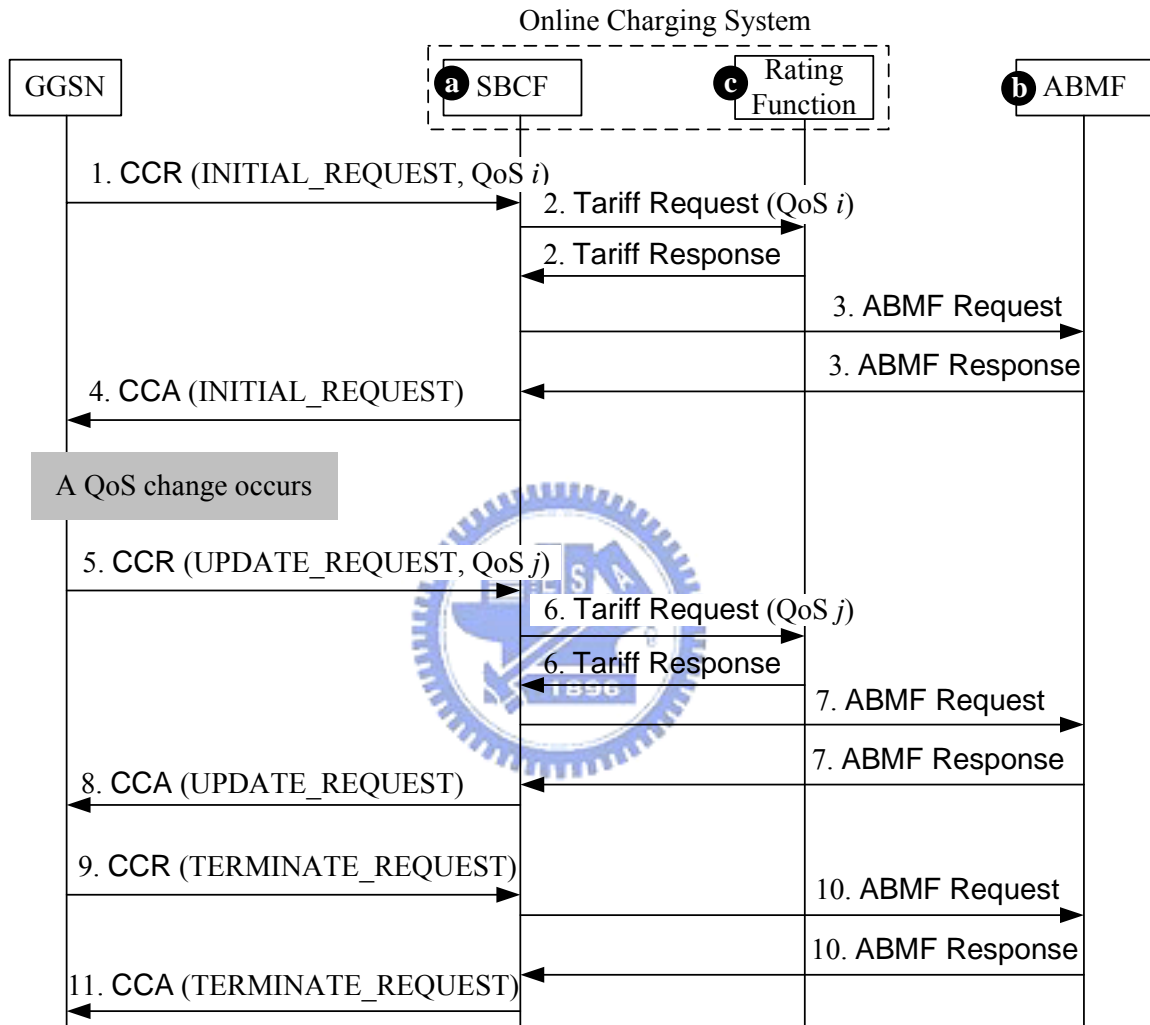


Figure 4.1: Message Flow for Credit Re-authorization Procedure.

Step 3. Based on the received tariff information, the SBCF grants $\alpha_i\tau_g$ credit units to the session (which can last for τ_g time units). By exchanging the **ABMF Request** and **Response** message pair with the ABMF, the SBCF reserves $\alpha_i\tau_g$ credit units in the subscriber's account.

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

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

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

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

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

Steps 9-11. When the video streaming service is complete, the GGSN terminates the session. The GGSN reports the used credit units to the OCS by sending the **TERMINATE_REQUEST CCR** message. The SBCF calculates the consumed credit units

and instructs the ABMF to debit the user account. Finally, the OCS acknowledges the GGSN with a CCA message.

For the discussion purpose, the above re-authorization procedure (i.e., Steps 6 and 7) in the OCS is referred to as the “basic” scheme. If the QoS of the session changes frequently, the signaling traffic incurred by the re-authorization procedure becomes significant. In this chapter, we propose a “threshold-based” scheme that utilizes a threshold parameter δ to reduce the signaling cost between the SBCF and the ABMF.

4.1.2 Threshold-based Scheme for Credit Re-authorization

In the basic scheme, the ABMF message exchanges (see Step 7 in Fig. 4.1) can be omitted if the remaining credit units are large enough to accommodate the new reservation. Based on this observation, we propose a threshold-based scheme to reduce the signaling cost. In this scheme, the SBCF determines whether to interact with the ABMF or not based on a threshold parameter δ . Under the QoS class for an in-progress GPRS session is changed from class i to class j , the ABMF message exchange is skipped if the remaining credit units $\alpha_i\tau_u$ is larger than $\delta\alpha_j\tau_g$. Details on the threshold-based scheme are described as follows:

Step 1. As described in Step 6 of Fig. 4.1, the SBCF retrieves the old tariff α_i and new tariff α_j for the bearer session from the Rating Function.

Step 2a. If $\alpha_i\tau_u \geq \delta\alpha_j\tau_g$, the SBCF directly allocates $\alpha_i\tau_u$ credit units to the GGSN. The ABMF message exchange (i.e., Step 7 in Fig. 4.1) is skipped.

Step 2b. Otherwise (i.e., $\alpha_i\tau_u < \delta\alpha_j\tau_g$), the SBCF allocates $\alpha_j\tau_g$ credit units to the GGSN. Then Step 7 in Fig. 4.1 is executed. That is, the SBCF debits $\alpha_i(\tau_g - \tau_u)$ credit units to the ABMF and makes new reservation for $\alpha_j\tau_g$ credit units.

In the OCS, a mobile operator (or a user) can check the account balance anytime. When there is no in-progress session, the OCS can accurately report the account balance of the user. On the other hand, when some credit units are reserved, the account balance reported by the OCS may not be up-to-date because some reserved credit units might already be used. When a “balance check” occurs, the OCS reports the account balance including the reserved credit units. Note that this reported value may be larger than the actual balance. Denote C as the inaccuracy of credit information, which is the difference

between the balance stored in the OCS (including the reserved credit units) and the actual balance (excluding the credit units already consumed by the GGSN). In other words, C is the credit units consumed by the GGSN between when the previous ABMF message exchange occurs and when the balance check occurs. In the threshold-based scheme, the inaccuracy of credit information C may increase because it skips some ABMF message exchanges. Therefore, it is important to select appropriate τ_g and δ values to optimize the performance of the threshold-based scheme in terms of the following output measures:

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

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

4.2 Analytic Modeling for the Basic Scheme

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

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

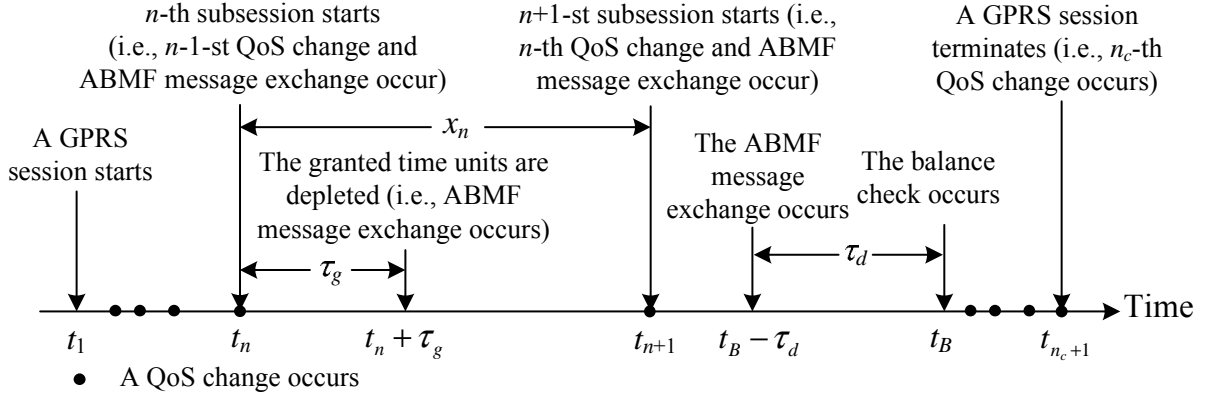


Figure 4.2: Timing Diagram for the Basic Scheme.

independent and identically distributed (i.i.d.) exponential random variable with mean $E[x_n] = 1/\lambda$.

In Fig. 4.2, a mobile user starts a GPRS streaming session at t_1 and terminates at t_{n_c+1} , where n_c QoS changes occur during the streaming session (including the last QoS change that terminates the session). These QoS changes partition the streaming session into n_c subsessions. In Fig. 4.2, the n -th subsession starts at t_n , where $1 \leq n \leq n_c$. The holding time of the n -th subsession is $x_n = t_{n+1} - t_n$. We assume that x_n ($1 \leq n \leq n_c$) are independent and identically distributed (i.i.d.) exponential random variables with mean $E[x_n] = 1/\lambda$.

4.2.1 Derivation for the Number of ABMF message Exchanges

After a subsession is complete, the GPRS session terminates with probability p_0 . Therefore the number n_c of the subsessions within a GPRS session has a geometric distribution, where

$$E[n_c] = \sum_{n_c=1}^{\infty} n_c (1 - p_0)^{n_c-1} p_0 = \frac{1}{p_0} \quad (4.1)$$

Let m_B be the number of ABMF message exchanges in a subsession for the basic scheme. $E[m_B]$ is derived as follows: In Fig. 4.2, the n -th subsession starts at t_n with QoS class i . The ABMF message exchange is always executed and the GGSN is granted $\alpha_i \tau_g$ credit units by the OCS. When these $\alpha_i \tau_g$ credit units are depleted at $t_n + \tau_g$, the GGSN requests next $\alpha_i \tau_g$ credit units from the OCS. The credit control requests issued within a subsession (excluding the request at the beginning of the subsession) can be modeled as a Poisson

stream with arrival rate μ . Applying Little's formula [34, 38], the expected number $E[m_B]$ for a subsession can be derived as

$$E[m_B] = 1 + \mu E[x_n] = 1 + \frac{\mu}{\lambda} \quad (4.2)$$

From (4.1) and (4.2),

$$M_B = E[n_c]E[m_B] = \frac{\mu + \lambda}{p_0\lambda} \quad (4.3)$$

4.2.2 Derivation for the Inaccuracy of Credit Information

This subsection derives the expected inaccuracy of credit information $E[C_B]$ when a balance check occurs in the basic scheme. In the long term behavior, the probabilities that a subsession starts with class i (for $1 \leq i \leq N$) are equally likely. Therefore, the expected credit units α consumed in one time unit can be computed as

$$\alpha = \left(\frac{1}{N}\right) \sum_{i=1}^N \alpha_i \quad (4.4)$$

In Fig. 4.2, assume that a balance check occurs at t_B and the previous ABMF message exchange occurs at $t_B - \tau_d$. Note that if no ABMF message exchange occurs in (t_{n+1}, t_B) , then $\tau_d = t_B - t_{n+1}$. $E[\tau_d]$ is derived as follows: In the basic scheme, the ABMF message exchange occurs either when a QoS change occurs (which is a Poisson stream with rate λ) or when the granted credit units are depleted (which is a Poisson stream with rate μ). Based on the superposition property of the Poisson process [40], the ABMF message stream is a Poisson process with rate $\mu + \lambda$. Assume that a balance check is a random observation point between two ABMF message exchanges. From the reverse residual life theorem [48], the expected value of τ_d is

$$E[\tau_d] = \frac{1}{\mu + \lambda} \quad (4.5)$$

From (4.4) and (4.5), $E[C_B]$ is expressed as

$$E[C_B] = \alpha E[\tau_d] = \left[\frac{1}{N(\mu + \lambda)} \right] \left(\sum_{i=1}^N \alpha_i \right) \quad (4.6)$$

4.3 Analytic Modeling for the Threshold-based Scheme

To analytically model the threshold-based scheme, we assume that $\delta = 0$ and $p_0 \rightarrow 0$ such that the stationary behavior for the ABMF message exchanges during a subsession can be

observed. For $p_0 \rightarrow 0$, it is clear that the output measure $M_T \rightarrow \infty$. Therefore, we shall derive the expected number m_T of the ABMF message exchanges for a subsession. This new measure will be used to partially validate our simulation model. Then we will use the simulation experiments to investigate M_T in Section 4.5. To simplify our discussion, the number of QoS classes is restricted to $N = 2$. These two QoS classes represent the low and the high bandwidth classes, respectively. The analytic model can be directly extended for $N > 2$.

4.3.1 Derivation for the Number of ABMF message Exchanges

The number m_T of the ABMF message exchanges in a subsession is determined by two factors: (i) the QoS class of the subsession and (ii) the remaining credits left at the end of the previous subsession. Consider the timing diagram in Fig. 4.3, where the n -th subsession starts at t_n . Let J_n be the QoS class of the n -th subsession. Since $N = 2$, a GPRS bearer session consists of subsessions with QoS class 1 and class 2 alternatively. Suppose that during subsession n , the granted credit units are depleted, and an ABMF message exchange occurs at $t_{n+1} - \tau_r$ (note that if $t_n < t_{n+1} - \tau_r$ as illustrated in Fig. 4.3, the QoS class is not changed). At t_{n+1} , the $n + 1$ -st subsession starts and the QoS class is changed from J_n to J_{n+1} . In Fig. 4.3, $J_n = 1$ and $J_{n+1} = 2$. At t_{n+1} , the SBCF retrieves the new tariff (i.e., α_2) from the Rating Function. Following the threshold-based scheme for $\delta = 0$, the SBCF directly assigns the remaining credit units to the new subsession without interacting with the ABMF. Suppose that the remaining credit units are not depleted in subsession $n + 1$. Therefore, no ABMF message exchange occurs during $[t_{n+1}, t_{n+2}]$. At t_{n+2} , the $n + 2$ -nd subsession starts and the QoS is changed to $J_{n+2} = 1$. For every J_n , we define a corresponding random variable I_n that represents the QoS class of the subsession immediately before the last ABMF message exchange occurs. In Fig. 4.3, $I_{n+1} = I_{n+2} = 1$. When subsession n starts, the amount of credit units left is affected by I_n . Define $Z_n = (I_n, J_n)$ as a two dimensional random variable. In Fig. 4.3, $Z_{n+1} = (1, 2)$ and $Z_{n+2} = (1, 1)$ for subsessions $n + 1$ and $n + 2$, respectively. It is clear that the process $\{Z_n, n \geq 1\}$ is a Markov chain.

Let $\pi_{(i,j)} = \lim_{n \rightarrow \infty} \Pr[Z_n = (i, j)]$ be the stationary distribution of the Markov chain, where $i, j \in \{1, 2\}$. Since the subsession holding time x_n is exponentially distributed, the QoS change occurrences form a Poisson stream. From Poisson Arrival See Time Average (PASTA) [34], $\pi_{(i,j)}$ also represents the proportion of time that the subsession resides

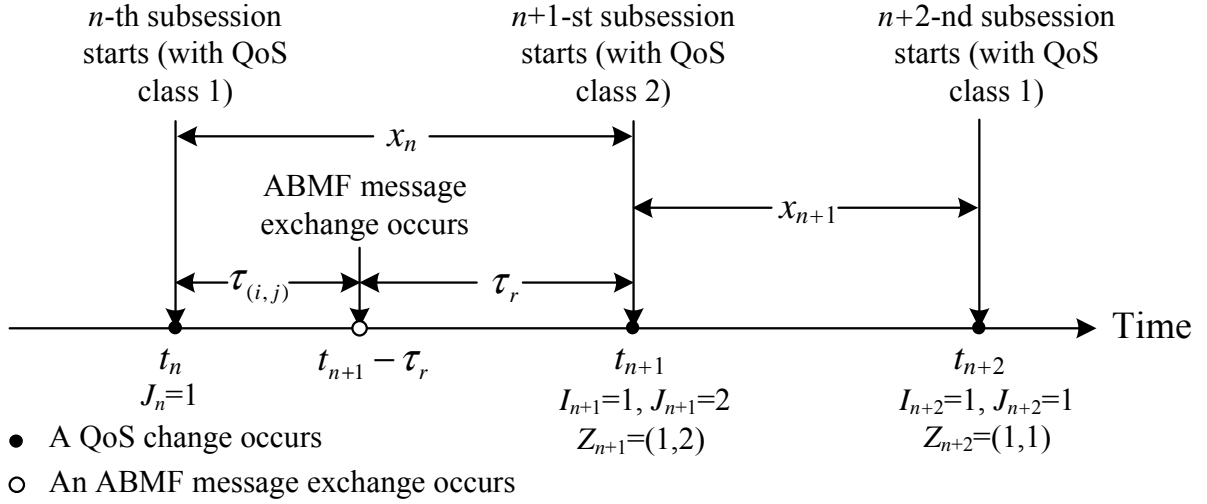


Figure 4.3: Timing Diagram for the Threshold-based Scheme.

in state (i, j) . Let $m_{(i,j)}$ be the expected number of ABMF message exchanges for a subsession in state (i, j) . Then m_T can be computed as

$$m_T = \sum_{i,j \in \{1,2\}} m_{(i,j)} \pi_{(i,j)} \quad (4.7)$$

To derive $\pi_{(i,j)}$, we first discuss the state transitions for the Markov chain $\{Z_n, n \geq 1\}$. There are four situations.

Situation I: $Z_n = (1, 2)$. If no ABMF message exchange occurs in this subsession, then $I_{n+1} = I_n = 1$ and $Z_{n+1} = (1, 1)$. Assume that the process moves from state $(1, 2)$ to state $(1, 1)$ with transition probability p_a . If any ABMF message exchange occurs in subsession n with QoS class 2, then I_{n+1} changes to 2. The Markov chain moves from state $(1, 2)$ to state $(2, 1)$ with probability $1 - p_a$.

Situation II: $Z_n = (2, 1)$. If no ABMF message exchange occurs in this subsession, then $I_{n+1} = I_n = 2$ and $Z_{n+1} = (2, 2)$. Assume that the process moves from state $(2, 1)$ to state $(2, 2)$ with transition probability p_b . If any ABMF message exchange occurs in subsession n with QoS class 1, then I_{n+1} changes to 1. The Markov chain moves from state $(2, 1)$ to state $(1, 2)$ with probability $1 - p_b$.

Situation III: $Z_n = (1, 1)$. During this subsession, whether an ABMF message exchange occurs or not, $I_{n+1} = I_n = 1$. Therefore, the Markov chain moves from state $(1, 1)$ to state $(1, 2)$ with probability 1.

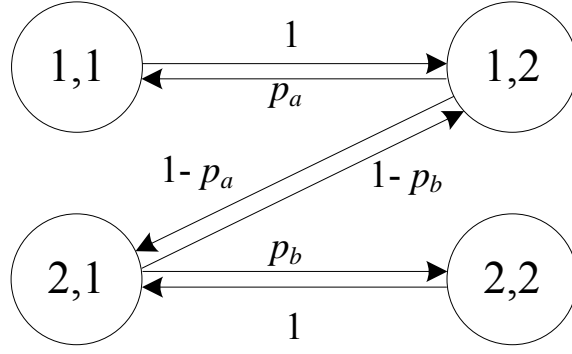


Figure 4.4: Probability Transition Diagram ($N = 2$).

Situation IV: $Z_n = (2, 2)$. Similar to Situation III, $I_{n+1} = I_n = 2$. The Markov chain moves from state $(2, 2)$ to state $(2, 1)$ with probability 1.

From the probability transition diagram in Fig. 4.4, the limiting probabilities are derived as

$$\left. \begin{aligned} \pi_{(1,1)} &= p_a \left[1 + p_a + \frac{(1-p_a)(1+p_b)}{1-p_b} \right]^{-1}, & \pi_{(2,1)} &= \left[1 + p_b + \frac{(1-p_b)(1+p_a)}{1-p_a} \right]^{-1} \\ \pi_{(1,2)} &= \left[1 + p_a + \frac{(1-p_a)(1+p_b)}{1-p_b} \right]^{-1}, & \pi_{(2,2)} &= p_b \left[1 + p_b + \frac{(1-p_b)(1+p_a)}{1-p_a} \right]^{-1} \end{aligned} \right\} \quad (4.8)$$

In (4.8), the transition probabilities p_a and p_b are derived as follows: Let $\alpha_i \tau_u$ be the credit units left at the end of the previous subsession. As illustrated in Fig. 4.3, suppose that the current subsession is in state (i, j) and $\alpha_j \tau_{(i,j)}$ credit units are assigned to this subsession, where

$$\tau_{(i,j)} = \alpha_i \tau_u / \alpha_j \quad (4.9)$$

The subsession holding time $x_n = t_{n+1} - t_n$ has the density function

$$f_x(x_n) = \lambda e^{-\lambda x_n} \quad (4.10)$$

and $x_n = \tau_{(i,j)} + \tau_r$. From (4.9), p_a and p_b can be expressed as

$$p_a = \Pr[\tau_{(1,2)} > x_n] = \Pr[\alpha_1 \tau_u / \alpha_2 > x_n] \text{ and } p_b = \Pr[\tau_{(2,1)} > x_n] = \Pr[\alpha_2 \tau_u / \alpha_1 > x_n] \quad (4.11)$$

Since τ_g is exponentially distributed with rate μ , and from the memoryless property [48], τ_u is also exponential distributed with the density function

$$f_u(\tau_u) = \mu e^{-\mu \tau_u} \quad (4.12)$$

Let $\hat{\alpha} = \alpha_2/\alpha_1$, then $\tau_{(1,2)} = \tau_u/\hat{\alpha}$ and $\tau_{(2,1)} = \hat{\alpha}\tau_u$. From (4.10)-(4.12), the transition probability p_a from state (1, 2) to state (1, 1) is derived as

$$p_a = \Pr[\tau_u/\hat{\alpha} > x_n] = \int_{x_n=0}^{\infty} f_x(x_n) \int_{\tau_u=\hat{\alpha}x_n}^{\infty} f_u(\tau_u) d\tau_u dx_n = \frac{\lambda}{\lambda + \hat{\alpha}\mu} \quad (4.13)$$

Similarly, the transition probability p_b from state (2, 1) to state (2, 2) can be derived as

$$p_b = \Pr[\hat{\alpha}\tau_u > x_n] = \frac{\hat{\alpha}\lambda}{\hat{\alpha}\lambda + \mu} \quad (4.14)$$

Substituting (4.13) and (4.14) into (4.8), the limiting probabilities for $\{Z_n, n > 0\}$ is re-written as

$$\left. \begin{aligned} \pi_{(1,1)} &= \frac{\lambda}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} & \pi_{(2,1)} &= \frac{\hat{\alpha}(\hat{\alpha}\lambda + \mu)}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} \\ \pi_{(1,2)} &= \frac{\lambda + \hat{\alpha}\mu}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} & \pi_{(2,2)} &= \frac{\hat{\alpha}^2\lambda}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} \end{aligned} \right\} \quad (4.15)$$

$m_{(i,j)}$ in (4.7) is derived as follows: For $i = j$, where $\tau_{(i,i)} = \tau_u$ is exponentially distributed with rate μ . When $\alpha_i\tau_{(i,i)}$ credit units are depleted, an ABMF message exchange occurs and the SBCF grants another $\alpha_i\tau_g$ credit units to the subsession. Therefore, the ABMF message arrivals during a subsession period x_n are a Poisson stream with rate μ . By Little's formula [34], $m_{(i,i)}$ is expressed as

$$m_{(i,i)} = \mu E[x_n] = \frac{\mu}{\lambda} \quad (4.16)$$

$m_{(2,1)}$ is derived as follows: Consider Situation II (i.e., $Z_n = (2, 1)$). As shown in Fig. 4.3, assume that the credit units $\alpha_1\tau_{(2,1)}$ are depleted at time $t_{n+1} - \tau_r$ and an ABMF message exchange occurs. From the memoryless property [48], τ_r is also exponential distributed with mean $E[\tau_r] = 1/\lambda$. In period τ_r , the ABMF message arrivals are a Poisson stream with rate μ . From the Little's formula, the expected number of ABMF message exchanges is $1 + \mu E[\tau_r]$. Since the probabilities that no ABMF message exchange in subsession n (i.e., $Z_{n+1} = (2, 2)$) or at least one ABMF message exchange occurs during subsession n (i.e., $Z_{n+1} = (1, 2)$) are denoted as p_b and $1 - p_b$, respectively. Therefore, $m_{(2,1)}$ is expressed as

$$m_{(2,1)} = p_b \times 0 + (1 - p_b)(1 + \mu E[\tau_r]) = \left(\frac{\mu}{\hat{\alpha}\lambda + \mu} \right) \left(1 + \frac{\mu}{\lambda} \right) \quad (4.17)$$

Similarly,

$$m_{(1,2)} = p_a \times 0 + (1 - p_a)(1 + \mu E[\tau_r]) = \left(\frac{\hat{\alpha}\mu}{\lambda + \hat{\alpha}\mu} \right) \left(1 + \frac{\mu}{\lambda} \right) \quad (4.18)$$

Substituting (4.15)-(4.18) into (4.7) to yield

$$m_T = \frac{2\hat{\alpha}\lambda\mu + \mu(\lambda + 2\hat{\alpha}\mu + \hat{\alpha}^2\lambda)}{2\lambda(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} \quad (4.19)$$

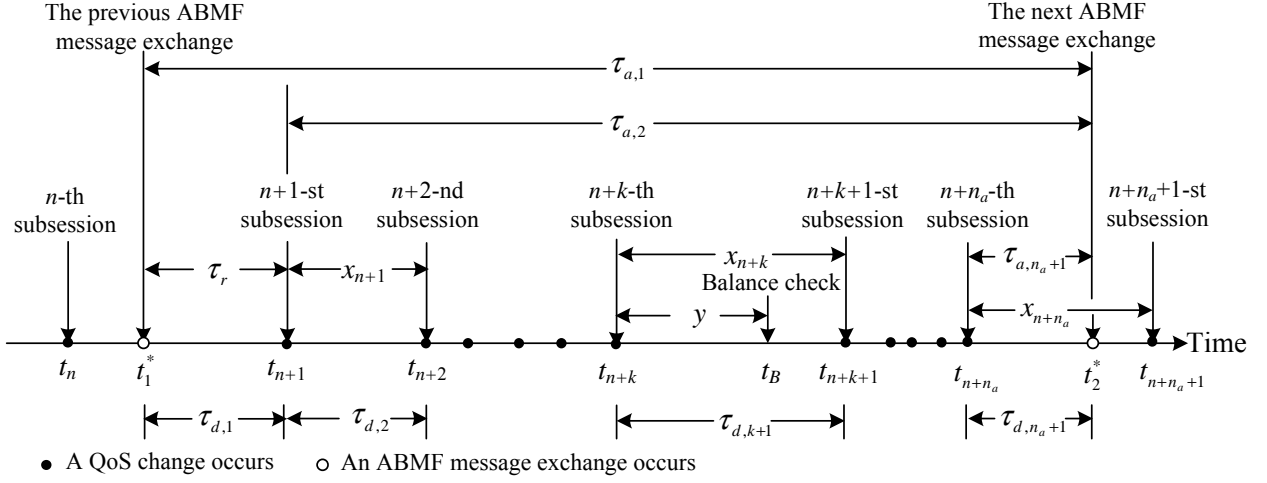


Figure 4.5: Timing Diagram for Deriving C_T .

4.3.2 Derivation for the Inaccuracy of Credit Information

This subsection derives the inaccuracy of credit information C_T in the threshold-based scheme. In Fig. 4.5, an ABMF message exchange occurs at t_1^* in subsession n . Suppose that the granted credit units are consumed at t_2^* in subsession $n + n_a$. Note that if $n_a = 0$, two consecutive ABMF message exchanges occur in subsession n . In $[t_1^*, t_2^*]$, a balance check occurs at t_B in subsession $n + k$, where $0 \leq k \leq n_a$. Let K be the QoS class of the session at t_1^* , and let $C_{T,i}$ be the C_T value under the condition that $K = i$. By considering whether $K = 1$ or $K = 2$, we can express C_T consumed in $[t_1^*, t_B]$ as

$$C_T = \sum_{i=1}^2 C_{T,i} \Pr[K = i] \quad (4.20)$$

In Fig. 4.5, if the previous ABMF message exchange occurs at $t_1^* \in [t_n, t_B]$, then $K = J_n$. When $k > 0$, $J_n = I_{n+k}$, and we can rewrite $K = I_{n+k}$. Therefore, $\Pr[K = i]$ can be derived in the following four situations.

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

Situation VI: $Z_{n+k} = (2, 1)$. Since the balance check is a random observer of subsession $n + k$, and from the reverse residual life theorem [48], $t_B - t_{n+k}$ is also exponential distributed with rate λ . Therefore, the probability that no ABMF message occurring in period $[t_{n+k}, t_B]$ (i.e., $K = 2$) is p_b . On the other hand, if any ABMF message occurs in period $[t_{n+k}, t_B]$, we have $K = 1$ (with probability $1 - p_b$).

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

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

Let $\pi_{(i,j)}$ be the probability that when a random observer arrives (i.e., the balance check occurs) at t_B , and the $n+k$ -th subsession is in state (i, j) . Based on the above discussion, and from (4.13)-(4.15), we have

$$\left. \begin{aligned} \Pr[K = 1] &= \pi_{(1,1)} + p_a\pi_{(1,2)} + (1 - p_b)\pi_{(2,1)} = \frac{2\lambda + \hat{\alpha}\mu}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} \\ \Pr[K = 2] &= \pi_{(2,2)} + p_b\pi_{(2,1)} + (1 - p_a)\pi_{(1,2)} = \frac{2\hat{\alpha}^2\lambda + \hat{\alpha}\mu}{2(\lambda + \hat{\alpha}\mu + \hat{\alpha}^2\lambda)} \end{aligned} \right\} \quad (4.21)$$

In (4.20), $C_{T,i}$ is derived as follows: As illustrated in Fig. 4.5, $\tau_r = t_{n+1} - t_1^*$. From the memoryless property [48], τ_r has the density function

$$f_r(\tau_r) = \lambda e^{-\lambda\tau_r} \quad (4.22)$$

In Fig. 4.5, the remaining granted credit units at t_1^* lasts for $\tau_{a,1}$ time units. Similarly, the remaining granted credit units at t_{n+l-1} ($2 \leq l \leq n_a + 1$) can last for $\tau_{a,l}$ time units. That is

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

Let

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

In Fig. 4.5, $\tau_{d,1} = \tau_r$, $\tau_{d,l} = x_{n+l-1}$ for $2 \leq l \leq n_a$, and $\tau_{d,n_a+1} = \tau_{a,n_a+1}$. Assume that a balance check occurs at t_B in subsession $n+k$. Let $y = t_B - \max(t_1^*, t_{n+k})$, then $t_B - t_1^* = \sum_{l=1}^k \tau_{d,l} + y$. Denote $P_{i,k}$ as the probability that exactly k QoS changes occur in $[t_1^*, t_B]$ under the condition that $K = i$. Let $\alpha_{i,l} = \alpha_{(i+l \bmod 2)+1}$ for $i \in \{1, 2\}$, $C_{T,i}$ can be expressed as

$$\begin{aligned} C_{T,i} &= P_{i,0}\alpha_{i,1}E[y|K = i, t_1^* < t_B < t_{n+1}] + \sum_{k=1}^{\infty} P_{i,k} \\ &\times \left\{ \sum_{l=1}^k \alpha_{i,l}E[\tau_{d,l}|K = i, t_{n+l} < t_2^*] + \alpha_{i,k+1}E[y|K = i, t_{n+k} < t_B < t_{n+k+1}] \right\} \end{aligned} \quad (4.23)$$

In (4.23), $E[\tau_{d,l}|K = i, t_{n+l} < t_2^*]$ is derived as follows: For $K = i$, let $f_{i,l}(\tau_{a,l})$ be the density function of $\tau_{a,l}$. We have

$$f_{1,l}(\tau_{a,l}) = \begin{cases} \mu e^{-\mu\tau_{a,l}}, & l \text{ is odd} \\ \hat{\alpha}\mu e^{-\hat{\alpha}\mu\tau_{a,l}}, & l \text{ is even} \end{cases} \quad \text{and} \quad f_{2,l}(\tau_{a,l}) = \begin{cases} \mu e^{-\mu\tau_{a,l}}, & l \text{ is odd} \\ \frac{\mu}{\hat{\alpha}} e^{-\frac{\mu}{\hat{\alpha}}\tau_{a,l}}, & l \text{ is even} \end{cases} \quad (4.24)$$

Eq. (4.24) is explained as follows: For $K = i$ and when l is odd, subsession $n + l - 1$ is served with QoS class i . Therefore, $f_{i,l}(\tau_{a,l})$ is exponentially distributed with rate μ . When l is even, subsession $n + l - 1$ is served with QoS classes 2 and 1 for $K = 1$ and $K = 2$, respectively. Therefore, $f_{i,l}(\tau_{a,l})$ is exponentially distributed with rates $\hat{\alpha}\mu$ and $\mu/\hat{\alpha}$, respectively. From (4.22) and (4.24), $E[\tau_{d,1}|K = i, t_{n+1} < t_2^*]$ can be derived as

$$\begin{aligned} E[\tau_{d,1}|K = i, t_{n+1} < t_2^*] &= E[\tau_r|\tau_r < \tau_{a,1}, K = i] \\ &= \frac{\int_{\tau_r=0}^{\infty} \tau_r f_r(\tau_r) \int_{\tau_{a,1}=\tau_r}^{\infty} f_{i,1}(\tau_{a,1}) d\tau_{a,1} d\tau_r}{\int_{\tau_r=0}^{\infty} f_r(\tau_r) \int_{\tau_{a,1}=\tau_r}^{\infty} f_{i,1}(\tau_{a,1}) d\tau_{a,1} d\tau_r} \end{aligned} \quad (4.25)$$

For $l \geq 2$, and from (4.10) and (4.24), $E[\tau_{d,l}|K = i, t_{n+l} < t_2^*]$ can be derived as

$$\begin{aligned} &E[\tau_{d,l}|K = i, t_{n+l} < t_2^*] \\ &= E[x_{n+l-1}|x_{n+l-1} < \tau_{a,l}, K = i] \\ &= \frac{\int_{x_{n+l-1}=0}^{\infty} x_{n+l-1} f_x(x_{n+l-1}) \int_{\tau_{a,l}=x_{n+l-1}}^{\infty} f_{i,l}(\tau_{a,l}) d\tau_{a,l} dx_{n+l-1}}{\int_{x_{n+l-1}=0}^{\infty} f_x(x_{n+l-1}) \int_{\tau_{a,l}=x_{n+l-1}}^{\infty} f_{i,l}(\tau_{a,l}) d\tau_{a,l} dx_{n+l-1}} \end{aligned} \quad (4.26)$$

Combining (4.25) and (4.26), $E[\tau_{d,l}|K = i, t_{n+l} < t_2^*]$ can be expressed as

$$E[\tau_{d,l}|K = i, t_{n+l} < t_2^*] = \begin{cases} \frac{1}{\lambda+\mu}, & l \text{ is odd} \\ \frac{1}{\lambda+\hat{\alpha}\mu}, & K = 1 \text{ and } l \text{ is even} \\ \frac{1}{\lambda+\mu/\hat{\alpha}}, & K = 2 \text{ and } l \text{ is even} \end{cases} \quad (4.27)$$

In (4.23), $E[y|K = i, t_1^* < t_B < t_{n+1}]$ and $E[y|K = i, t_{n+k} < t_B < t_{n+k+1}]$ are derived as follows: Since the balance check is a random observation point of $\tau_{d,k+1}$, and from reverse residual life theorem for the exponential random variables [48], $E[y|K = i, \max(t_1^*, t_{n+k}) < t_B < t_{n+k+1}] = E[\tau_{d,k+1}]$. If $t_1^* < t_B < t_{n+1}$, we have

$$\begin{aligned} &E[y|K = i, t_1^* < t_B < t_{n+1}] \\ &= \int_{\tau_r=0}^{\infty} \tau_r f_r(\tau_r) \int_{\tau_{a,1}=\tau_r}^{\infty} f_{i,1}(\tau_{a,1}) d\tau_{a,1} d\tau_r + \int_{\tau_{a,1}=0}^{\infty} \tau_{a,1} f_{i,1}(\tau_{a,1}) \int_{\tau_r=\tau_{a,1}}^{\infty} f_r(\tau_r) d\tau_r d\tau_{a,1} \end{aligned} \quad (4.28)$$

For $t_{n+k} < t_B < t_{n+k+1}$, we have

$$\begin{aligned}
& E[y|K = i, t_{n+k} < t_B < t_{n+k+1}] \\
&= \int_{x_{n+k}=0}^{\infty} x_{n+k} f_x(x_{n+k}) \int_{\tau_{a,k+1}=x_{n+k}}^{\infty} f_{i,k+1}(\tau_{a,k+1}) d\tau_{a,k+1} dx_{n+k} \\
&+ \int_{\tau_{a,k+1}=0}^{\infty} \tau_{a,k+1} f_{i,k+1}(\tau_{a,k+1}) \int_{x_{n+k}=\tau_{a,k+1}}^{\infty} f_x(x_{n+k}) dx_{n+k} d\tau_{a,k+1} \quad (4.29)
\end{aligned}$$

Combining (4.28) and (4.29), $E[y|K = i, \max(t_1^*, t_{n+k}) < t_B < t_{n+k+1}]$ can be derived as

$$\begin{aligned}
E[y|K = i, \max(t_1^*, t_{n+k}) < t_B < t_{n+k+1}] &= E[\tau_{d,k+1}|K = i] \\
&= \begin{cases} \frac{1}{\lambda+\mu}, & k \text{ is even} \\ \frac{1}{\lambda+\hat{\alpha}\mu}, & K = 1 \text{ and } k \text{ is odd} \\ \frac{1}{\lambda+\mu/\hat{\alpha}}, & K = 2 \text{ and } k \text{ is odd} \end{cases} \quad (4.30)
\end{aligned}$$

In (4.23), $P_{i,k}$ is the probability that $\max(t_1^*, t_{n+k}) \leq t_B < \min(t_{n+k+1}, t_2^*)$ under the condition that $K = i$ (see Fig. 4.5). To derive $P_{i,k}$, we consider $[t_1^*, t_2^*]$ as a renewal interval and define a continuous-time stochastic process $\Delta(t)$, where

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

At time t , there are exactly $\Delta(t) - 1$ QoS changes occur in $[t_1^*, t]$. Therefore, $P_{i,k} = \Pr[\Delta(t_B) = k + 1|K = i]$. In Fig. 4.5, $t_2^* > t_{n+k}$ and the length of time such that $\Delta(t) = k + 1$ is $\tau_{d,k+1}$. Since t_B is uniformly distributed over $[t_1^*, t_2^*]$, and from the alternating renewal theory [48], $P_{i,k}$ can be computed as

$$P_{i,k} = \Pr[\Delta(t_B) = k + 1|K = i] = \frac{\Pr[t_2^* > t_{n+k}|K = i] E[\tau_{d,k+1}|K = i]}{E[t_2^* - t_1^*|K = i]} \quad (4.31)$$

In (4.31), $\Pr[t_2^* > t_{n+k}|K = i]$ is derived as follows: For $k = 0$, it is clear that $\Pr[t_2^* > t_n|K = i] = 1$. For $k > 1$, we have

$$\begin{aligned}
\Pr[t_2^* > t_{n+k}|K = i] &= \Pr[\tau_{a,1} > \tau_r|K = i] \prod_{l=2}^k \Pr[\tau_{a,l} > x_{n+l-1}|K = i] \\
&= \left[\int_{\tau_r=0}^{\infty} \int_{\tau_{a,1}=\tau_r}^{\infty} f_r(\tau_r) f_{i,1}(\tau_{a,1}) d\tau_{a,1} d\tau_r \right] \\
&\quad \times \prod_{l=2}^k \left[\int_{x_{n+l-1}=0}^{\infty} \int_{\tau_{a,l}=x_{n+l-1}}^{\infty} f_x(x_{n+l-1}) f_{i,l}(\tau_{a,l}) d\tau_{a,l} dx_{n+l-1} \right] \quad (4.32)
\end{aligned}$$

From (4.10), (4.22) and (4.24), (4.32) can be expressed as

$$\Pr[t_2^* > t_{n+k} | K = i] = \begin{cases} \frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil} (\lambda+\hat{\alpha}\mu)^{\lfloor k/2 \rfloor}}, & K = 1 \\ \frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil} (\lambda+\mu/\hat{\alpha})^{\lfloor k/2 \rfloor}}, & K = 2 \end{cases} \quad (4.33)$$

In Fig. 4.5, $t_2^* - t_1^* = \sum_{l=1}^{n_a+1} \tau_{d,l}$. Therefore, the expected length of renewal interval $[t_1^*, t_2^*]$ can be computed as

$$\begin{aligned} E[t_2^* - t_1^* | K = i] &= \sum_{l=0}^{\infty} \Pr[n_a = l | K = i] E[t_2^* - t_1^* | K = i, n_a = l] \\ &= \sum_{l=0}^{\infty} \Pr[n_a \geq l | K = i] E[\tau_{d,l+1} | K = i] \\ &= \sum_{l=1}^{\infty} \Pr[t_2^* > t_{n+l-1} | K = i] E[\tau_{d,l} | K = i] \end{aligned} \quad (4.34)$$

Following the derivation for (4.32) and (4.33), we have

$$\Pr[t_2^* > t_{n+l-1} | K = i] = \begin{cases} \frac{\lambda^{l-1}}{(\lambda+\mu)^{\lceil (l-1)/2 \rceil} (\lambda+\hat{\alpha}\mu)^{\lfloor (l-1)/2 \rfloor}}, & K = 1 \\ \frac{\lambda^{l-1}}{(\lambda+\mu)^{\lceil (l-1)/2 \rceil} (\lambda+\mu/\hat{\alpha})^{\lfloor (l-1)/2 \rfloor}}, & K = 2 \end{cases} \quad (4.35)$$

Following the derivation for (4.28)-(4.30), we have

$$E[\tau_{d,l} | K = i] = \begin{cases} \frac{1}{\lambda+\mu}, & l \text{ is odd} \\ \frac{1}{\lambda+\hat{\alpha}\mu}, & K = 1 \text{ and } l \text{ is even} \\ \frac{1}{\lambda+\mu/\hat{\alpha}}, & K = 2 \text{ and } l \text{ is even} \end{cases} \quad (4.36)$$

Substituting (4.35) and (4.36) into (4.34) to yield

$$E[t_2^* - t_1^* | K = i] = \begin{cases} \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)} \right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)} \right]^{-1}, & K = 1 \\ \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})} \right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})} \right]^{-1}, & K = 2 \end{cases} \quad (4.37)$$

From (4.30), (4.33) and (4.37), (4.31) can be expressed as

$$P_{1,k} = \begin{cases} \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)} \right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)} \right] \left[\frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil} (\lambda+\hat{\alpha}\mu)^{\lfloor k/2 \rfloor + 1}} \right], & k \text{ is odd} \\ \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)} \right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\hat{\alpha}\mu)} \right] \left[\frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil + 1} (\lambda+\hat{\alpha}\mu)^{\lfloor k/2 \rfloor}} \right], & k \text{ is even} \end{cases} \quad (4.38)$$

and

$$P_{2,k} = \begin{cases} \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})} \right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})} \right] \left[\frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil} (\lambda+\mu/\hat{\alpha})^{\lfloor k/2 \rfloor + 1}} \right], & k \text{ is odd} \\ \left[\frac{1}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})} \right]^{-1} \left[1 - \frac{\lambda^2}{(\lambda+\mu)(\lambda+\mu/\hat{\alpha})} \right] \left[\frac{\lambda^k}{(\lambda+\mu)^{\lceil k/2 \rceil + 1} (\lambda+\mu/\hat{\alpha})^{\lfloor k/2 \rfloor}} \right], & k \text{ is even} \end{cases} \quad (4.39)$$

Substituting (4.27), (4.30), (4.38) and (4.39) in (4.23), $C_{T,i}$ can be obtained. Finally, C_T can be computed from Eqs. (4.20), (4.21) and (4.23).

Table 4.1: Comparison of the Analytic and Simulation Results ($\delta = 0$, $N = 2$ and $\hat{\alpha} = 2$)

τ_g (Unit: $1/\lambda$)	M_B			C_B (unit: $1/\lambda$)		
	Sim.	Ana.	Error	Sim.	Ana.	Error
0.1	1100.66	1100.0	-0.06%	0.1364	0.1364	0.00%
0.25	500.55	500.0	-0.11%	0.3001	0.3000	-0.03%
0.5	298.03	300.0	0.66%	0.4996	0.5000	0.09%
0.75	232.39	233.3	0.40%	0.6429	0.6429	0.00%
1	198.89	200.0	0.56%	0.7499	0.7500	0.01%
2.5	140.08	140.0	-0.05%	1.0712	1.0714	0.02%
5	119.49	120.0	0.42%	1.2507	1.2500	-0.05%
7.5	112.51	113.3	0.73%	1.3234	1.3235	0.01%
10	109.29	110.0	0.64%	1.3636	1.3636	0.01%

(a) The Basic Scheme ($p_0 = 0.01$)

τ_g (Unit: $1/\lambda$)	m_T			C_T (unit: $1/\lambda$)		
	Sim.	Ana.	Error	Sim.	Ana.	Error
0.1	9.8007	9.8000	-0.01%	0.1604	0.1604	-0.01%
0.25	3.8461	3.8462	0.00%	0.4188	0.4188	0.01%
0.5	1.8884	1.8889	0.03%	0.8620	0.8619	-0.01%
0.75	1.2465	1.2464	-0.01%	1.3086	1.3089	0.02%
1	0.9283	0.9286	0.03%	1.7579	1.7571	-0.04%
2.5	0.3654	0.3655	0.03%	4.4546	4.4526	-0.05%
5	0.1814	0.1815	0.03%	8.9552	8.9502	-0.06%
7.5	0.1207	0.1207	0.03%	13.4434	13.4491	0.04%
10	0.0904	0.0904	0.01%	17.9514	17.9486	-0.02%

(b) The Threshold-based Scheme ($p_0 \rightarrow \infty$)

4.4 Simulation Validation

The analytic model developed in this section is used to validate against the simulation experiments. The input parameter τ_g and the output measures C_B and C_T are normalized by the mean $1/\lambda$ of the subsession holding time. The discrepancies between analytic analysis (specifically, Eqs. (4.3), (4.6), (4.19) and (4.20)) and simulation are within 1% in Table 4.1. The simulation model follows the discrete event approach described in Chapter 2, and the details are omitted.

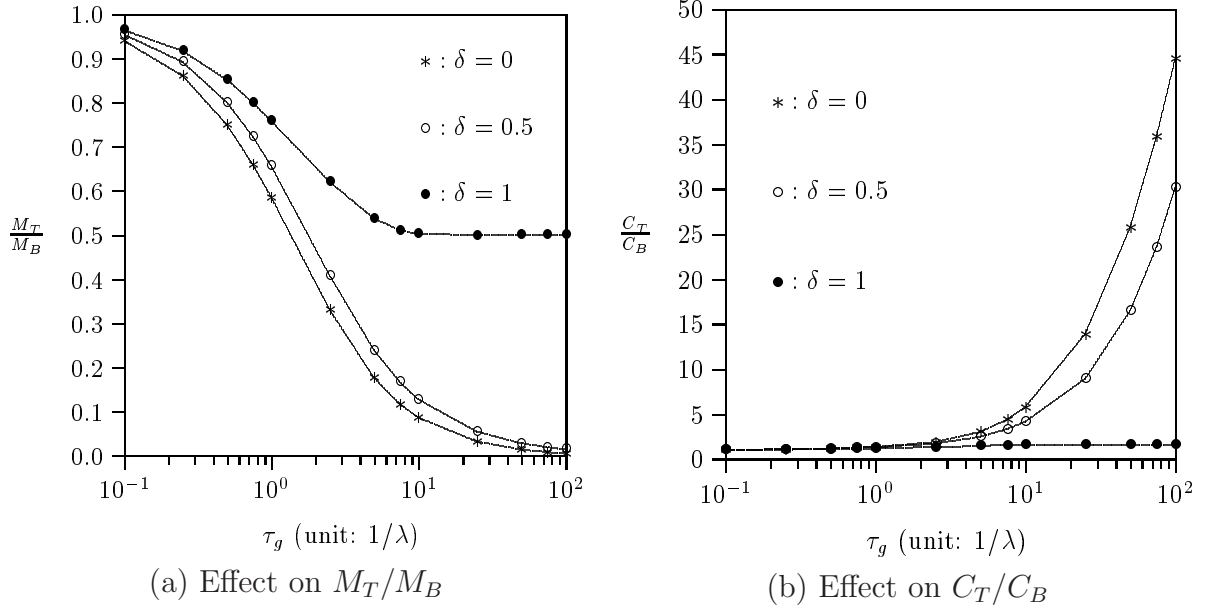


Figure 4.6: Effects of τ_g ($N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$)

4.5 Numerical Examples

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

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

Effects of the granted time units τ_g . Fig. 4.6 plots M_T/M_B and C_T/C_B against the threshold parameter δ and the granted time units τ_g , where $N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$. Fig. 4.6 shows the trivial result that M_T/M_B is a decreasing function of τ_g , and C_T/C_B is an increasing function of τ_g . The non-trivial result is that when $\tau_g \leq 5/\lambda$, the threshold-based scheme significantly reduces the ABMF signaling overhead while the inaccuracy of the credit information insignificantly increases.

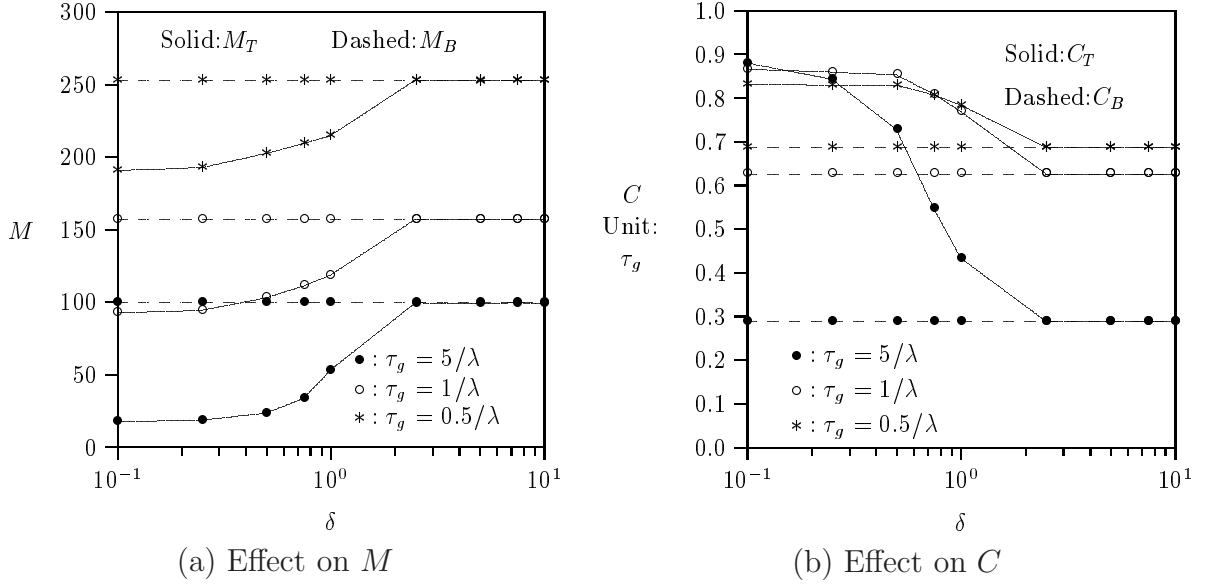


Figure 4.7: Effects of δ ($N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$)

Effects of the threshold parameter δ . Fig. 4.7 plots M and C against the threshold parameter δ , where $N = 2$, $\alpha_2 = 2\alpha_1$ and $p_0 = 0.01$. The output measures for the basic scheme are not affected by δ . For the threshold-based scheme, M_T increases and C_T decreases as δ increases. When δ is large, the basic scheme and the threshold-based scheme have the same performance. In Fig. 4.7, the performance of the threshold-based scheme is similar to that of the basic scheme when $\delta \geq 2.5$.

Fig. 4.7 (b) quantitatively indicates how δ and τ_g affect C . In the OCS, the guideline for selecting the granted time units τ_g can be found in [52]. When a specific τ_g is selected, for example, $\tau_g = 5/\lambda$, we observe that $C_B = 0.29\tau_g$ in the basic scheme. If the mobile operator can tolerate a larger inaccuracy of the credit information, e.g., $C \leq 0.5\tau_g$, then we can choose $\delta = 1$ in the threshold-based scheme. In this case, $C_T = 0.43\tau_g$ and M is decreased by 46.45% as compared with the basic scheme.

Effects of the session termination probability p_0 . Fig. 4.8 plots M_T/M_B and C_T/C_B as functions of p_0 and δ , where $N = 2$, $\alpha_2 = 2\alpha_1$ and $\tau_g = 1/\lambda$. As p_0 increases, both M_T and M_B decrease. However, the proportion M_T/M_B of the ABMF message exchanges saved by the threshold-based scheme is the same as illustrated in Fig. 4.8 (a). Fig. 4.8 (b) shows that C_T/C_B is also insignificantly affected by p_0 .

Effects of the number N of QoS classes. Fig. 4.9 plots M_T/M_B and C_T/C_B as functions of N and δ , where $\alpha_i = i\alpha_1$, $p_0 = 0.01$ and $\tau_g = 1/\lambda$. This figure shows that

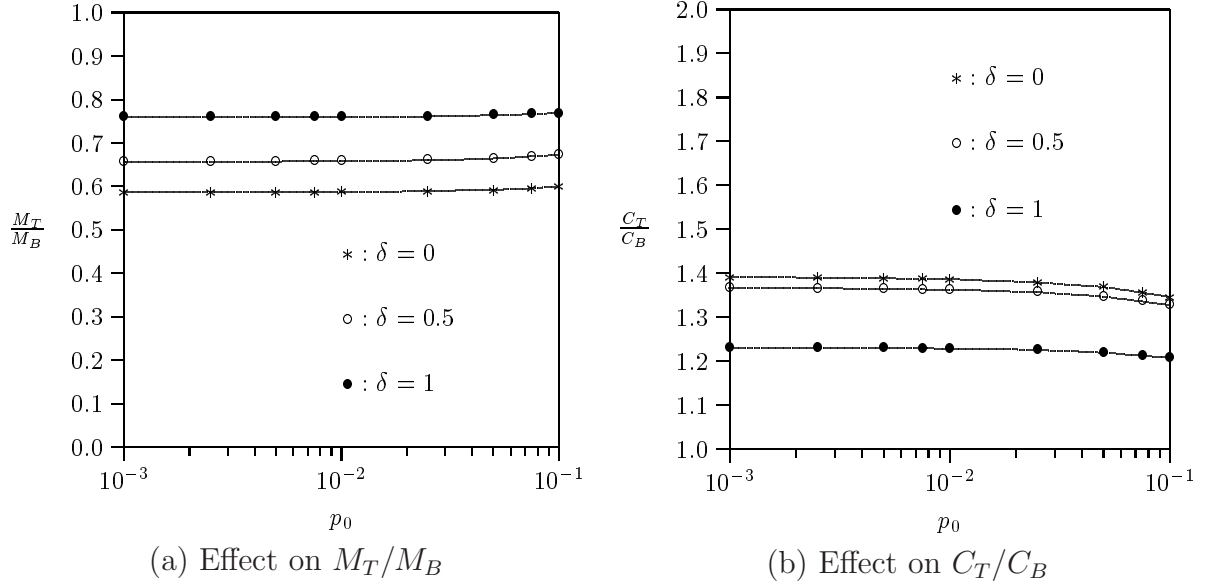


Figure 4.8: Effects of δ and p_0 ($N = 2$, $\alpha_2 = 2\alpha_1$ and $\tau_g = 1/\lambda$)

both M_T/M_B and C_T/C_B are not significantly affected by the number N of QoS classes. For $N = 4$, when δ decreases from 0.5 to 0, M_T/M_B decreases by 9.79% and C_T/C_B only increases by 2.71%. Therefore, it is better to set $\delta = 0$ than $\delta = 0.5$ in this case. Note that the threshold-based scheme with $\delta = 0$ is not equivalent to the basic scheme.



4.6 Summary

This chapter studied *Online Charging System* (OCS) in UMTS. We proposed a threshold-based scheme with parameter δ to reduce the traffic signaling for the OCS credit re-authorization procedure. An analytic model is developed to investigate the performance on the basic scheme [51] and our proposed threshold-based scheme. Basically, the threshold-based scheme reduces the number M of ABMF message exchanges during a session at the cost of increasing the inaccuracy of credit information C when a balance check occurs. These two conflicting output measures are affected by the threshold parameter δ and the granted time units τ_g . We make the following observations, where the subscripts “B” and “T” in the output measures M and C represent the basic scheme and the threshold-based scheme, respectively.

- the ratio M_T/M_B is a decreasing function of τ_g , and the ratio C_T/C_B is an increasing function of τ_g . When τ_g is small, the threshold-based scheme significantly

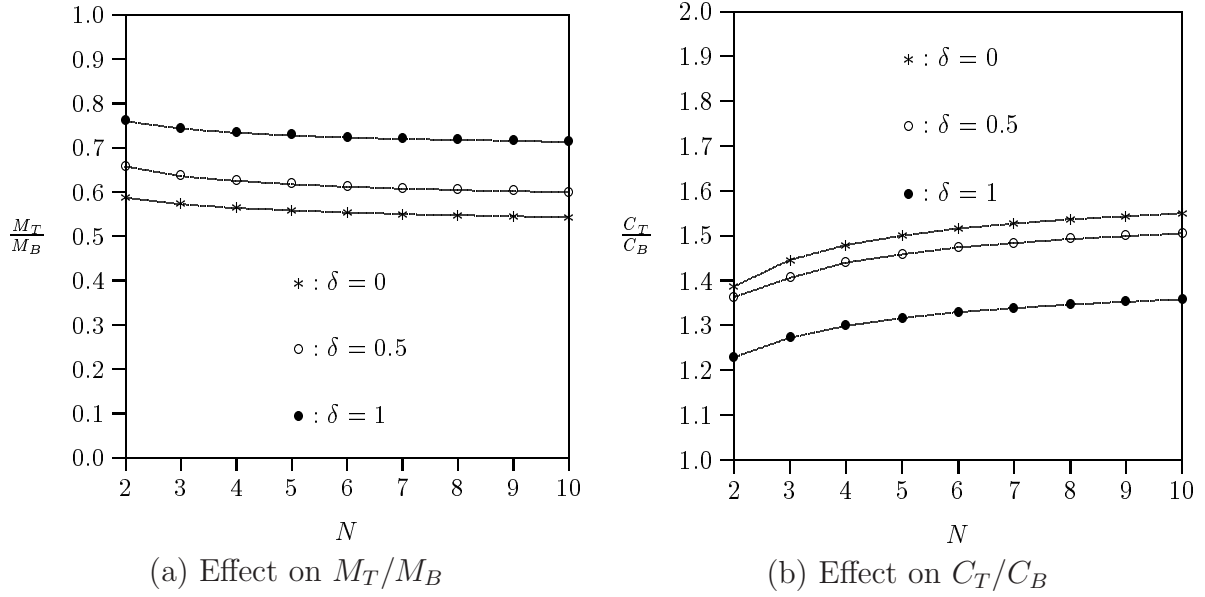


Figure 4.9: Effects of δ and N ($\alpha_i = i\alpha_1$, $p_0 = 0.01$ and $\tau_g = 1/\lambda$)

reduces the ABMF signaling overhead while the inaccuracy of the credit information insignificantly increases.

- As the threshold parameter δ increases, M_T increases and C_T decreases. When δ is large, the basic scheme and the threshold-based scheme have the same performance.
- As the session termination probability p_0 increases, both M_T and M_B decrease. However, the ratio M_T/M_B is not affected. Also, the ratio C_T/C_B is insignificantly affected by p_0 .
- Both M_T/M_B and C_T/C_B are not significantly affected by the number N of QoS classes.

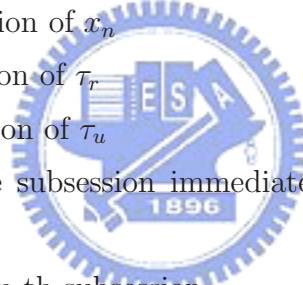
Based on the above discussion, the mobile operator can select the appropriate δ and τ_g values for various traffic conditions based on our model.

4.7 Notation

The notation used in this chapter is listed below.

- α : the expected credit units consumed in one time unit
- α_i : the number of credit units charged for every time units in a class i session
- $\alpha(l) = \alpha_{(l+i \bmod 2)+1}$

- $\hat{\alpha} = \alpha_2/\alpha_1$
- δ : the parameter for the threshold-based scheme
- $\Delta(t)$: the number of QoS changes occur in $[t_1^*, t]$ is $\Delta(t) - 1$
- C_B : the expected inaccuracy of credit information when a balance check occurs during an in-progress session in the basic scheme
- C_T : the expected inaccuracy of credit information when a balance check occurs during an in-progress session in the threshold-based scheme
- $E[m_B]$: the expected number of ABMF message exchanges for a subsession in the basic scheme
- $E[m_T]$: the expected number of ABMF message exchanges for a subsession in the threshold-based scheme
- M_B : the expected number of ABMF message exchanges for a GPRS session in the basic scheme
- M_T : the expected number of ABMF message exchanges for a GPRS session in the threshold-based scheme
- $f_x(x_n)$: the density function of x_n
- $f_r(\tau_r)$: the density function of τ_r
- $f_u(\tau_u)$: the density function of τ_u
- I_n : the QoS class of the subsession immediately before the last ABMF message exchange occurs
- J_n : the QoS class of the n -th subsession
- λ : the rate of x_n
- μ : the rate of τ_g
- k : a balance check occurs at t_B in subsession $n + k$
- K : the QoS class of the session at t_1^*
- $m(i, j)$: the number of ABMF message exchanges for a subsession in state (i, j)
- N : the number of QoS classes
- n_c : the number of QoS changes occur during the streaming session (including the last QoS change that terminates the session)
- n_a : the number of QoS changes occur during $[t_1^*, t_2^*]$
- p_0 : the probability that an in-progress session terminates when a QoS change occurs
- p_a : the transition probability from state $(1, 2)$ to state $(1, 1)$
- p_b : the transition probability from state $(2, 1)$ to state $(2, 2)$



- $P_{i,k}$: the probability that exactly k QoS changes occurs in $[t_1^*, t_B]$ under the condition that $K = i$
- $\pi_{(i,j)}$: the stationary distribution of the Markov chain $\{Z_n, n \geq 1\}$
- $\tau_{a,1}$: the time units that can be supported by the remaining granted credit units at t_1^*
- $\tau_{a,l}$: the time units that can be supported by the remaining granted credit units at t_{n+l-1} ($l \geq 2$)
- $\tau_{d,1} = \min(\tau_r, \tau_{a,1})$
- $\tau_{d,l} = \min(x_{n+l-1}, \tau_{a,l})$ for $l \geq 2$
- τ_d : the time period between when the previous ABMF message exchange occurs and when the balance check occurs
- τ_g : the time units that the SBCF grants to the GGSN through the ABMF message exchange
- $\tau(i, j)$: the unused time units at the beginning of the subsession in state (i, j)
- τ_r : the period between when the ABMF message exchange occurs and when the next subsession starts
- τ_u : the time units left at the end of the previous subsession
- t_1^* : the time that the previous ABMF message exchange before t_B
- t_2^* : the time that the next ABMF message exchange after t_B
- t_B : the time that the balance check occurs
- t_n : the time that the n subsession starts
- x_n : the subsession holding time
- $y = t_B - \max(t_1^*, t_{n+k})$
- $Z_n = (I_n, J_n)$ is a two dimensional random variable for the Markov chain

Chapter 5

Conclusions and Future Work

The *IP Multimedia Core Network Subsystem* (IMS) provides real-time multimedia services for *Universal Mobile Telecommunications System* (UMTS). Traditional voice and basic data services such as *Short Message Service* (SMS) and *Multimedia Messaging Service* (MMS) deliveries are the principal sources of revenue for most mobile operators. The deployment of IMS will bring more competition for more specialized services and content. However, before the benefits of IMS and all-IP can be realized, a fundamental issue must be addressed is charging/billing, which is one of the most important activities in telecommunications. In this dissertation, we first studied how to design an IMS prepaid application server to integrate existing IP-based services. Then we investigated how to improve the performance in UMTS online charging system. This chapter concludes our work presented in this dissertation, and briefly discusses future directions of our work.

5.1 Concluding Remarks

In Chapter 2, we proposed a prepaid application server to handle both the prepaid calls and messaging services in the IMS. When both voice and messaging are simultaneously offered, a potential problem is that the delivery of a message during a call may result in force-termination of that call due to credit depletion. To address this issue, we proposed a strategy to determine if a prepaid message can be sent out during a call session. We presented an analytic model to investigate the performance of this strategy. The analytic model developed in this chapter is validated against the simulation experiments. Our study provides guidelines to select appropriate input parameters for the prepaid application server.

In 3GPP Release 6, the IP-based *Online Charging System* (OCS) is defined to support

online charging for GPRS and IMS services. In Chapter 3, we studied the online credit reservation mechanism for the online charging system. Through the Recharge Threshold-based Credit Reservation (RTCR) mechanism, prepaid IMS services can be supported by the OCS in UMTS. In RTCR, when the remaining amount of prepaid credit is below a threshold, the OCS reminds the user to recharge the prepaid account. It is essential to choose an appropriate recharge threshold to reduce the probability that the in-progress service sessions are forced-terminated. An analytic model is developed to investigate the performance of RTCR for OCS. We also developed simulations experiments to evaluate the performance of the proposed scheme. Based on our study, the mobile operator can select the appropriate parameter values for various traffic conditions.

During an online charging *General Packet Radio Service* (GPRS) session, a number of mid-session events, such as *Quality of Service* (QoS), could dynamically affect the rating of the in-progress service. When such events occur, the GPRS support node needs to re-authorize the granted credit units with the OCS. If the QoS changes frequently, the signaling traffic incurred by the re-authorization procedure increases heavily. In Chapter 4, we proposed a threshold-based scheme that utilizes a threshold parameter δ to reduce the signaling traffic for the OCS credit re-authorization procedure. Simulations and analytic model are developed to evaluate the performance of the proposed scheme. By selecting an appropriate δ value, our study indicates that the signaling overhead in the OCS can be significantly reduced while the inaccuracy of the credit information insignificantly increases. Based on our model, the mobile operator can choose appropriate parameter values in the threshold-based scheme for various traffic conditions.

5.2 Future Work

Based on the research results of this dissertation, we suggest the following topics for future research:

Performance of the Tariff Switch Mechanism. During a service session, the tariff information may be changed when a specified event occurs (i.e., a tariff switch is reached). In such case, all active user sessions should report their session usage by the end of the validity period of the current request and should receive new quota for resource usage for the new tariff period. It is important to set the tariff switch time appropriately such that the OCS can handle all the requests without delaying the service continuity. In order to avoid the tremendous signaling messages caused

by simultaneous quota refresh, the OCS should handle the credit allocation with more flexibility. For example, the granted credit units can be split into resource usage before a tariff switch and resource usage after a tariff switch. Different kinds of credit allocation mechanisms for tariff switch can be further investigated and analysed.

Reducing Signaling Messages for Failure Handling. In the Diameter CCA message, the *Credit-Control-Failure-Handling* AVP determines what to do if the sending of Diameter credit control messages to the OCS has been temporarily prevented. The usage of *Credit-Control-Failure-Handling* AVP gives flexibility to have different failure handling for online credit control session. As defined in RFC 4006 [30], the Tx timer is introduced to limit the waiting time in the *Charging Trigger Function* (CTF) for the CCR response from the OCS. When the Tx timer elapses, the CTF takes an action to the end user according to the value of the *Credit-Control-Failure-Handling* AVP. It is possible that several concurrent online charging sessions are handled in the same CTF. In this case, a mechanism is needed to reduce the concurrent retransmissions of credit control messages exchanged between the CTF and the OCS.

Policy and Charging Control Integration. The QoS control in IMS/GPRS is realized by the *Session-Based Local Policy* (SBLP) functionality defined in 3GPP Release 5 [3]. In SBLP, the *Policy Decision Function* (PDF) is responsible for making policy decisions based on session and media-related information obtained from the CSCF. On the GGSN side, the *Policy Enforcement Function* (PEF) is implemented to be responsible for QoS control of the IP service flows. When the QoS policy control and the flow-based charging functionalities are used as separate mechanisms, it will increase the interworking cost between the network nodes (e.g., GGSN and P-CSCF) and the charging nodes. Through the *Policy and Charging Control* (PCC) defined in 3GPP Release 7 [19], integration of QoS policy and charging rules can be realized in the IMS network or other bearer network (e.g., WLAN, WiMAX). The design of the PCC architecture to support roaming among heterogeneous wireless network is for further study.

Bibliography

- [1] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; Architectural Requirements for Release 1999. Technical Report 3G TS 23.121 Version 3.6.0 (2002-06), 3GPP, 2002.
- [2] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Multimedia Messaging Service (MMS); Stage 1. 3rd Generation Partnership Project . Technical Specification 3G TS 22.140 Version 6.7.0 (2005-03), 2005.
- [3] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; End-to-End QoS Concept and Architecture. Technical Specification 3G TS 23.207 Version 5.10.0 (2005-09), 2005.
- [4] 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. Technical Specification 3G TS 32.215 Version 5.9.0 (2005-06), 2005.
- [5] 3GPP. 3rd Generation Partnership Project; Technical Specification Group; Telecommunication management; Charging management; Charging data description for application services. Technical Specification 3G TS 32.235 Version 5.5.0 (2005-09), 2005.
- [6] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; IP Multimedia Call Control Protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3. Technical Specification 3G TS 24.229 Version 5.18.0 (2006-09), 2006.

- [7] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Policy control over Gq interface. Technical Specification 3G TS 29.209 Version 6.6.0 (2006-09), 2006.
- [8] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Core Network; IP Multimedia (IM) session handling; IM call model; Stage 2. Technical Specification 3G TS 23.218 Version 5.9.0 (2006-06), 2006.
- [9] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Service and System Aspects; Telecommunication management; Charging management; Charging data description for the IP Multimedia Subsystem (IMS). Technical Specification 3G TS 32.225 Version 5.11.0 (2006-03), 2006.
- [10] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Service and System Aspects; Telecommunication management; Charging management; On-line Charging System (OCS): Applications and interfaces. 3rd Generation Partnership Project . Technical Specification 3G TS 32.296 Version 6.3.0 (2006-09), 2006.
- [11] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Telecommunication management; Charging management; Charging architecture and principles. Technical Specification 3G TS 32.240 Version 6.4.0 (2006-09), 2006.
- [12] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; General Packet Radio Service (GPRS); Service Description; Stage 2. Technical Specification 3G TS 23.060 Version 5.13.0 (2006-12), 2006.
- [13] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Terminals; Multimedia Messaging Service (MMS); Functional description; Stage 2. Technical Report 3G TS 23.140 Version 6.14.0 (2006-09), 3GPP, 2006.
- [14] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Terminals; Technical realization of the Short Message Service (SMS). Technical Specification 3G TS 23.040 Version 6.8.1 (2006-10), 2006.
- [15] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Service and System Aspects; Telecommunication management; Charging management; Di-

- ameter charging applications. Technical Specification 3G TS 32.299 Version 6.11.0 (2007-06), 2007.
- [16] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Service and System Aspects; Telecommunication management; Charging Management; IP Multimedia Subsystem (IMS) charging. Technical Specification 3G TS 32.260 Version 6.8.0 (2007-03), 2007.
- [17] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Service and System Aspects; Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and codecs (Release 6) . Technical Specification 3G TS 26.234 Version 6.12.0 (2007-09), 2007.
- [18] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; IP Multimedia Subsystem Stage 2. Technical Specification 3G TS 23.228 Version 6.16.0 (2007-03), 2007.
- [19] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; Policy and charging control architecture. Technical Specification 3G TS 23.203 Version 7.4.0 (2007-09), 2007.
- [20] Bhushan, B., Hall, J., Kurtansky, P., and Stiller, B. OSS Functions for Flexible Charging and Billing of Mobile Services in a Federated Environment. *9th IFIP/IEEE International Symposium on Integrated Network Management*, May 2005.
- [21] Calhoun, P., Loughney, J., Guttman, E., Zorn, G. and Arkko, J. Diameter Base Protocol. IETF RFC 3588, September 2003.
- [22] Campbell, B., Rosenberg, J., Schulzrinne, H., Huitema, C. and Gurle, D. Session Initiation Protocol (SIP) Extension for Instant Messaging. IETF RFC 3428, December 2002.
- [23] Chang, M.-F. and Yang, W.-Z. Performance of Mobile Prepaid and Priority Call Services. *IEEE Communications Letters*, 6(2):61–63, 2002.
- [24] Chang, M.-F., Lin, Y.-B., and Yang, W.-Z. Performance of Hot Billing Mobile Prepaid Service. *Computer Networks Journal*, 36:269–290, 2001.

- [25] Chang, M.-F., Yang, W.-Z., and Lin, Y.-B. Performance of Hot Billing Mobile Prepaid Service. *IEEE Transactions on Vehicular Technology*, 51(3):597–612, 2002.
- [26] Clemente M.A. and Velez F.J. Parameters for Tele-traffic Characterisation in Enhanced UMTS. <http://seacorn.ptinovacao.pt/>.
- [27] Collins, D. *Carrier Grade Voice over IP*. McGraw-Hill, 2003.
- [28] Convergys. On-line Charging: Delivering Pre- and Post-Paid Convergence. 2005.
- [29] Ghys, F. and Vaaraniemi, A. Component-based Charging in a Next-Generation Multimedia Network. *IEEE Communications Magazine*, 41(1):99–102, 2003.
- [30] Hakala, H., Mattila, L., Koskinen, J.-P., Stura, M. and Loughney, J. Diameter Credit-Control Application. IETF RFC 4006, August 2005.
- [31] Handley, M., Jacobson, V., Perkins, C. and Johnston, A. SDP: Session Description Protocol. IETF RFC 4566, July 2006.
- [32] Hung, H.-N., Lin, Y.-B., Peng, N.-F. and Sou, S.-I. Connection Failure Detection Mechanism of UMTS Charging Protocol. *IEEE Transactions on Wireless Communications*, 5(5):1180–1186, 2006.
- [33] Kelly, F.P. *Reversibility And Stochastic Networks*. John Wiley & Sons, 1979.
- [34] Kleinrock, L. *Queueing Systems: Volume I – Theory*. New York: Wiley, 1976.
- [35] Lin, P., Lin, Y.-B., Yen, C.S., and Jeng, J.-Y. Credit Allocation for UMTS Prepaid Service. *IEEE Trans. on Vehicular Technology*, 55(1):306–316, 2006.
- [36] Lin, Y.-B., and Chlamtac, I. *Wireless and Mobile Network Architectures*. John Wiley & Sons, 2001.
- [37] Lin, Y.-B., and Pang, A.-C. *Wireless and Mobile All-IP Networks*. Wiley, 2005.
- [38] Little J. D. C. A Proof for the Queuing Formula: $L = \lambda W$. *Operations Research*, 9(3):383–387, 1961.
- [39] MARBEAN. MARBEN Diameter Product. 2007.
- [40] Mitrani, I. *Modeling of Computer And Communication Systems*. Cambridge University Press, 1987.

- [41] Novak, L. and Svensson, M. MMS-Building on the success of SMS. *Ericsson Review*, 3:102–109, 2001.
- [42] RADVISION. IMS DIAMETER Toolkit. 2007.
- [43] Rao, C.H., Chang, D.-F., and Lin, Y.-B. iSMS: An Integration Platform for Short Message Service and IP Networks. *IEEE Network*, 15(2):48–55, 2001.
- [44] Rao, C.H., Cheng, Y.-H., Chang, K.-H., and Lin, Y.-B. iMail: A WAP Mail Retrieving System. *Information Sciences*, 151:71–91, 2003.
- [45] Rigney, C. and Livingston. RADIUS Accounting. IETF RFC 2866, June 2000.
- [46] Rigney, C., Willens, S., Rubens, A. and Simpson W. Remote Authentication Dial In User Service (RADIUS). IETF RFC 2865, June 2000.
- [47] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M. and Schooler, E. SIP: Session Initiation Protocol. IETF RFC 3261, June 2002.
- [48] Ross, S.M. *Stochastic Processes*. John Wiley & Sons, 1996.
- [49] Schulzrinne, H., and Rosenberg, J. The IETF Internet Telephony Architecture. *IEEE Communications Magazine*, pages 18–23, May 1999.
- [50] Schulzrinne, H., Casner, S., Frederick, R. and Jacobson, V. RTP: A Transport Protocol for Real-Time Applications. IETF RFC 1889, January 1996.
- [51] Sou, S.-I. Performance Analysis of Credit Re-authorization Schemes in UMTS Online Charging System. *International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2007.
- [52] Sou, S.-I., Hung, H.-N., Lin, Y.-B., Peng, N.-F. and Jeng, J.-Y. Modeling Credit Reservation Procedure for UMTS Online Charging System. Accepted for publication in *IEEE Transactions on Wireless Communications*.
- [53] Sou, S.-I., Lin, Y.-B., Wu, Q., and Jeng, J.-Y. Modeling Prepaid Application Server of VoIP and Messaging Services for UMTS. *IEEE Transactions on Vehicular Technology*, 56(3):1434–1441, 2007.

- [54] Sou, S.-I., Wu, Q., Lin, Y.-B., and Yeh, C.-H. Prepaid Mechanism of VoIP and Messaging Services. *IEEE International Conference on Information Technology Research and Education (ITRE)*, 2005.
- [55] Vovida. <http://www.vovida.org>.
- [56] Yang, W.-Z., Lu, F.-S., and Chang, M.-F. Performance Modeling of Integrated Mobile Prepaid Services. *IEEE Trans. on Vehicular Technology*, 56(2):899–906, 2007.



Publication List

- **Book**

1. Lin Y.-B. and Sou S.-I., *Charging for Mobile All-IP Telecommunications*, to be published by Wiley Publishing Co. in October, 2008.

- **Journal Publications**

1. Hung, H.-N., Lin, Y.-B., Peng, N.-F. and Sou, S.-I. Connection Failure Detection Mechanism of UMTS Charging Protocol. *IEEE Transactions on Wireless Communications*, 5(5): 1180-1186, 2006.
2. Sou, S.-I., and Lin, Y.-B. Modeling of Mobility Database Failure Restoration using Checkpointing Schemes. *IEEE Transactions on Wireless Communications*, 6(1): 313-319, 2007.
3. Sou, S.-I., Lin, Y.-B., Wu, Q., and Jeng, J.-Y. Modeling Prepaid Application Server of VoIP and Messaging Services for UMTS *IEEE Transactions on Vehicular Technology*, 56(3): 1434-1441, 2007.
4. Sou, S.-I., Hung, H.-N., Lin, Y.-B., Peng, N.-F., and Jeng, J.-Y. Modeling Credit Reservation Procedure for UMTS Online Charging System. *IEEE Transactions on Wireless Communications*, 6(11): 4129-4135, 2007.
5. Sou, S.-I., and Lin, Y.-B. Broadcast Approach for UMTS Mobility Database Recovery, *IEEE Transactions on Mobile Computing*, 6(8): 865-871, 2007.
6. Sou, S.-I., Lin, Y.-B. and Jeng, J.-Y. Reducing Credit Re-authorization Cost in UMTS Online Charging System. Accepted and to appear in *IEEE Transactions on Wireless Communications*.

- **Conference Papers**

1. Sou, S.-I. Performance Analysis of Credit Re-authorization Schemes in UMTS Online Charging System, *ACM International Wireless Communications and Mobile Computing Conference (IWCMC)*, Hawaii, USA, 2007.
2. Sou, S.-I., Wu Q., Lin Y.-B. and Yeh C.-H. Prepaid Mechanism of VoIP and Messaging Services, *IEEE International Conference on Information Technology Research and Education (ITRE)*, Hsinchu, Taiwan, 2005.

3. Sou, S.-I., Wu Q., Lin Y.-B. and Chen W.-E. SIP-based VoIP Prepaid System on NTP VoIP Platform, *Proceedings of the Taiwan Academic Network Conference (TANET)*, Taitung, Taiwan, 2004.

- **Book Chapter**

1. Sou S.-I. “IMS Charging Management in Mobile Telecommunication Networks,” to appear in *VoIP Handbook: Applications, Technologies, Reliability, and Security*, by Syed A. Ahson and Mohammad Ilyas (editor), CRC Press, 2008.

