國 立 交 通 大 學

資訊科學與工程研究所

博 士 論 文

IMS 之通話控制研究

A Study on IMS Call Control

研 究 生：蔡孟勳

指導教授：林一平　教授

中 華 民 國 九 十 八 年 九 月

IMS 之通話控制研究
A Study on IMS Call Control

研 究 生：蔡孟勳　　Student：Meng-Hsun Tsai

指導教授：林一平 博士　　Advisor：Dr. Yi-Bing Lin

國 立 交 通 大 學
資 訊 科 學 與 工 程 研 究 所

博 士 論 文

A Dissertation

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

September 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年九月

本 論 文 係 接 受
聯 發 科 技 獎 學 金 贊 助

This work was supported by the MediaTek Fellowship.

IMS 之通話控制研究

學生: 蔡孟勳　　　　　　　　　　　　　　　指導教授: 林一平 博士

國立交通大學資訊工程與科學研究所博士班

摘　　要

全球行動通訊系統 (UMTS) 為第三代行動通訊的主流規格之一。UMTS 的核心網路包含線路交換領域 (Circuit-Switched Domain) 與封包交換領域 (Packet-Switched Domain)。為了在行動通訊系統上整合網際網路技術，3GPP 在封包交換領域上提出了 IP 多媒體子系統 (IMS)，該子系統利用 VoIP 技術提供網際網路多媒體服務。

在 UMTS 與網際網路的整合過程中，通話控制成為最關鍵的研究議題之一。隨著 IMS 的演化，許多新穎的通話控制功能相繼被提出。本論文探討 IMS 通話控制對三個主要應用服務的影響：緊急電話、隨按即說 (PoC) 以及語音通話連續性 (VCC)。

IMS 緊急電話服務允許緊急應變中心透過位置輪詢機制來持續追蹤發話者的位置，此機制具有追蹤誤差與重複詢問的問題。我們針對 IMS 的位置追蹤進行效能分析，並提出可即時追蹤與不會重複詢問的主動位置回報機制。

隨按即說服務提供類似 “walkie-talkie” 無線對講機的群組通訊模式，並利用發言權控制機制來仲裁發言權。我們在 IMS 平台上實作了隨按即說服務，並分析發言權控制機制的效能。研究結果顯示，藉由緩衝機制 (讓未獲發言權的請求稍作等待)，隨按即說服務可以在維持相同發言權獲得率的情況下，支援到無緩衝機制時兩倍數量以上的群組成員。

語音通話連續性服務允許使用者在通話過程中切換領域。然而，領域轉換可能會產生大量的信令交換與資源保留，造成冗長的轉換延遲時間。針對此議題，我們提出

在 IMS 平台上的通道保留機制，以達到快速的領域轉換。研究結果顯示，通道保留機制的優異效能在使用者行爲愈不規則時成效愈顯著。

在前述 IMS 平台上衍生出來的三個研究子題中，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。本論文的研究成果可提供行動電信業者各項通話控制與應用服務之參數建議。實驗結果證實了我們提出的方法能有效提升通話控制的效能。

關鍵字: 通話控制、 緊急電話、 IP 多媒體子系統、 隨按即說、 全球行動通訊系統、 語音通話連續性

A Study on IMS Call Control

Student: Meng-Hsun Tsai   Advisor: Dr. Yi-Bing Lin

Institute of Computer Science and Engineering

National Chiao Tung University

Abstract

Universal Mobile Telecommunications System (UMTS) is one of the major standards for the third generation (3G) mobile telecommunications. In UMTS, the core network consists of two service domains: the *circuit-switched* (CS) and the *packet-switched* (PS) domains. To integrate with the Internet, the *Third Generation Partnership Project* (3GPP) proposed the *IP Multimedia Core Network Subsystem* (IMS) in the PS domain to provide Internet services through the *Voice over Internet Protocol* (VoIP) technology.

In integration of UMTS and Internet, call control has become one of the most fundamental issues in telecommunications. As IMS evolves, more new call control features are supported. This dissertation investigates how the IMS call control affects three major applications: emergency call, *Push-to-talk over Cellular* (PoC), and *Voice Call Continuity* (VCC).

IMS emergency call provides the capability for the emergency center to track the caller's location through a polling procedure. The problems of existing mechanism include mis-tracking and redundant query. We study the performance of location tracking in the IMS, and then propose the Active Location Reporting (ALR) scheme which reports the caller's location in real time without redundant query.

PoC provides a "walkie-talkie"-like group communication method. The speak permission is arbitrated through a talk burst control mechanism. We implement a prototype in the IMS and study the performance of the talk burst control mechanism. Our study indicates that by including the buffer mechanism for PoC (where an ungranted request can be buffered in the queue, and may be granted later), more than twice clients can be supported while maintaining the same granting probability of a request.

VCC allows a user to switch from one domain to another during a call. Domain transfer may produce a large number of message exchanges and resource reservation that result in long switching latency. To resolve this issue, we propose the Bearer Reservation with Preemption (BRP) scheme in the IMS to support fast domain transfer. Our study indicates that when the user behavior (either in terms of call holding time or movement pattern) is more irregular, the advantage of the BRP scheme becomes more significant.

We also develop analytic models and simulation experiments to investigate the effects of input parameters. Our study provides suggestions to set up the parameters for mobile operations. Numerical experiments indicate that the proposed schemes significantly improve the performance of IMS call control.

Keywords: Call Control, Emergency Call, IP Multimedia Core Network Subsystem (IMS), Push-to-talk over Cellular (PoC), Universal Mobile Telecommunications System (UMTS), Voice Call Continuity (VCC)

Acknowledgements

I would like to express my deep and sincere gratitude to my advisor, Prof. Yi-Bing Lin for his continuous support, encouragement, and guidance throughout my graduate study. His extensive knowledge and creative thinking have been an invaluable help for me. Without his supervision and perspicaious advice, I can not complete this dissertation.

Great appreciation to MediaTek Fellowship. Without MediaTek's support, this dissertation would not have been possible.

Also, I would like to deliver a special gratitude to Prof. Ming-Feng Chang and Prof. Yu-Chee Tseng for their stimulating suggestions and guidance. Special thanks to my committee members, Prof. Ming-Feng Chang, Prof. Hsi-Lu Chao, Prof. Jean-Lien C. Wu, Prof. Han-Chieh Chao, Prof. Chu-Sing Yang and Dr. Herman Chung-Hwa Rao for their valuable comments and suggestions.

Furthermore, I am grateful to the colleagues in Laboratory 117 for their friendship and helpful discussions. Thanks also to Svilen Ivanov, Georg Lukas and André Herms, my best friends in Germany, for their friendship and help during my internship in Otto-von-Guericke-Universität Magdeburg, Germany. Also, I would like to thank Rebecca Chen, Li-Fen Li, Webbor Lee and Matt Lee for their friendship and help during my internship in IBM. Special thanks to Hsiao-Han Wang, Hui-Wen Dai, Wei-Sheng Cheng, Hsiao-Yun Tseng and Zheng-Han Wu. Their positive attitude encourages me a lot.

Finally, I am grateful to my dear parents, my sister and my girl friend for their unfailing love and firmly support in these years.

Contents

List of Tables

List of Figures

Notation

The notation used in this dissertation is listed below.

- α: the probability of mis-tracking for an SA crossing (Chapter 2)
- β: the probability that redundant queries exist between two SA crossings (Chapter 2)
- C: the capacity (number of channels) of the MSC (Chapter 4)
- $1/\delta$: the expected domain residence time (Chapter 4)
- $1/\eta$: the expected value of the sojourn time t_s (Chapter 4)
- $f_m(t_m)$: the density function of the SA residence time t_m (Chapter 2)
- $f_m^*(s)$: the Laplace transform of $f_m(t_m)$ (Chapter 2)
- γ: the revoking rate (Chapter 3)
- K: the number of requesting and permitted PoC clients (Chapter 3)
- K_0: the number of calls in the MSC seen by a low-priority call arrival (Chapter 4)
- K_i: the number of high-priority calls in the MSC when the i-th high-priority call arrives (Chapter 4)
- K_Q: the number of requests in the queue (Chapter 3)
- K_S: the number of permitted PoC client (Chapter 3)

- λ: job arrival rates which represent
  - the location query rate (Chapter 2)
  - the talk burst request rate (Chapter 3)
  - the VCC CS call arrival rate (Chapter 4)
- λ^*: the expected talk burst request rate (Chapter 3)
- M: the number of PoC clients in a PoC group (Chapter 3)
- $1/\mu$: service times which represent
  - the expected SA residence time (Chapter 2)
  - the expected speak time (Chapter 3)
  - the expected call holding time (Chapter 4)
- N: the number of queries between two SA crossings (Chapter 2)
- N_R: the expected number of queries between two SA crossings under the condition that $N > 1$ (therefore, $N_R - 1$ is the expected number of redundant queries) (Chapter 2)
- $1/\omega$: the expected patient time (Chapter 3)
- P_D: the probability that the request of a PoC client is not granted (Chapter 3)
- p_f: the probability that there is no available resource when the UE switches back to the CS domain (Chapter 4)
- $p_{(m,n)}$: the one-step transition probability from state $K_i = m$ to state $K_{i+1} = n$ (Chapter 4)
- $p_{(m,n)}^{(l)}$: the probability that the stochastic process moves from state m to state n with exact l steps (Chapter 4)
- p_n: the probability that the reserved CS bearer is not preempted (Chapter 4)
- p_r: the probability that the reserved CS bearer has been preempted, but the resource becomes available when the UE switches back to the CS domain (Chapter 4)

- P_R: the probability that a PoC client obtains the permission, but is revoked before it finishes the talk (Chapter 3)
- $r_q(\tau_q)$: the density function of the invalid period τ_q (Chapter 2)
- τ_a: the inter-request time (Chapter 3)
- τ_p: the patient time (Chapter 3)
- τ_q: the invalid period (Chapter 2)
- τ_s: the speak time (Chapter 3)
- t_c: the call holding time (Chapter 4)
- t_d: the domain residence time (Chapter 4)
- t_h: the inter-arrival time between the i-th and the $(i+1)$-th high-priority call arrivals (Chapter 4)
- t_q: the inter-query interval (Chapter 2)
- θ_{P2C}: the percentage of re-connection overhead saved by our approach for PS-to-CS domain transfer as compared with the 3GPP approach (Chapter 4)
- θ_{C2P}: the percentage of re-connection overhead saved by our approach for CS-to-PS domain transfer as compared with the 3GPP approach (Chapter 4)
- t_m: the SA residence time (Chapter 2)
- T_i: the expected invalid period (Chapter 2)
- T_R: the revoking timer (Chapter 3)
- t_s: the sojourn time; i.e., the period that a high-priority (low-priority) call utilizes (reserves) a channel at the MSC before it is switched to another domain or is completed (Chapter 4)
- V_c: the variance of the call holding times (Chapter 4)
- V_d: the variance of the domain residence times (Chapter 4)

- V_p: the variance of the patient times (Chapter 3)
- V_m: the variance of the SA residence times (Chapter 2)
- V_i: the variance of the invalid periods (Chapter 2)
- V_s: the variance of the speak times (Chapter 3)
- W: the expected waiting time of a request (Chapter 3)

Chapter 1

Introduction

Universal Mobile Telecommunications System (UMTS) is one of the major standards for the third generation (3G) mobile telecommunications. In UMTS, the core network consists of two service domains: the *circuit-switched* (CS) and the *packet-switched* (PS) domains [1]. The CS domain evolved from *Global System for Mobile Communications* (GSM), and the PS domain evolved from *General Packet Radio Service* (GPRS). Compared to GSM and GPRS, UMTS provides higher data rate, larger system capacity, and customized services with quality of services. Currently, the UMTS service is commercially available in 89 countries supported by more than 207 mobile operators.

With the explosive growth of the Internet, supporting Internet telephony through the *Voice over Internet Protocol* (VoIP) technology has become a promising trend in mobile operations. The *Third Generation Partnership Project* (3GPP) proposed the *IP Multimedia Core Network Subsystem* (IMS) in the PS domain to provide Internet services by utilizing the *Session Initiation Protocol* (SIP) [2, 3].

In integration of UMTS and Internet, call control has become one of the most fundamental issues in telecommunications. As IMS evolves, more new call control features are supported. This dissertation investigates how the IMS call control affects three major applications: emergency call, *Push-to-talk over Cellular* (PoC), and *Voice Call Continuity* (VCC). In the emergency call and the VCC applications, we modify the standard IMS call control to improve their performance. In the PoC application, we implement a prototype as well as proposing analytic model and simulation experiments to investigate its performance.

IMS emergency call provides the capability for the emergency center to track the user's location after emergency call setup through a polling procedure, where the emergency center periodically queries the user's location. In this dissertation, we propose analytic and simulation models to study the performance of location tracking, and then propose the Active Location Reporting (ALR) scheme to improve the performance of location tracking.

PoC provides a "walkie-talkie"-like group communication method in the IMS. In this service, several predefined PoC group members participate in one PoC session. The PoC communication is half-duplex, where only one PoC member speaks at a time, and the others listen. The speak permission is arbitrated through a talk burst control mechanism; the design of the talk burst control mechanism significantly affects the performance of PoC. In this dissertation, we propose an analytic model to study the performance of the talk burst control mechanism, and provide suggestions to set up the parameters for IMS PoC.

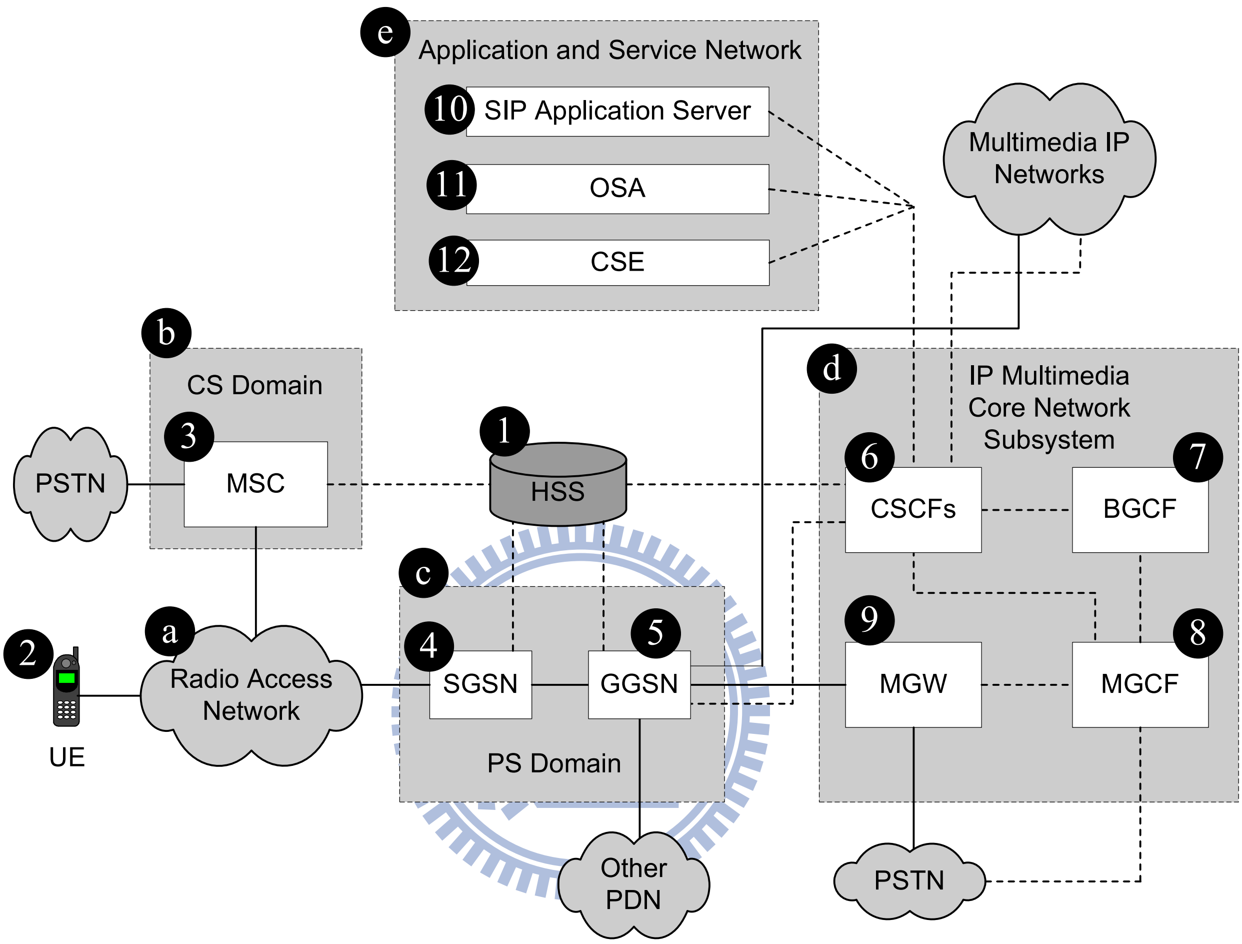
VCC allows a user to switch from one domain to another during a call. A UMTS handset can initiate or receive a call in either the CS or the PS domain. By anchoring the call in the IMS, the user may switch from one domain to another during the call. This process is called *domain transfer*. Domain transfer may produce a large number of message exchanges and resource reservation that result in long switching latency. In this dissertation, we propose the Bearer Reservation with Preemption scheme to support fast domain transfer.

In the following sections, we first present the UMTS network architecture, and then elaborate on the IMS call control issues.

1.1 The UMTS Network

Figure 1.1 illustrates the UMTS network architecture [4]. This architecture consists of a radio access network (RAN; Figure 1.1 (a)), the CS domain (Figure 1.1 (b)), the PS domain, i.e., the GPRS network (Figure 1.1 (c)), the IMS network (Figure 1.1 (d)), and the application and service network (Figure 1.1 (e)). In this architecture, *Home Subscriber Server* (HSS; Figure 1.1 (1)) is the master database containing all user-related subscription information, which supports mobility management of the users. A mobile user utilizes a *User Equipment* (UE; Figure 1.1 (2)) to access CS and PS services. In the CS domain, the *Mobile Switching Center* (MSC; Figure 1.1 (3)) is responsible for call control, including the processing of user data and control signals.

In the PS domain, the *Serving GPRS Support Node* (SGSN; Figure 1.1 (4)) and the *Gateway GPRS Support Node* (GGSN; Figure 1.1 (5)) provide mobility management and session management to the UEs. The GPRS network connects to the external *Packet Data Network* (PDN) or the IMS network through the GGSNs. In the IMS, the transport of user data is separated from that for control signals. The *Call Session Control Functions* (CSCFs; Figure 1.1 (6)) are SIP servers, which are responsible for call control. The *Breakout Gateway Control Function* (BGCF; Figure 1.1 (7)) is responsible for selecting



PSTN: Public Switched Telephone Network
CS: Circuit-Switched
MSC: Mobile Switching Center
UE: User Equipment
SIP: Session Initiation Protocol
OSA: Open Service Access
CSE: CAMEL Service Environment
HSS: Home Subscriber Server

SGSN: Serving GPRS Support Node
GGSN: Gateway GPRS Support Node
PS: Packet-Switched
PDN: Packet Data Network
CSCF: Call Session Control Function
BGCF: Breakout Gateway Control Function
MGCF: Media Gateway Control Function
MGW: Media Gateway

Figure 1.1: The UMTS Network Architecture (dashed lines: signaling; solid lines: signaling/data)

an appropriate *Public Switched Telephone Network* (PSTN) breakout point based on the received SIP request from the CSCFs. The *Media Gateway Control Function* (MGCF; Figure 1.1 (8)) is the same as the Media Gateway Controller in a VoIP network, which controls the connection for media channels in a *Media Gateway* (MGW; Figure 1.1 (9)). The GGSN and MGW together are the network border toward the PSTN and legacy mobile networks.

The application and service network supports flexible services through a service platform. 3GPP defines three possible ways to provide flexible and global services:

SIP Application Server (Figure 1.1 (10)) is used by mobile operators to provide new multimedia SIP applications. The SIP applications are either developed by the mobile operators or purchased from trusted third parties.

Open Service Access (OSA) (Figure 1.1 (11)) is used to give third parties controlled access to the operator's network [5]. The third parties can run applications in their own application servers without concerning the underlying network environment.

Customized Application Mobile Enhanced Logic (CAMEL) Service Environment (CSE) (Figure 1.1 (12)) is used by the mobile operators to provide popular CAMEL services (e.g., prepaid service) to the IMS users [6]. Note that similar services for the CS domain have already been provided via the CAMEL platform.

1.2 IMS Emergency Call

Emergency call is one of the essential telephony services for telecommunications operations. To support IMS emergency call, three network nodes are deployed: the *Gateway*

Mobile Location Center (GMLC), the *Public Safety Answering Point* (PSAP) and the *Emergency-CSCF* (E-CSCF). The GMLC supports *Location Service* (LCS). A PSAP dispatches emergency calls according to the types of emergent events (e.g., fire). When the UE originates an emergency call, the call is established by an E-CSCF, which dispatches the call to the nearest PSAP according to the location of the UE.

The LCS utilizes one or more positioning methods [7, 8] between the RAN and the UE to determine the location of a UE. Four positioning methods are specified in UMTS: *Cell-ID-based*, *Observed Time Difference of Arrival* (OTDOA), *Uplink Time Difference of Arrival* (U-TDOA) and *Assisted Global Positioning System* (A-GPS). The Cell-ID-based method determines the UE's position based on the coverage of Service Areas (SAs). An SA includes one or more cells (base stations). The OTDOA and the U-TDOA methods utilize trilateration to determine the UE's position based on the time differences among the downlink and the uplink signal arrivals, respectively. The A-GPS method speeds up GPS positioning by downloading GPS information through the RAN. Without loss of generality, we consider the Cell-ID-based method in this dissertation.

After emergency call setup, the PSAP may need to monitor the UE's location in real time. The UE's location is tracked by the PSAP through Location Polling, where the PSAP periodically queries the UE's location. In the Location Polling scheme, if the UE does not change its location between two queries, the second query is wasted. On the other hand, if the UE has moved to several new locations between two location queries, then the PSAP may lose track of the UE. To resolve this issue, we propose the Active Location Reporting scheme that reports the UE's location upon change of its SA.

1.3 Push-to-talk over Cellular

PoC is a "walkie-talkie"-like service designed for mobile networks [9]. In this service, several predefined PoC group members participate in one PoC session. The PoC communication is half-duplex; only one PoC member speaks at a time, and the others listen. When a PoC member attempts to speak, he/she presses the push-to-talk button of his/her mobile terminal to ask for the permission; the speak permission is arbitrated through a talk burst control mechanism defined in the Open Mobile Alliance (OMA) PoC specifications [10].

The *PoC server* is responsible for handling PoC session management (create or delete a PoC session) and talk burst control. A PoC session can be initiated by any group member. When the PoC server receives a PoC session invitation, it obtains the group information and its group member list, and then forwards the PoC session invitation to each PoC group member. After the PoC session is established, each PoC group member builds a Real-time Transport Protocol (RTP) [11] session with the PoC server. If a PoC group member obtains the permission, his/her voice is sent to the PoC server through the RTP session. The PoC server then forwards the voice packets to all group members.

In the OMA specification, the talk burst control mechanism is implemented by finite state machines in both the server and the client sides (see Chapter 3 for the detailed descriptions). The talk burst control messages between the PoC group members and the PoC server are carried by RTP Control Protocol (RTCP) packets [11]. To obtain the speak permission, a PoC group member sends a talk burst request message to the PoC server for arbitration. The PoC server may grant the permission or deny the request depending on the current talk burst status (indicated by the server-side finite state machine). A queueing option is provided in the talk burst control mechanism. If this option is selected, then the ungranted requests are buffered in the queue at the PoC server instead of being denied. In Chapter 3, we propose analytic and simulation models to study the performance of the talk burst control mechanism with queueing and without queueing, and provide suggestions to set up the parameters for IMS PoC.

1.4 Voice Call Continuity

VCC allows a user to switch from one domain to another during a call [12]. In existing commercial operation, the CS domain provides full service coverage with limited bandwidth. On the other hand, the PS domain typically provides zonal coverage with larger bandwidth and cheaper services. Therefore, when the PS connection is available, the UE will switch to the PS domain. When the PS connection is not available, the UE switches to the CS domain. A UMTS handset can attach to the CS and the PS domains individually or simultaneously, and initiate or receive a call in either domain. The user may switch from one domain to another during the call. The technique to transfer a voice call between the CS and the PS domains is called VCC.

The VCC call path is partitioned into two segments: the UE-MGW segment and the MGW-Callee segment. When the UE makes a call to a call party in a different network, the call is first routed to the MGW. Then the MGW routes the call to the callee through the destination network. When the UE moves from one domain to another during a call, the UE-MGW segment is switched, and the MGW-Callee segment remains unchanged.

When the UE switches from one domain to another during a voice call, the call is domain-transferred by the *VCC Application Server* (VCC AS). In order to maintain call continuation, the connection in the old domain is released, and a connection is established in the new domain. For example, when a UE decides to transfer its VCC call from the PS domain to the CS domain, a new call is initiated in the CS domain with a specific called number. This number is used to route the call from the CS domain to the VCC AS in the PS domain. When the call setup signal arrives at the VCC AS, the VCC AS updates the UE-MGW segment.

After the call has been successfully switched to the CS domain, the UE may decide to switch the call back to the PS domain again. In this case, the released PS bearer must be re-established. Such bearer re-establishment contributes extra overload to the domain transfer. To speed up the subsequent switchings, we may not release the PS bearer at the PS-to-CS domain transfer, and postpone the bearer release until the VCC call is complete. If the user moves back from the CS domain to the PS domain, the bearer re-establishment is eliminated. Based on the above intuition, we propose the Bearer Reservation with Preemption scheme to support fast domain transfer.

1.5 Organization of the Dissertation

Based on the above discussion, we investigate how the IMS call control affects the emergency call, the PoC, and the VCC applications. This dissertation is organized as follows.

In Chapter 2, we study the location tracking procedure for IMS emergency call. After emergency call setup, the UE's location is tracked by the PSAP through Location Polling, where the PSAP periodically queries the UE's location. We present an analytic model and a simulation model to investigate the performance of Location Polling. To improve the performance of location tracking, we propose the Active Location Reporting scheme that reports the UE's location upon change of its SA.

In Chapter 3, we study the performance of the talk burst control mechanism for IMS PoC. The PoC communication is half-duplex; only one PoC member speaks at a time, and the others listen. The speak permission is arbitrated through the talk burst control mechanism. We propose analytic and simulation models to study the performance of the talk burst control mechanism. Our study provides guidelines to set up the parameters for IMS PoC.

In Chapter 4, we focus on the domain transfer procedures between the CS and the PS domains. Domain transfer may produce a large number of message exchanges and resource reservation that result in long switching latency. To resolve this issue, we propose the Bearer Reservation with Preemption scheme to support fast domain transfer. We also present both analytic model and simulation experiments to investigate the performance of the Bearer Reservation with Preemption scheme.

Finally, we conclude this dissertation in Chapter 5 by discussing our contributions and future work.

Chapter 2

Active Location Reporting for Emergency Call

Emergency call is one of the essential telephony services for telecommunications operations. In IMS, an emergency call is established by an *Emergency-Call Session Control Function* (E-CSCF). The E-CSCF dispatches the call to the nearest *Public Safety Answering Point* (PSAP) according to the location of the caller. After emergency call setup, the PSAP may need to monitor the UE's location in real time. The UE's location is tracked by the PSAP through Location Polling, where the PSAP periodically queries the UE's location. In the Location Polling scheme, if the UE does not change its location between two queries, the second query is wasted. On the other hand, if the UE has moved to several new locations between two location queries, then the PSAP may lose track of the UE. This chapter investigates the performance of location tracking. Then we propose the Active Location Reporting scheme to improve the performance of location tracking. Our study indicates that the Active Location Reporting scheme may significantly outperform the Location Polling scheme.

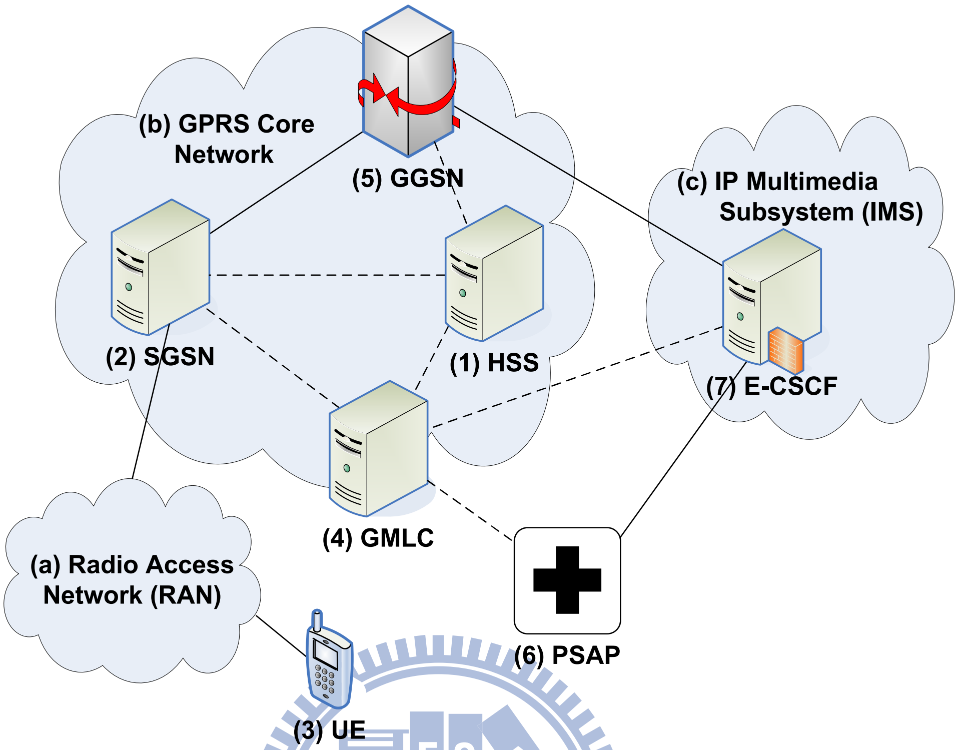


Figure 2.1: The UMTS Network Architecture for IMS Emergency Call

2.1 Introduction to IMS Emergency Call

Figure 2.1 illustrates a simplified UMTS network architecture that accommodates IMS emergency call [2, 4, 1]. This architecture consists of a *Radio Access Network* (RAN; Figure 2.1 (a)), the *General Packet Radio Service* (GPRS) core network (Figure 2.1 (b)) and the IMS network (Figure 2.1 (c)). The *Home Subscriber Server* (HSS; Figure 2.1 (1)) is the master database containing all user-related subscription information. The *Serving GPRS Support Node* (SGSN; Figure 2.1 (2)) is responsible for packet delivery to and from *User Equipments* (UEs; Figure 2.1 (3)). The *Gateway Mobile Location Center* (GMLC; Figure 2.1 (4)) supports *Location Service* (LCS). The GPRS core network connects to the IMS network through *Gateway GPRS Support Nodes* (GGSNs; Figure 2.1 (5)). To establish a data session, a *Packet Data Protocol* (PDP) *Context* must be activated before a UE can access the IMS network. A *Public Safety Answering Point* (PSAP; Figure 2.1 (6)) dispatches emergency calls according to the types of emergent events (e.g., fire). When the UE originates an emergency call, the call is established by an *Emergency-Call Session Control Function* (E-CSCF; Figure 2.1 (7)), which dispatches the call to the nearest PSAP according to the location of the UE.

The LCS utilizes one or more positioning methods [7, 8] between the RAN and the UE to determine the location of a UE. Four positioning methods are specified in 3GPP TS 25.305 [7]. These methods are briefly described as follows:

The Cell-ID-based method determines the UE's position based on the coverage of Service Areas (SAs). An SA includes one or more cells (base stations). At most one-cell-sized accuracy (about 500 meters) can be achieved when the SA includes only one cell.

The Observed Time Difference of Arrival (OTDOA) method utilizes trilateration to determine the UE's position. At least three concurrent downlink signals from different cells are measured in the UE. The time differences among the signal arrivals are calculated to form hyperbolic curves. The intersection of these curves is then used to indicate the UE's position. This method provides location accuracy within 50-150 meters.

The Assisted Global Positioning System (A-GPS) method speeds up GPS positioning by downloading GPS information through the RAN. Execution of A-GPS positioning only requires several seconds while execution of normal GPS positioning requires 30 seconds to several minutes. GPS modules are installed in both the UE and the RAN. This method provides location accuracy within 5-15 meters.

The Uplink Time Difference of Arrival (U-TDOA) method evolves from the OTDOA method. This method utilizes uplink signals instead of downlink signals. A normal uplink signal from the UE is measured in different cells, and no extra signal is required. Same calculation process as OTDOA is then conducted to find out the UE's position. Since the measurement and the calculation process are exercised only in the RAN, this method does not require any modification to the mobile phone. This method provides location accuracy within 50-150 meters.

Without loss of generality, we consider the Cell-ID-based method in this dissertation (the Cell-ID-based method supports all kinds of UEs, while other methods may require UEs to measure extra signal). The conclusions also apply to other positioning methods except that the accuracy of location measured in Cell-ID-based method is SA, while the accuracy measured in, for example, OTDOA is meter.

2.2 Emergency Call Setup and Location Tracking

This section describes the IMS emergency call setup and location tracking procedures. We first elaborate on the call setup procedure and the Location Polling scheme proposed in 3GPP [13, 14]. Then we propose the Active Location Reporting scheme that improves the performance of location tracking.

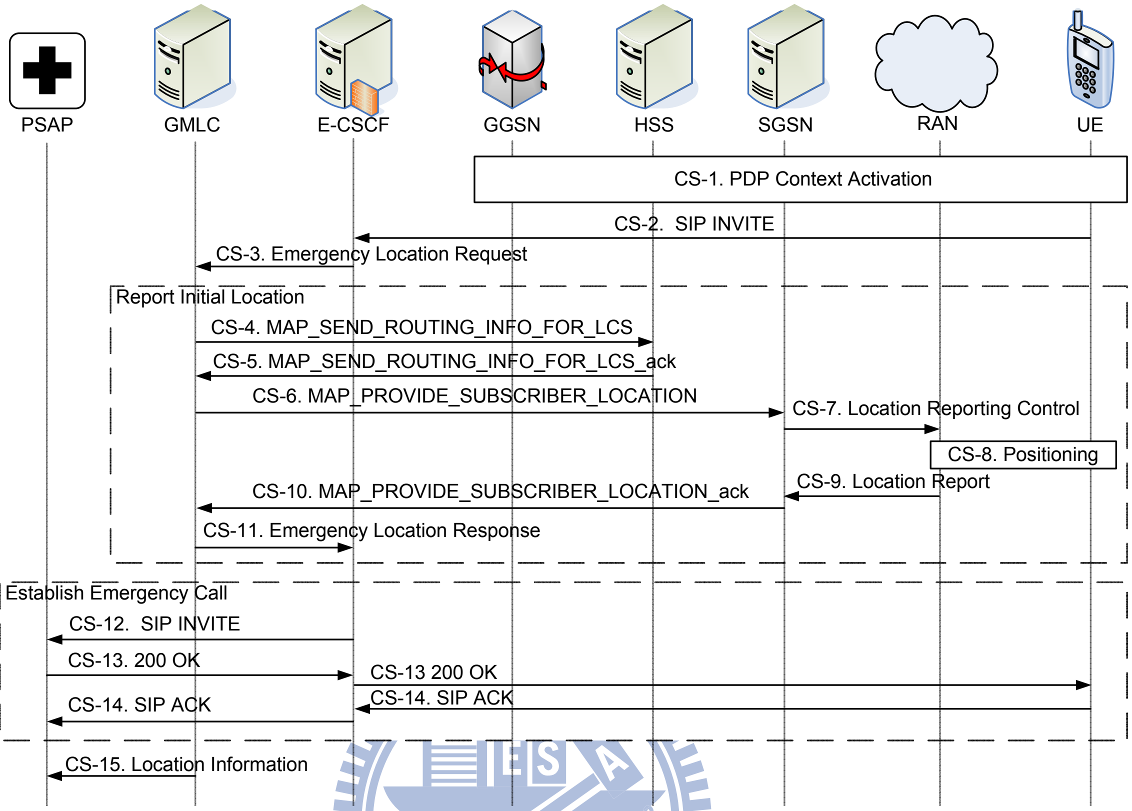


Figure 2.2: IMS Emergency Call Setup

2.2.1 Emergency Call Setup

Before the IMS emergency call is set up, the UE has attached to the network through the RAN. Figure 2.2 illustrates IMS emergency call setup message flow defined in 3GPP [13] with the following steps:

Step CS-1. The UE performs PDP context activation that establishes the IP connectivity to the IMS through the GPRS network [4].

Step CS-2. The UE sends the SIP INVITE message to the E-CSCF. This message includes the supported positioning methods of the UE (i.e., Cell-ID-based in our example).

Step CS-3. The E-CSCF uses the received information to determine a GMLC and sends the Emergency Location Request message to the GMLC, which includes all UE-related information received at Step CS-2.

Steps CS-4 and 5. The GMLC exchanges the MAP_SEND_ROUTING_INFO_FOR_LCS and MAP_SEND_ROUTING_INFO_FOR_LCS_ack message pair with the HSS to obtain the SGSN address for the UE.

Step CS-6. The GMLC sends the MAP_PROVIDE_SUBSCRIBER_LOCATION message to the SGSN to request the UE's location. In this message, the locationEstimateType parameter is set to "initialLocation".

Step CS-7. Upon receipt of the MAP_PROVIDE_SUBSCRIBER_LOCATION message, the SGSN sends a Location Reporting Control message to the RAN to trigger the positioning procedure. In this message, the Request Type is set to "report directly".

Step CS-8. The RAN and the UE exercise the Cell-ID-based positioning procedure to obtain the location estimate information of the UE (i.e., the SA identity of the UE).

Step CS-9. The RAN returns the Location Report message with the SA identity to the SGSN.

Step CS-10. The SGSN returns the SA identity to the GMLC through the MAP_PROVIDE-_SUBSCRIBER_LOCATION_ack message.

Step CS-11. The GMLC selects a suitable PSAP according to the SA of the UE and replies the Emergency Location Response message with the selected PSAP address to the E-CSCF.

Steps CS-12-14. The E-CSCF forwards the SIP INVITE to the PSAP. The PSAP and the UE exchange the 200 OK and the SIP ACK messages through the E-CSCF. After the PSAP has received the SIP ACK message, the emergency call is established.

Step CS-15. The GMLC sends the location information obtained at Step CS-10 to the PSAP after the call has been established.

2.2.2 Location Polling

A UE may move during an emergency call, and the PSAP may need to monitor the UE's location in real time. In 3GPP TS 23.271 [14], the UE's location is monitored through a polling procedure where the PSAP periodically queries the UE's location. In each polling query, the following steps are executed (see Figure 2.3).

Step LP-1. The PSAP sends the Location Information Request message to the GMLC.

Steps LP-2-8. These steps are similar to Steps CS-4-10 in Figure 2.2 except that the parameter locationEstimateType in the MAP_PROVIDE_SUBSCRIBER_LOCATION message is set to "currentLocation".

Step LP-9. The GMLC returns the SA identity of the UE to the PSAP.

Steps LP-10-12. When the emergency call is terminated, the E-CSCF exchanges the Emergency Location Release and Response message pair with the GMLC to terminate location tracking.

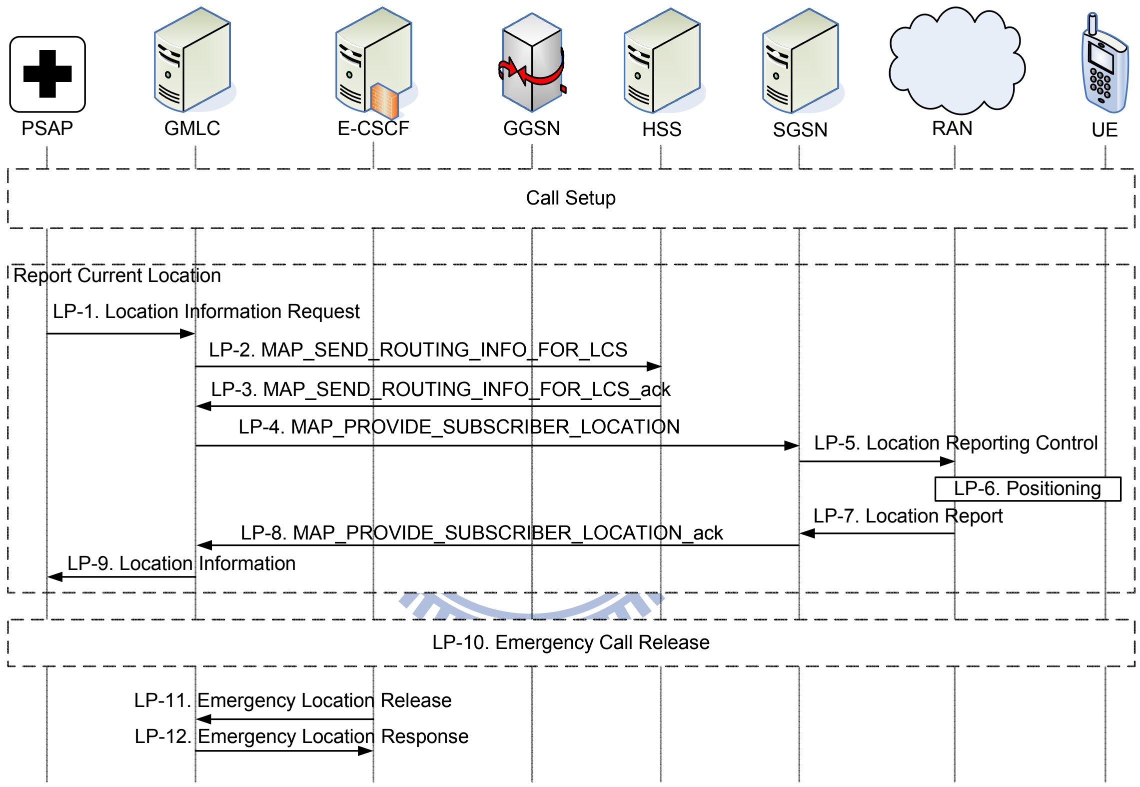


Figure 2.3: Location Polling

2.2.3 Active Location Reporting

In the Location Polling scheme, if the UE does not change its location between two queries, the second query is wasted. On the other hand, if the UE has moved to several new locations between two location queries, then the PSAP may lose track of the UE. To resolve this issue, we propose the Active Location Reporting scheme that reports the UE's location upon change of its SA. Our scheme introduces two new locationEstimateTypes "initiateActiveReport" (to trigger Active Location Reporting) and "terminateActiveReport" (to terminate Active Location Reporting) in the MAP_PROVIDE_SUBSCRIBER_LOCATION message, and one new LCS event type "ActiveReporting" (to indicate the Active Location Reporting event). At emergency call setup, the locationEstimateType is set to "initiateActiveReport" at Step CS-6, and the Request Type is set to "change of service area" at Step CS-7. Since the IP connectivity exists during the IMS emergency call, the UE is in the Cell-Connected state and is tracked by the RAN at the cell level [4]. Therefore, the RAN can detect the movement of the UE at the cell level and report the new SA identity to the SGSN. In this approach, the GMLC maintains a UE-PSAP mapping table, and the (UE, PSAP) pair is stored in the GMLC at Step CS-11. The GMLC does not need to query the HSS to obtain the SGSN address of the UE (therefore, Steps LP-2 and LP-3 are eliminated). The Active Location Reporting scheme is illustrated in Figure 2.4 with the following steps:

Step ALR-1. When the UE moves to a new SA, the RAN detects this movement at the cell tracking mode [4] and then triggers the positioning procedure.

Step ALR-2. After the positioning procedure is executed, the UE's SA identity is obtained.

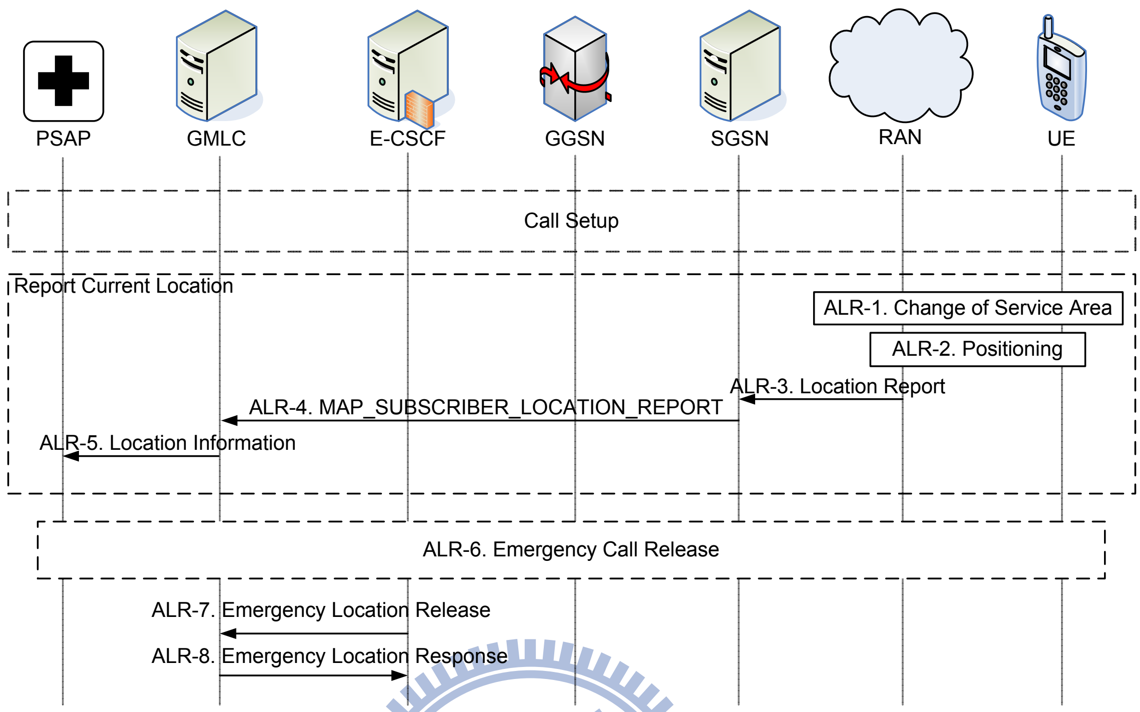


Figure 2.4: Active Location Reporting

Step ALR-3. The RAN sends the Location Report message with the SA identity of the UE to the SGSN.

Step ALR-4. The SGSN sends the MAP_SUBSCRIBER_ LOCATION_REPORT message with the SA identity to the GMLC.

Step ALR-5. From the UE-PSAP mapping table, the GMLC retrieves the PSAP address of the UE stored at Step CS-11 and then sends the updated location information to the PSAP.

When the emergency call is terminated, the following steps are executed.

Step ALR-6. When the IMS call is released, the UE moves from the Cell-Connected mode to the Idle mode, and the RAN no longer tracks the movement of the UE [4].

Step ALR-7. The E-CSCF sends the Emergency Location Release message to the GMLC to terminate location tracking.

Step ALR-8. The GMLC returns the Emergency Location Response message to the E-CSCF and then deletes the (UE, PSAP) mapping from the UE-PSAP table.

The major difference between Active Location Reporting and Location Polling is at Steps ALR-1 and ALR-2. Active Location Reporting is triggered when the RAN detects the movement of the UE (through the standard tracking procedure at the Cell-Connected mode). Note that the HSS query (see Steps LP-2 and LP-3 in Figure 2.3) is not required for the Active Location Reporting scheme because of the active reporting of the RAN. The GMLC needs to maintain the (UE, PSAP) mapping so that when a UE changes the SA, the GMLC can report the location update to the corresponding PSAP.

2.3 Analytic Modeling of Location Polling

This section proposes an analytic model to study the performance of location tracking. Let N be the number of queries between two SA crossings. Five output measures are considered.

- α: the probability of *mis-tracking* for an SA crossing. An SA crossing is mis-tracked if there is no query between this SA crossing and the next SA crossing, and therefore the system does not know that the user has moved to this SA. It is clear that $\alpha = \Pr[N = 0]$.
- T_i: the expected *invalid* period. The invalid period is defined as the period between when an SA crossing occurs and when the next query arrives under the condition

that $N \geq 1$. In this period, the system does not know that the user has moved (i.e., the location known by the system is obsolete and therefore is "invalid").

- V_i: the variance of the invalid periods

- β: the probability that *redundant* queries exist between two SA crossings (i.e., $\beta = \Pr[N > 1]$). It is clear that redundant queries create extra network traffic without providing useful location information.

- N_R: the expected number of queries between two SA crossings under the condition that $N > 1$. In other words, $N_R - 1$ is the expected number of redundant queries.

The smaller the above output measure values, the better the performance of location tracking. It is clear that for the Active Location Reporting scheme, optimal performance is achieved for these output measures, that is, $\alpha = 0$, $T_i = 0$, $V_i = 0$, and $\beta = 0$. On the other hand, the above output measure values are not 0 for the Location Polling scheme. This section derives α, T_i, V_i, β, and N_R for Location Polling. Then we investigate if Active Location Reporting (the optimal case) significantly outperforms Location Polling.

Figure 2.5 illustrates the relationship between location queries and user movements. The UE changes its SA at t_1, t_5 and t_6, and the PSAP queries the UE's location at t_0, t_2, t_3 and t_4. In this example, $N = 3$ between t_1 and t_5, and $N = 0$ between t_5 and t_6. Let the SA residence time interval $t_m = t_5 - t_1$ be a random variable with the density function $f_m(\cdot)$ and the Laplace transform $f_m^*(\cdot)$. Let the inter-query interval $t_q = t_2 - t_0$ be a random variable with the exponential distribution with the mean $1/\lambda$ (i.e., the query stream forms a Poisson process). Then α is derived as

$$\alpha = \Pr[N = 0] = \int_{t_m=0}^{\infty} e^{-\lambda t_m} f_m(t_m) dt_m = f_m^*(\lambda) \tag{2.1}$$

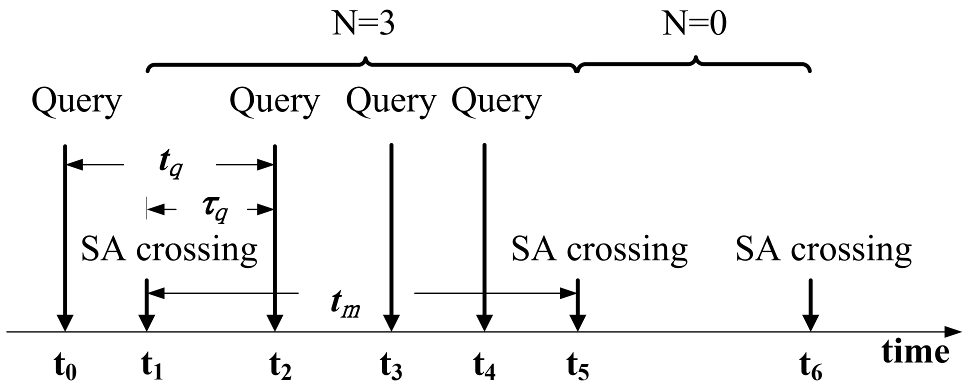


Figure 2.5: Timing Diagram for Location Polling

If t_m has Gamma distribution with the mean $1/\mu$ and the variance V_m, then (2.1) is re-written as

$$\alpha = \left(\frac{1}{V_m\mu\lambda + 1}\right)^{\frac{1}{V_m\mu^2}}$$

The Gamma distribution is selected because it has been shown that the distribution of any positive random variable can be approximated by a mixture of Gamma distributions (see Lemma 3.9 in [15]). Following the past experiences [16, 17, 18], we can obtain the SA residence time samples from the commercial mobile telecommunication operation and then use the statistical tools to fit the sample data by the Gamma distribution.

In Figure 2.5, $\tau_q = t_2 - t_1$ is the invalid period for the SA residence time interval $[t_1, t_5]$. In this period, the PSAP is not aware of the SA crossing at t_1. Since the query stream is a Poisson process, the density function $r_q(\cdot)$ for invalid period τ_q is the same as that for t_q because of the memoryless property of the exponential distribution. Consider the conditional density function $r_{q|N\geq 1}(\tau_q)$ where the PSAP issues more than one query in the SA residence time interval. From (2.1) and because $\Pr[N \geq 1] = \Pr[\tau_q \leq t_m]$, we have

$$r_{q|N\geq 1}(\tau_q) = \left[\frac{1}{1 - f_m^*(\lambda)}\right] \int_{t_m=\tau_q}^{\infty} r_q(\tau_q) f_m(t_m) dt_m \tag{2.2}$$

Based on (2.2) and the memoryless property of the exponential distribution, we derive the expected invalid period T_i as

$$\begin{aligned} T_i &= E[\tau_q | N \geq 1] \\ &= \int_{\tau_q=0}^{\infty} \tau_q r_{q|N\geq 1}(\tau_q) d\tau_q \\ &= \left[\frac{1}{1-f_m^*(\lambda)}\right] \left[\left.\frac{df_m^*(s)}{ds}\right|_{s=\lambda}\right] + \frac{1}{\lambda} \end{aligned} \tag{2.3}$$

If t_m has a Gamma distribution, (2.3) is re-written as

$$T_i = \frac{1}{\lambda} - \frac{1}{\mu(V_m\mu\lambda+1)^{\frac{1}{V_m\mu^2}+1} - V_m\mu^2\lambda - \mu}$$

Similar to the derivation for T_i, we derive the variance V_i of the invalid periods as

$$\begin{aligned} V_i &= V[\tau_q | N \geq 1] \\ &= \frac{2(V_m\mu\lambda+1)^{\frac{1}{V_m\mu^2}}}{\lambda^2\left[(V_m\mu\lambda+1)^{\frac{1}{V_m\mu^2}} - 1\right]} - \frac{3V_m\lambda^2\mu^2 + 2\lambda\mu + \lambda^2 + 2\mu^2(V_m\mu\lambda+1)^2}{\lambda^2\mu^2(V_m\lambda\mu+1)^2\left[(V_m\mu\lambda+1)^{\frac{1}{V_m\mu^2}} - 1\right]} \\ &\quad - \left[\frac{1}{\lambda} - \frac{1}{\mu(V_m\mu\lambda+1)^{\frac{1}{V_m\mu^2}+1} - V_m\mu^2\lambda - \mu}\right]^2 \end{aligned}$$

The probability β of redundant queries is derived as

$$\begin{aligned} \beta &= \Pr[N > 1] \\ &= 1 - \Pr[N=1] - \Pr[N=0] \\ &= 1 - \int_{t_m=0}^{\infty} \lambda t_m e^{-\lambda t_m} f_m(t_m) dt_m - \alpha \\ &= 1 + \lambda\left[\left.\frac{df_m^*(s)}{ds}\right|_{s=\lambda}\right] - f_m^*(\lambda) \end{aligned} \tag{2.4}$$

If t_m has a Gamma distribution, (2.4) is re-written as

$$\beta = 1 - \left(\frac{1}{V_m\lambda\mu+1}\right)^{\frac{1}{V_m\mu^2}}\left[1 - \frac{\lambda}{\mu(V_m\lambda\mu+1)}\right]$$

Table 2.1: Comparison of Analytic and Simulation Models ($V_m = 1/\mu^2$)

λ	0.1μ	10μ
α (Analytic)	0.90909	0.090909
α (Simulation)	0.90914	0.090910
Error	0.0063 %	0.0013 %
T_i (Analytic)	$0.909/\mu$	$0.0909/\mu$
T_i (Simulation)	$0.907/\mu$	$0.0908/\mu$
Error	0.1374 %	0.1119 %
V_i (Analytic)	$0.8264/\mu^2$	$0.008265/\mu^2$
V_i (Simulation)	$0.8257/\mu^2$	$0.008271/\mu^2$
Error	0.0907 %	0.0804 %
β (Analytic)	0.00826	0.82645
β (Simulation)	0.00828	0.82657
Error	0.1875%	0.0153 %
N_R (Analytic)	2.1	12
N_R (Simulation)	2.103	11.99
Error	0.1794 %	0.0687 %

Since the query stream is a Poisson process, N has Poisson distribution with mean λt_m. Therefore N_R is derived as

$$\begin{aligned} N_R &= E[N|N>1] \\ &= \sum_{n=2}^{\infty} n \left[\frac{\int_{t_m=0}^{\infty} \frac{(\lambda t_m)^n}{n!} e^{-\lambda t_m} f_m(t_m) dt_m}{\beta} \right] \\ &= \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu} \left[\frac{V_m\mu^2+1}{\mu(V_m\mu\lambda+1)^{\frac{1}{V_m\mu^2}+1} - V_m\mu^2\lambda - \mu - \lambda} \right] \end{aligned} \tag{2.5}$$

The above analytic model is validated against the discrete event simulation experiments. The discrete event simulation model is described in Appendix A. As shown in Table 2.1 (where $V_m = 1/\mu^2$), the analytic analysis is consistent with the simulation results.

2.4 Numerical Examples

Based on the simulation experiments validated against the analytic model, this section investigates the performance of Location Polling. Two types of inter-query intervals can be considered. *Fixed* polling queries the UE's location with fixed period $1/\lambda$. On the other hand, in *exponential* polling, the inter-query interval has the exponential distribution with the mean $1/\lambda$. In this section, we first compare fixed polling with exponential polling. Our study will indicate that fixed polling outperforms exponential polling. Then we compare fixed polling with the Active Location Reporting scheme. We will show that Active Location Reporting scheme outperforms fixed polling.

Assume that the SA residence time interval t_m has the exponential distribution with the mean $1/\mu$ (i.e., the SA crossings form a Poisson process). From (2.1), the probability α for exponential polling is expressed as

$$\alpha = f_m^*(\lambda) = \frac{\mu}{\lambda + \mu} \tag{2.6}$$

Since the SA crossings form a Poisson process, the number X of SA crossings in an arbitrary fixed inter-query interval $1/\lambda$ has a Poisson distribution with the mean μ/λ. When $X > 0$, there are $X - 1$ SA residence time intervals without any query. Therefore, the probability α for fixed polling is expressed as

$$\alpha = \frac{E[X-1|X>0]}{E[X|X>0]} = 1 - \frac{\lambda}{\mu}\left(1 - e^{-\mu/\lambda}\right) \tag{2.7}$$

Figure 2.6 plots α for fixed and exponential polling approaches based on (2.6) and (2.7). The figure indicates that fixed polling outperforms exponential polling in terms of the α measure. For all cases considered, fixed polling outperforms exponential polling when λ is large (e.g., $\lambda = 10\mu$), while both approaches have similar performance when λ is small.

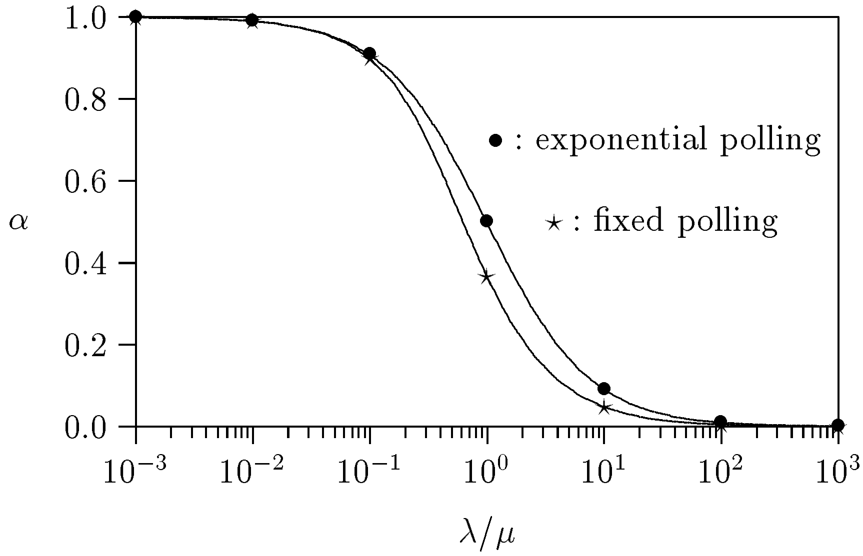


Figure 2.6: Comparing Fixed and Exponential Inter-Query Interval (Poisson SA Crossing Stream with the Rate μ)

Detailed comparison between fixed polling and exponential polling will not be presented in this dissertation. Instead, this dissertation uses the analytic model based on exponential polling to validate against the simulation model. In the remainder of this chapter, we only consider fixed polling through simulation experiments. General conclusions drawn from this chapter also apply to exponential polling. The effects of the input parameters are investigated as follows.

Effects of λ and V_m on the mis-tracking probability α : Figure 2.7 plots α for Gamma SA residence time intervals with different variance values. The figure indicates that α increases as V_m increases. This phenomenon is explained as follows. When the SA residence time intervals become more irregular (i.e., V_m increases), we will observe more SA residence time intervals without any query. On the other hand, the number of SA residence time intervals with queries will not increase as V_m increases, but the number of queries in an SA residence time interval will increase. Therefore, larger α

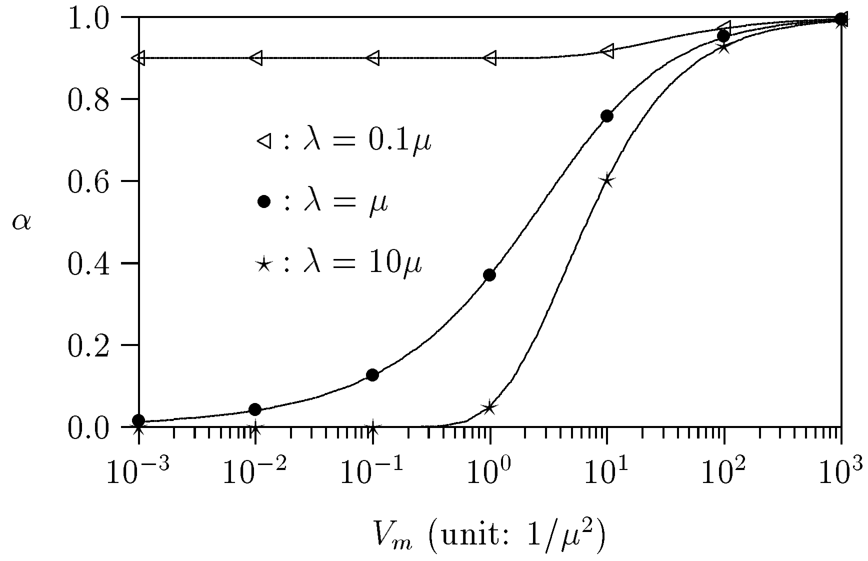


Figure 2.7: Effects of λ and V_m on α

is observed. The figure also indicates that when V_m is very large, α is not sensitive to the λ values, and poor accuracy is always observed (i.e., α is large).

Effects of λ and V_m on the expected invalid period T_i : Figure 2.8 plots T_i against λ and V_m. When λ is small (e.g., $\lambda = 0.1\mu$), T_i increases as V_m increases. This phenomenon is explained as follows. As V_m increases, more long and short SA residence time intervals are observed. For short SA residence time intervals, it is likely that $N = 0$, and these intervals will not contribute to τ_q. In other words, when V_m increases, more long τ_q intervals are observed. Therefore, T_i increases as V_m increases. When λ is large (e.g., $\lambda = 10\mu$), T_i becomes less sensitive to V_m.

Effects of λ and V_m on the variance V_i of the invalid periods : Figure 2.9 plots V_i against λ and V_m. When λ is small (e.g., $\lambda = 0.1\mu$), V_i increases as V_m increases. When λ is large (e.g., $\lambda = 10\mu$), V_i becomes less sensitive to V_m. This phenomenon is similar to that for T_i.

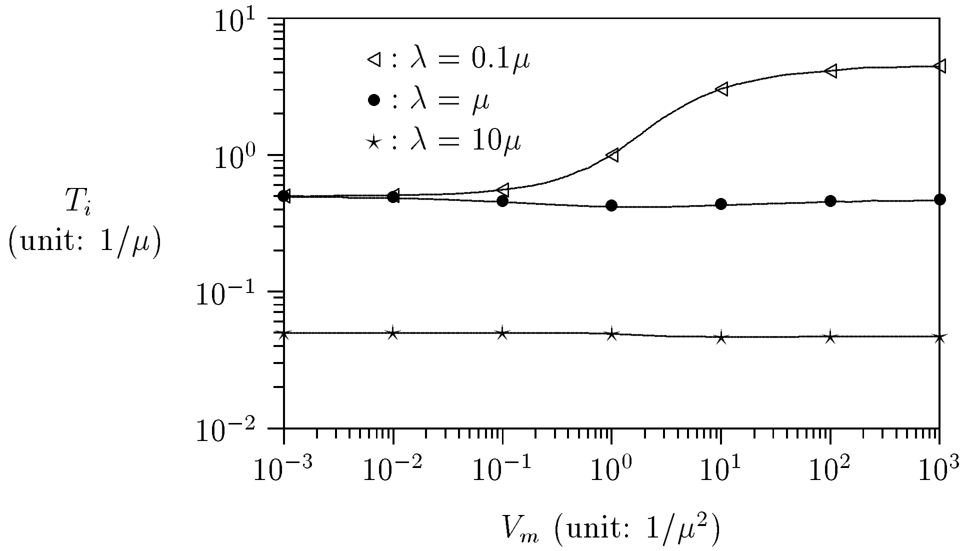


Figure 2.8: Effects of λ and V_m on T_i

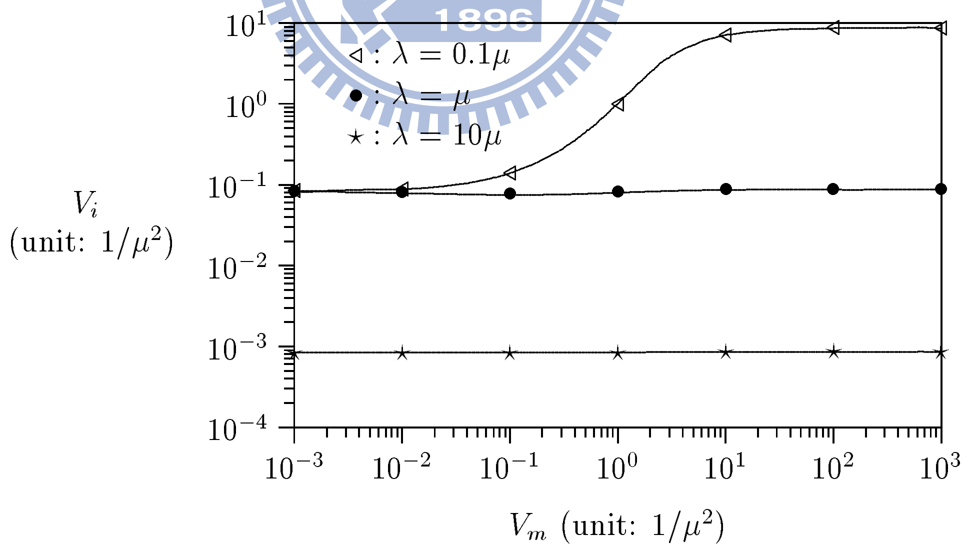


Figure 2.9: Effects of λ and V_m on V_i

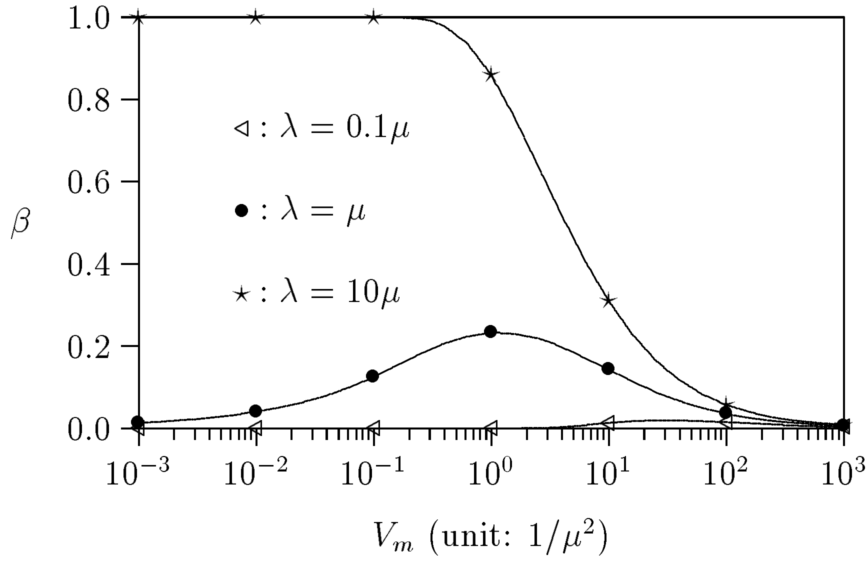


Figure 2.10: Effects of λ and V_m on β

Effects of λ and V_m on the probability β of redundant query : Figure 2.10 plots β against λ and V_m. As V_m increases, two effects are observed: (I) More SA residence time intervals without any query are observed, which results in smaller β, (II) More SA residence time intervals with more than one query are observed, which results in larger β. When λ is large (e.g., $\lambda = 10\mu$), Effect (I) is more significant than Effect (II). Therefore, β is a decreasing function of V_m. For $\lambda = \mu$, β increases and then decreases as V_m increases. When V_m is small, Effect (II) is more significant, while Effect (I) is more significant when V_m is large. Therefore, β increases and then decreases as V_m increases. When λ is small (e.g., $\lambda = 0.1\mu$), both Effects (I) and (II) are insignificant.

Effects of λ and V_m on N_R : Figure 2.11 shows that N_R is an increasing function of V_m. This phenomenon is explained as follows. As V_m increases, more long SA residence time intervals are observed. Since query events are more likely to fall on long SA residence time intervals, larger N_R is observed. Therefore, N_R increases as

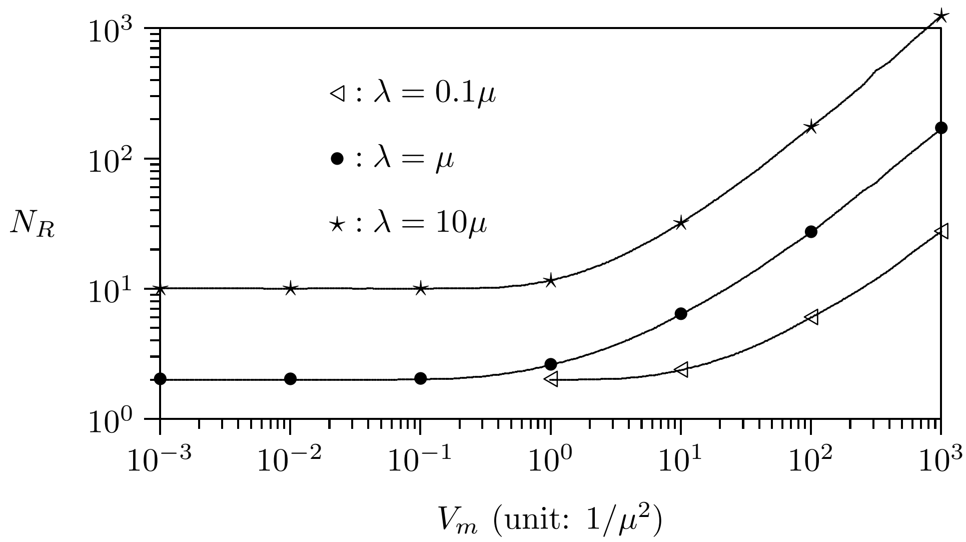


Figure 2.11: Effects of λ and V_m on N_R

V_m increases. When V_m is small, (i.e., $V_m \le 1/\mu^2$), N_R is not sensitive to the change of V_m.

2.5 Summary

This chapter investigated emergency call mechanism for IMS. After an IMS emergency call is established, the caller's location is tracked by the PSAP through Location Polling. We proposed the Active Location Reporting scheme to improve the performance of location tracking. Our study indicated that the Active Location Reporting scheme significantly outperforms the Location Polling scheme. We observed the following results:

- When the query frequency is low (i.e., λ is small) or when the movement is irregular (i.e., when V_m is large), Active Location Reporting significantly outperforms Location Polling in terms of the α (mis-tracking probability) performance.

- When the query frequency is low, Active Location Reporting significantly outperforms Location olling in terms of the T_i and V_i (for the invalid period) performance.

- When the query frequency is high and when the movement is regular, Active Location Reporting significantly outperforms Location Polling in terms of the β (redundant query probability) performance.

Chapter 3

Talk Burst Control for Push-to-Talk over Cellular

Push-to-talk over Cellular (PoC) provides a "walkie-talkie"-like group communication method in the IMS. The PoC communication is half-duplex, where only one PoC member speaks at a time, and the others listen. The speak permission is arbitrated through a talk burst control mechanism. In the PoC service, the talk burst control mechanism is implemented by finite state machines in both the server and the client sides. To obtain the speak permission, a PoC group member sends a talk burst request message to the PoC server for arbitration. The PoC server may grant the permission or deny the request depending on the current talk burst status (indicated by the server-side finite state machine). A queueing option is provided in the talk burst control mechanism. If this option is selected, then the ungranted requests are buffered in the queue at the PoC server instead of being denied. In this chapter, we propose an analytic model to study the performance of the talk burst control mechanism with queueing and without queueing. This analytic model is validated against simulation experiments. Through numerical examples, our study provides guidelines to set up the parameters for IMS PoC service.

3.1 Introduction to PoC Service

PoC is a "walkie-talkie"-like service defined in Open Mobile Alliance (OMA) specifications [9, 10]. In this service, several predefined PoC group members participate in one PoC session. Since the PoC communication is half-duplex, only one PoC member speaks at a time, and the others listen. When a PoC member attempts to speak, he/she presses the push-to-talk button of his/her mobile terminal to ask for the permission. This mobile terminal installed with the PoC application is called the PoC client. The speak permission is arbitrated through the talk burst control mechanism [10].

A PoC group includes a predefined set of group members, and is identified by a Telephone Universal Resource Identifier (TEL URI; e.g., tel: +88635131350) or a SIP URI (e.g., sip:PoCGroup1@pcs1.csie.nctu.edu.tw). SIP URI of each PoC group member is maintained in a group member list. The PoC group information and its group member list are stored in the *PoC XML Document Management Server* (XDMS).

The *PoC server* is responsible for handling PoC session management (create or delete a PoC session) and talk burst control. Session Initiation Protocol (SIP) [3] and Session Description Protocol (SDP) [19] are utilized for session establishment. To support the PoC service, a new parameter "tb_grant" is added in SDP's attribute field such that PoC server can arbitrate the speak permission during session establishment. If "tb_grant=1", the PoC client is granted the permission to speak. Otherwise (i.e., "tb_grant=0"), the PoC client is not permitted to talk.

Based on the OMA specifications, an Open Service Access (OSA)-based PoC system is implemented in the *Industrial Technology Research Institute* (ITRI) and *National Chiao-Tung University* (NCTU) Joint Research Center. The OSA-based PoC server architecture

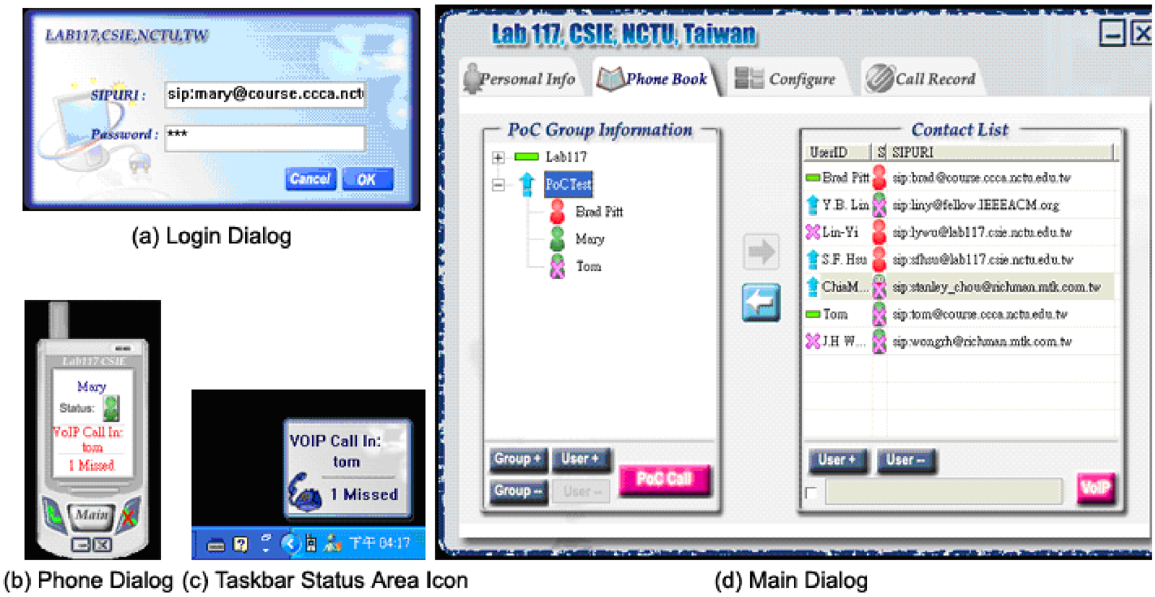


Figure 3.1: The User Interface of the PoC Client

is described in [20]. Figure 3.1 shows the user interface of the PoC client. The Login dialog (Figure 3.1 (a)) is popped up after the PoC client program is executed. It waits for the user to input the SIP URI (for the PoC client) and the password, and then executes the SIP registration procedure. After the user is authenticated, the PoC client enters the standby mode, and the Phone dialog (Figure 3.1 (b)) is activated. After closing the phone dialog, the user can select to open the Main dialog (Figure 3.1 (d)) or the taskbar status area icon (Figure 3.1 (c)). In the taskbar status area icon mode, when a call arrives or a call is missed, a message window is popped up to notify the user. The Main dialog contains five pages: the personal information page, the phone book page, the configuration page, the call record page, and the PoC/VoIP call page (not shown in Figure 3.1 (d)). The PoC/VoIP call page is created when a call is established. This page presents the call information, and allows the PoC user to request for the speak permission. The readers are referred to [21] for the detailed descriptions.

PoC is a new service in cellular networks. To our knowledge, all PoC studies have focused on call setup time and transmission delay of voice packets [22, 23, 24]. These studies did not consider the performance of talk burst control mechanism. Many talk burst-related questions are not answered in these previous studies. These questions include the best setting of revoking time T_R, the maximum number M of group members that a mobile operator should support, etc. In the following sections, we describe the talk burst control mechanism, and then propose an analytic model to study the talk burst control mechanism.

3.2 Talk Burst Control Mechanism

In the OMA specification, talk burst control mechanism is implemented by finite state machines (FSMs) in both the server and the client sides. Figures 3.2 and 3.3 illustrate simplified talk burst control FSMs for PoC server (called FSM_G) and PoC client (called FSM_U), where "U" represents "user" (i.e., for a specific PoC client) and "G" represents "general" (i.e., for the whole group). The prefixes "S:" and "R:" of the transitions represent "send" and "receive", respectively. For a PoC session, there is one FSM_G in the PoC server and an FSM_U in each of the PoC clients. To clearly describe the PoC procedures, several terms are defined:

- The *session initiator* is the PoC client who initiates a PoC session.
- An *invited PoC client* is a PoC group member other than the session initiator.
- A *requesting PoC client* is a PoC client who requests for speak permission.
- The *permitted PoC client* is the PoC client who is allowed to speak.

- A *listening PoC client* is a PoC client who is not permitted to speak.

- A *queued PoC client* is a PoC client whose request is queued in the PoC server.

For the PoC clients and the PoC server involved in a PoC session, their FSMs are initialized at the **Start-stop** state (see Figures 3.2 and 3.3). When the session initiator sends a SIP INVITE message to the PoC server, the PoC server broadcasts SIP INVITE messages with "tb_grant=0" to other group members (the invited PoC clients). Each of the invited PoC clients answers with a SIP 200 OK message and its FSM_U enters **U: has no permission** (transition 2 in Figure 3.3). This state means that the PoC client is not permitted to speak.

After receiving the first SIP 200 OK message from the invited PoC clients, the PoC server replies a SIP 200 OK with "tb_grant=1" to the session initiator and FSM_G enters **G: TB_Taken** (transition 1 in Figure 3.2). This state means that some PoC client (the session initiator in this case) has obtained the permission. FSM_U of the session initiator enters **U: has permission** (transition 1 in Figure 3.3). This state means that the PoC client is allowed to speak. The session initiator becomes the permitted PoC client, and the invited PoC clients become listening PoC clients. At this moment, the session initiator speaks and all invited PoC clients listen.

After the PoC session is established, each of the PoC clients has built a Real-time Transport Protocol (RTP) [11] session with the PoC server. The talk burst control messages (with the prefix "TB") between the PoC clients and the PoC server are carried by RTP Control Protocol (RTCP) packets [11]. After finishing the talk, the permitted PoC client releases the permission by sending the TB_Release message to the PoC server and its FSM_U enters **U: pending TB_Release** (transition 6 in Figure 3.3). In this

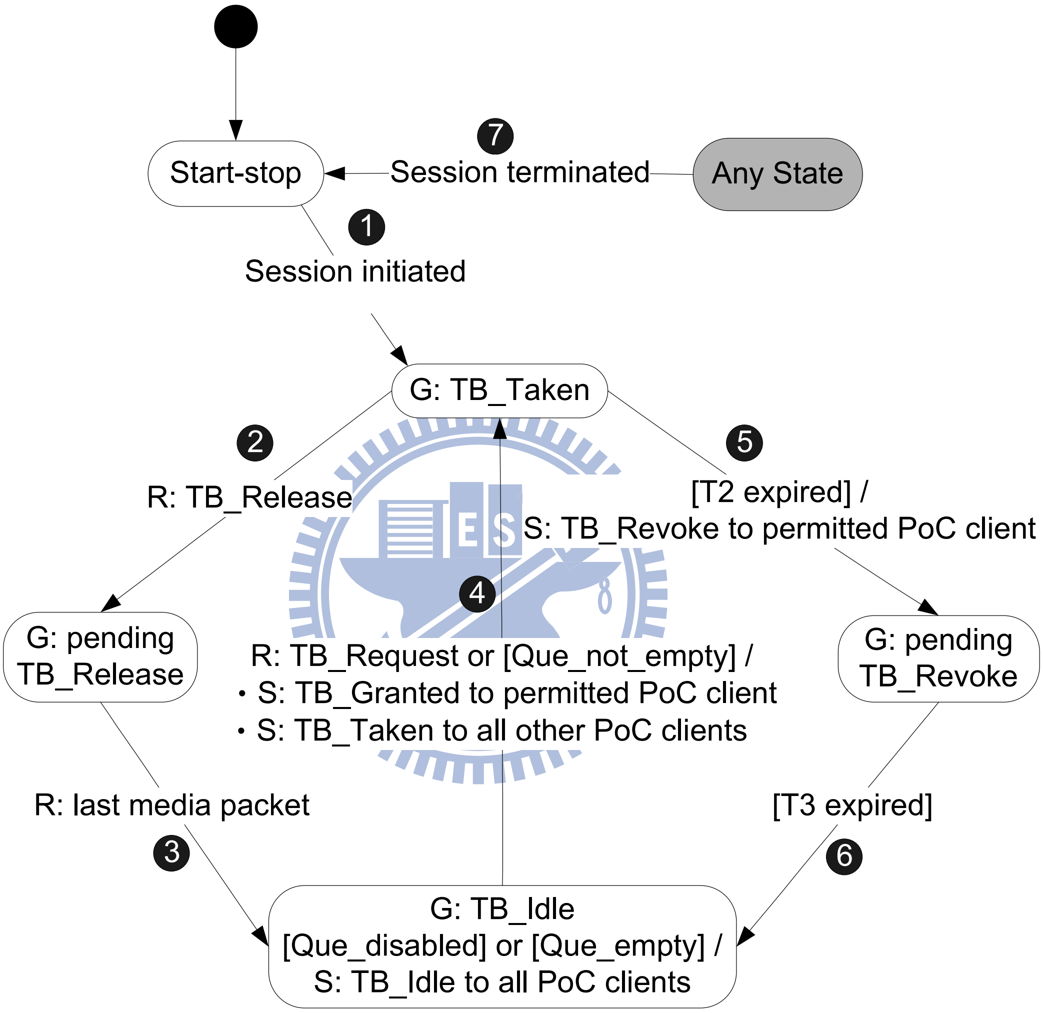


Figure 3.2: Talk Burst Control Finite State Machine for PoC Server (FSM_G)

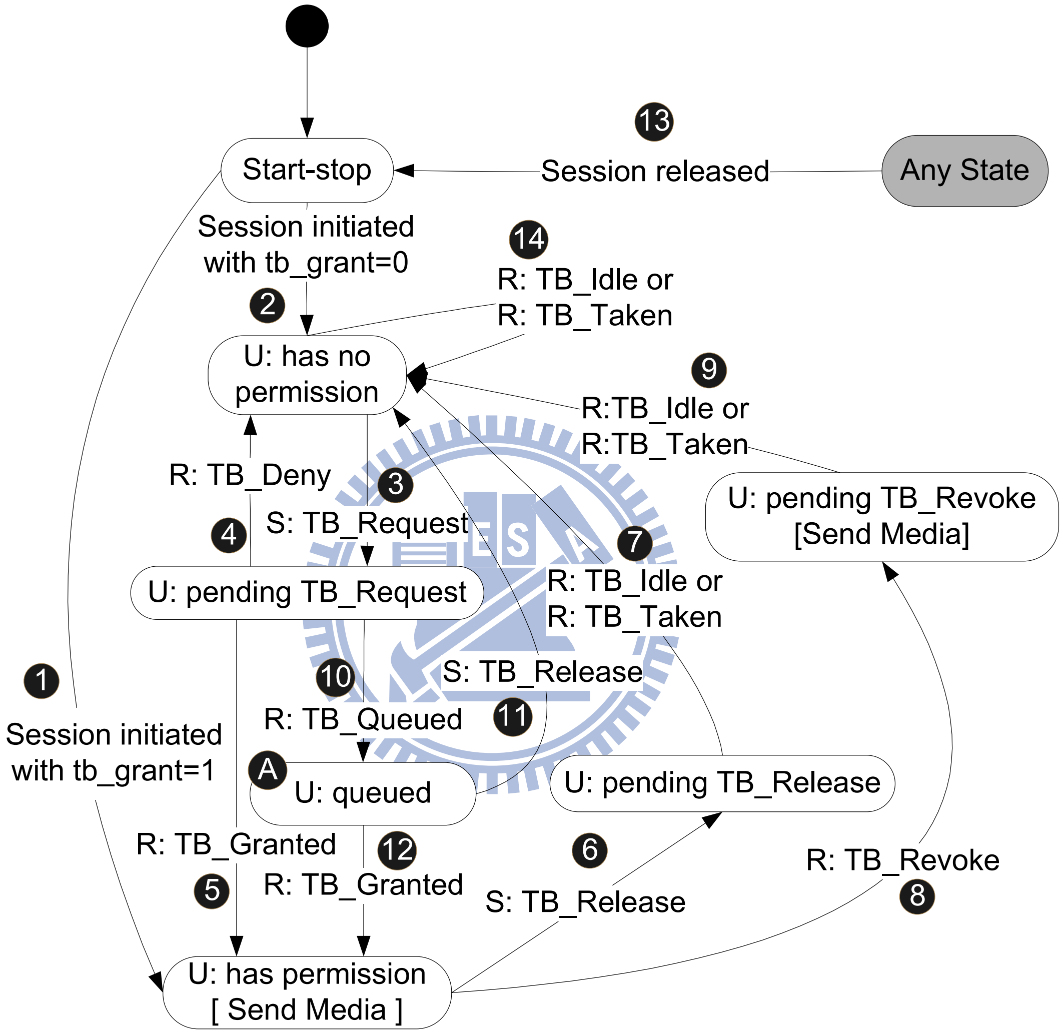


Figure 3.3: Talk Burst Control Finite State Machine for PoC Client (FSM_U)

state, the PoC client stops sending media packets and is waiting for the response from the PoC server. The sequence number of the last delivered media packet is included in the TB_Release message. FSM_G enters **G: pending TB_Release** after receiving the TB_Release message (transition 2 in Figure 3.2). In this state, the PoC server keeps forwarding the transient media packets delivered before the TB_Release message is issued from the permitted PoC client. When the last transient media packet has been processed, FSM_G enters **G: TB_Idle** (transition 3 in Figure 3.2). This state means that no PoC client is granted the permission to speak. The PoC server broadcasts the TB_Idle message to all PoC clients. FSM_U of the permitted PoC client enters **U: has no permission** upon receipt of the TB_Idle message (transition 7 in Figure 3.3). A listening PoC client remains in **U: has no permission** when it receives the TB_Idle message (transition 14 in Figure 3.3). At this point, all PoC clients can compete for the permission to speak.

To obtain the permission, a listening PoC client sends the TB_Request message to the PoC server. The PoC client becomes a requesting PoC client, where its FSM_U enters **U: pending TB_Request** (transition 3 in Figure 3.3). This state means that the PoC client is waiting for the arbitration from the PoC server. If some other PoC client has been granted the permission, the PoC server sends the TB_Deny message to the requesting PoC client, and FSM_U of the requesting PoC client moves back to **U: has no permission** (transition 4 in Figure 3.3). If the PoC server grants the permission to the requesting PoC client, it sends the TB_Granted message to this requesting PoC client and the TB_Taken message to other PoC clients. A timer T2 at the PoC server is started; this timer is used to determine whether the permitted PoC client speaks too long (and therefore should be revoked). FSM_G enters **G: TB_Taken** (transition 4 in Figure 3.2), and FSM_U of the requesting PoC client enters **U: has permission** (transition 5 in Figure 3.3). When a listening PoC client receives TB_Taken, its FSM_U remains at **U: has no permission**, and is not allowed to request for the permission. The requesting PoC client becomes the permitted PoC client.

If the permitted PoC client speaks longer than the T2 timeout period, the PoC server will start timer T3 and send the TB_Revoke message to stop the permitted PoC client. FSM_U of the permitted PoC client enters **U: pending TB_Revoke** upon receipt of the TB_Revoke message (transition 8 in Figure 3.3). In this state, the PoC client can keep sending media packets until T3 expires. FSM_G enters **G: pending TB_Revoke** (transition 5 in Figure 3.2). In this state, the PoC server keeps forwarding media packets until T3 expires. Then FSM_G enters **G: TB_Idle** (transition 6 in Figure 3.2). The PoC server sends the TB_Idle message to all PoC clients. FSM_U of the permitted PoC client enters **U: has no permission** (transition 9 in Figure 3.3) upon receipt of the TB_Idle message. When a listening PoC client receives TB_Idle, its FSM_U remains at **U: has no permission**. At this point, all PoC clients can compete for the permission to speak.

A queueing option is provided in the talk burst control mechanism. If this option is selected, then the ungranted requests are buffered in the queue at the PoC server instead of being denied. After the permitted PoC client finishes talking, the PoC server grants the next request from the queue. The state **U: queued** (see Figure 3.3 (A)) in FSM_U indicates that a request of the client is queued in the PoC server and will be granted later.

After a PoC client X has obtained the permission, the PoC server may receive the TB_Request message from another requesting PoC client Y. With the queueing option, FSM_G is in the **G: TB_Taken** state, and FSM_U of client Y is in the **U: pending TB_Request** state. The PoC server buffers the TB_Request message in the queue and replies the TB_Queued message to client Y. FSM_G stays in **G: TB_Taken** and FSM_U

of client Y enters **U: queued** (transition 10 in Figure 3.3). The PoC client of a queued request (called *queued PoC client*) may send the TB_Release message to the PoC server to cancel the request, and its FSM_U moves back to **U: has no permission** (transition 11 in Figure 3.3). In this case, the PoC server removes the corresponding request from the queue. After the permission is released (or revoked), FSM_G will enter **G: TB_Idle**, and FSM_U of the permitted PoC client will enter **U: pending TB_Release** (or **U: pending TB_Revoke**). If the queue is not empty, the PoC server processes the next queued request instead of sending out the TB_Idle message. Then the granting procedure is performed at the PoC server. Therefore, the next queued PoC client will receive the TB_Granted message from the PoC server, and its FSM_U enters **U: has permission** (transition 12 in Figure 3.3). This queued PoC client becomes the next permitted PoC client. At the same time, all other PoC clients receive the TB_Taken message from the PoC server. The previously-permitted PoC client becomes a listening PoC client, and its FSM_U enters **U: has no permission** (transition 7 or 9 in Figure 3.3). FSM_U of every listening PoC client remains at **U: has no permission**, and FSM_U of every queued PoC client remains at **U: queued**.

When a PoC client leaves the PoC session, its FSM_U moves back to **Start-stop** (transition 13 in Figure 3.3). The PoC session remains active for other PoC clients. After all PoC clients leave the PoC session, the PoC session is implicitly terminated. FSM_G moves back to **Start-stop** (transition 7 in Figure 3.2).

3.3 Analytic Modeling

This section models the PoC talk burst control mechanisms with queueing (Approach Q) and without queueing (Approach NQ). Let M be the number of PoC clients in a PoC group. We investigate two timers defined in FSM_G: (1) T2 is used to determine whether the PoC client speaks too long; (2) T3 is used to gracefully terminate the talk burst. For simplicity, we define T_R as the revoking timer where

$$T_R = T2 + T3$$

Three output measures are considered in our study:

- P_D: the probability that the request of a PoC client is not granted because the PoC client is not patient in Approach Q or because the PoC server rejects the request in Approach NQ
- P_R: the probability that a PoC client obtains the permission, but is revoked before it finishes the talk
- W (for Approach Q only): the expected waiting time of a request, i.e., the expected time between when a PoC client issues a request and when it is granted the permission to speak or when it leaves the queue without being granted the permission

For a PoC client, we define the following input parameters:

- The inter-request time random variable τ_a with mean $1/\lambda$
- The speak time random variable τ_s with mean $1/\mu$ and variance V_s

- The patient time random variable τ_p with mean $1/\omega$ and variance V_p (note that a queued PoC client will drop the request if it does not receive the permission within τ_p)
- The revoking time T_R (when $\tau_s > T_R$, the permitted PoC client is revoked); T_R is a fixed period

We assume that τ_a, τ_s and τ_p are exponentially distributed (i.e., $V_s = 1/\mu^2, V_p = 1/\omega^2$) in this section. The purpose of analytic model is two folds: to validate the simulation model and to provide mean value analysis. Mean value analysis based on exponential assumptions provides understanding on the "trend" of performance. The validated simulation relaxes the exponential assumptions to accommodate more general (and therefore more practical) scenarios such as those described in [25].

3.3.1 Modeling for Approach Q

The PoC talk burst control mechanism is modeled as a stochastic process. Figure 3.4 illustrates the state transition rate diagram of the stochastic process where the state $K = k$ denotes that $k-1$ requests are waiting in the queue at the PoC server besides the permitted PoC client. In this figure, we define $\gamma(t)h$ to be the probability that the permitted PoC client is revoked during the interval $(t, t+h)$, where $h \to 0$. When the process is in state k, where $0 < k \leq M$, there are $M-k$ listening PoC clients, $k-1$ queued PoC clients, and a permitted PoC client. State 0 represents that all PoC clients are listening PoC clients. For any state k, there are $M-k$ listening PoC clients that may request for permission to speak. Therefore, the transition rate from state k to state $k+1$ is $(M-k)\lambda$. For any state k, where $0 < k \leq M$, there are $k-1$ queued PoC clients that

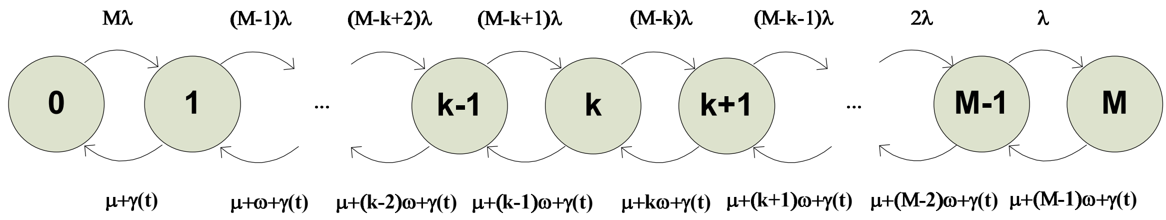


Figure 3.4: State Transition Rate Diagram for Approach Q

may leave the queue with rate $(k-1)\omega$, and the permitted PoC client that may finish the talk with rate μ. Furthermore, the permitted PoC client may be revoked during the interval $(t, t+h)$ with probability $\gamma(t)h$. Therefore, at time t the transition rate from state k to state $k-1$ is $\mu + (k-1)\omega + \gamma(t)$.

Let $\pi_k(t)$ denote the probability that the number of requests in the system is k at time t. Then based on the state transition rate diagram in Figure 3.4, we obtain the following differential difference equations [26]:

$$\pi_0'(t) = \pi_1(t)[\mu + \gamma(t)] - \pi_0(t)M\lambda \tag{3.1}$$

$$\begin{aligned} \pi_k'(t) &= \pi_{k-1}(t)(M-k+1)\lambda + \pi_{k+1}(t)[\mu + k\omega + \gamma(t)] \\ &\quad -\pi_k(t)[\mu + (k-1)\omega + \gamma(t) + (M-k)\lambda] \quad (1 \le k \le M-1) \end{aligned} \tag{3.2}$$

Under the assumption of statistical equilibrium (i.e., $t \to \infty$), Equations (3.1) and (3.2) are re-written as

$$0 = \pi_1(\mu + \gamma) - \pi_0 M\lambda \tag{3.3}$$

$$\begin{aligned} 0 &= \pi_{k-1}(M-k+1)\lambda + \pi_{k+1}(\mu + k\omega + \gamma) \\ &\quad -\pi_k[\mu + (k-1)\omega + \gamma + (M-k)\lambda] \quad (1 \le k \le M-1) \end{aligned} \tag{3.4}$$

where $\pi_i = \lim_{t\to\infty} \pi_i(t)$ and $\gamma = \lim_{t\to\infty} \gamma(t)$ is the revoking rate in equilibrium and will be derived later. Equations (3.3) and (3.4) are re-arranged to yield

$$\pi_1 = \frac{\pi_0 M\lambda}{\mu+\gamma} \tag{3.5}$$

$$\pi_{k+1} = \left[\frac{\mu+(k-1)\omega+\gamma+(M-k)\lambda}{\mu+k\omega+\gamma}\right]\pi_k - \left[\frac{(M-k+1)\lambda}{\mu+k\omega+\gamma}\right]\pi_{k-1} \tag{3.6}$$

From (3.5) and (3.6), we have

$$\pi_k = \left\{\frac{\lambda^k\left[\prod_{j=0}^{k-1}(M-j)\right]}{\prod_{m=1}^{k}[\mu+(m-1)\omega+\gamma]}\right\}\pi_0 \tag{3.7}$$

where $1 \le k \le M$. Since $\sum_{k=0}^{M}\pi_k = 1$, (3.7) is solved to yield

$$\pi_0 = \left\{1+\sum_{k=1}^{M}\frac{\lambda^k\left[\prod_{j=0}^{k-1}(M-j)\right]}{\prod_{m=1}^{k}[\mu+(m-1)\omega+\gamma]}\right\}^{-1} \tag{3.8}$$

The revoking rate γ is derived as follows. Let $p(\tau,t)h$ be the probability that the elapsed speak time (i.e., the member has talked for the period τ, and the talk has not been finished yet) is between $(\tau-h)$ and τ (where $h \le \tau \le T_R$, and $h \to 0$) at time t. The probability that the permitted PoC client finishes the talk in any interval h is μh. Since the elapsed speak time can advance from interval $(\tau-h,\tau)$ to interval $(\tau,\tau+h)$ only if the permitted PoC client does not finish the talk (with probability $1-\mu h$) during the interval $(t,t+h)$, we have

$$p(\tau+h,t+h)h = p(\tau,t)h(1-\mu h) \tag{3.9}$$

From (3.9), we have

$$\frac{p(\tau+h,t+h)-p(\tau,t)}{h} = -\mu p(\tau,t) \tag{3.10}$$

Let $h \to 0$, (3.10) is expressed as

$$\frac{\partial p(\tau,t)}{\partial \tau}+\frac{\partial p(\tau,t)}{\partial t} = -\mu p(\tau,t) \tag{3.11}$$

When $t \to \infty$, $\frac{\partial p(\tau,t)}{\partial t} = 0$, and (3.11) is re-written as

$$\frac{\partial p(\tau,\infty)}{\partial \tau} = -\mu p(\tau,\infty) \tag{3.12}$$

Equation (3.12) has the general solution

$$p(\tau,\infty) = Ae^{-\mu\tau} \tag{3.13}$$

From (3.13) and since $\int_0^{T_R} p(\tau,\infty)d\tau = 1$, we have

$$A = \frac{\mu}{1-e^{-\mu T_R}} \tag{3.14}$$

From (3.13) and (3.14), we have

$$p(\tau,\infty) = \frac{\mu e^{-\mu\tau}}{1-e^{-\mu T_R}} \tag{3.15}$$

For any interval $(t, t+h)$, a permitted PoC client is revoked if and only if the member has talked for the period τ at time t, where $T_R - h < \tau < T_R$. Therefore, for any interval h in equilibrium, the revoking probability γh is equal to $p(T_R,\infty)h$. In other words, the revoking rate $\gamma = p(T_R,\infty)$. From (3.7), (3.8) and (3.15), we have

$$\begin{aligned} \pi_k &= \left\{ \frac{\lambda^k \left[\prod_{j=0}^{k-1}(M-j)\right]}{\prod_{m=1}^{k}[\mu+(m-1)\omega+\frac{\mu e^{-\mu T_R}}{1-e^{-\mu T_R}}]} \right\} \\ &\quad \times \left\{ 1+\sum_{k=1}^{M} \frac{\lambda^k \left[\prod_{j=0}^{k-1}(M-j)\right]}{\prod_{m=1}^{k}[\mu+(m-1)\omega+\frac{\mu e^{-\mu T_R}}{1-e^{-\mu T_R}}]} \right\}^{-1} \end{aligned} \tag{3.16}$$

where $1 \le k \le M$.

The expected queue length is $E[K] = \sum_{k=0}^{M} k\pi_k$. From (3.16), we have

$$\begin{aligned} E[K] &= \left\{ \sum_{k=1}^{M} \frac{k\lambda^k \left[\prod_{j=0}^{k-1}(M-j)\right]}{\prod_{m=1}^{k}[\mu+(m-1)\omega+\frac{\mu e^{-\mu T_R}}{1-e^{-\mu T_R}}]} \right\} \\ &\quad \times \left\{ 1+\sum_{k=1}^{M} \frac{\lambda^k \left[\prod_{j=0}^{k-1}(M-j)\right]}{\prod_{m=1}^{k}[\mu+(m-1)\omega+\frac{\mu e^{-\mu T_R}}{1-e^{-\mu T_R}}]} \right\}^{-1} \end{aligned}$$

Let λ^* be the expected request rate to the PoC server. Then

$$\lambda^* = \sum_{k=0}^{M}(M-k)\lambda\pi_k = \lambda\left(M - E[K]\right) \tag{3.17}$$

Since a PoC client leaves the system without being granted the permission to speak if its waiting time in the queue is longer than its patient time, P_D can be expressed by

$$P_D = \frac{\text{the number of requests leaving the queue due to impatience}}{\text{the number of total request arrivals}} \tag{3.18}$$

From (3.17) and (3.18), we have

$$P_D = \frac{\sum_{k=1}^{M}(k-1)\omega\pi_k}{\lambda^*} = \frac{\omega\{E[K]-(1-\pi_0)\}}{\lambda\{M-E[K]\}}$$

Similarly, P_R can be expressed by

$$P_R = \frac{\text{the number of revoked requests}}{\text{the number of total request arrivals}} \tag{3.19}$$

From (3.17) and (3.19), we have

$$P_R = \frac{\sum_{k=1}^{M}\gamma\pi_k}{\lambda^*} = \frac{\mu e^{-\mu T_R}(1-\pi_0)}{\lambda\{M-E[K]\}(1-e^{-\mu T_R})}$$

Let random variable K_Q be the number of requests in the queue at the PoC server and random variable K_S be the number of permitted PoC clients, then $K = K_Q + K_S$ and

$$E[K] = E[K_Q] + E[K_S] \tag{3.20}$$

Note that $K_S = 1$ when system is busy (i.e., some PoC client is speaking), and $K_S = 0$ when system is idle (i.e., no PoC client is speaking). Therefore,

$$E[K_S] = 1 \times \left(\sum_{k=1}^{M}\pi_k\right) + 0 \times \pi_0 = 1 - \pi_0 \tag{3.21}$$

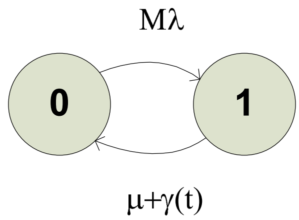


Figure 3.5: State Transition Rate Diagram for Approach NQ

From Little's result [26], $E[K_Q]$ is expressed as

$$E[K_Q] = \lambda^* W \tag{3.22}$$

From (3.17), (3.20), (3.21) and (3.22), we have

$$W = \frac{E[K_Q]}{\lambda^*} = \frac{E[K] - (1 - \pi_0)}{\lambda\{M - E[K]\}}$$

3.3.2 Modeling for Approach NQ

Figure 3.5 illustrates the state transition rate diagram for Approach NQ. State $K = 0$ represents that no PoC client has the permission, and state $K = 1$ represents that a PoC client is speaking. When the stochastic process is in state 0, there are M listening PoC clients that may request for the permission to speak. Therefore, the transition rate from state 0 to state 1 is $M\lambda$. When the process is in state 1, the permitted PoC client may finish the talk with rate μ, and may be revoked during the interval $(t, t + h)$ with probability $\gamma(t)h$. Therefore, the transition rate from state 1 to state 0 is $\mu + \gamma(t)$ at time t. Similar to the derivation for Approach Q, the following relation at equilibrium is derived:

$$\pi_1 = \frac{M\lambda\pi_0}{\mu + \frac{\mu e^{-\mu T_R}}{1-e^{-\mu T_R}}} = \frac{M\lambda(1 - e^{-\mu T_R})\pi_0}{\mu} \tag{3.23}$$

Since $\pi_0 + \pi_1 = 1$, (3.23) is solved to yield

$$\pi_0 = \frac{\mu}{\mu + M\lambda(1 - e^{-\mu T_R})} \quad \text{and} \quad \pi_1 = \frac{M\lambda(1 - e^{-\mu T_R})}{\mu + M\lambda(1 - e^{-\mu T_R})} \tag{3.24}$$

Since a requesting PoC client leaves the system as soon as it finds that the system is busy (i.e., some PoC client is speaking), P_D can be expressed by

$$P_D = \frac{\text{number of request arrivals when system is busy}}{\text{number of total request arrivals}} \tag{3.25}$$

From (3.17), (3.24) and (3.25), we have

$$P_D = \frac{(M-1)\lambda\pi_1}{\lambda^*} = \frac{(M-1)\lambda(1 - e^{-\mu T_R})}{\mu + (M-1)\lambda(1 - e^{-\mu T_R})}$$

Similar to the derivation for Approach Q, we derive P_R from (3.17) and (3.24) as

$$P_R = \frac{\gamma\pi_1}{\lambda^*} = \frac{\mu e^{-\mu T_R}}{\mu + (M-1)\lambda(1 - e^{-\mu T_R})}$$

The above analytic model is validated against the discrete event simulation experiments. The discrete event simulation model for Approach Q is described in Appendix B; the simulation model for Approach NQ is similar to that for Approach Q, and the details are omitted. As shown in Table 3.1 (where $V_s = 1/\mu^2$, $M = 10$, $\omega = \mu$, $V_p = 1/\omega^2$, and $T_R = 3/\mu$), the analytic analysis is consistent with the simulation results.

3.4 Numerical Examples

Based on the simulation experiments validated against the analytic model, this section investigates the performance of the talk burst control mechanism. Suppose that τ_a is exponentially distributed with mean $1/\lambda$ [27], τ_s has the Gamma distribution with mean $1/\mu$ and variance V_s, and τ_p has the Gamma distribution with mean $1/\omega$ and variance V_p.

Table 3.1: Comparison of Analytic and Simulation Models ($V_s = 1/\mu^2$, $M = 10$, $\omega = \mu$, $V_p = 1/\omega^2$, $T_R = 3/\mu$)

(a) Approach Q

λ	0.005μ	0.05μ
$P_D(Analytic)$	0.0205	0.1779
$P_D(Simulation)$	0.0201	0.1755
Error	1.81 %	1.32 %
$P_R(Analytic)$	0.0488	0.0409
$P_R(Simulation)$	0.04881	0.0410
Error	0.02 %	0.24 %
$W(Analytic)$	$0.0205/\mu$	$0.1779/\mu$
$W(Simulation)$	$0.0202/\mu$	$0.1748/\mu$
Error	1.61 %	1.76 %

(b) Approach NQ

λ	0.005μ	0.05μ
$P_D(Analytic)$	0.0410	0.2995
$P_D(Simulation)$	0.0416	0.3003
Error	1.46 %	0.27 %
$P_R(Analytic)$	0.0477	0.0349
$P_R(Simulation)$	0.04775	0.03484
Error	0.10 %	0.16 %

The Gamma distribution is selected because it has been shown that the distribution of any positive random variable can be approximated by a mixture of Gamma distributions (see Lemma 3.9 in [15]). Following the past experience [16], we can measure the PoC speak times and impatience times from the field and then generate the Gamma distribution from the measured data. In our study, the range for M is selected based on commercial operation. For example, ChungHwa Telecom (CHT) has limited M to 20 (that is, at most 20 members can be defined in a group before the PoC session is initiated). Also note that finite population with Poisson arrival is widely used in commercial PoC network planning by mobile operators (e.g., FarEasTone or FET [16]) and is followed in this chapter. The PoC members are allowed to join in and to leave the PoC session at any time during the session, which can be accommodated in the simulation model. The effects of the input parameters are investigated as follows.

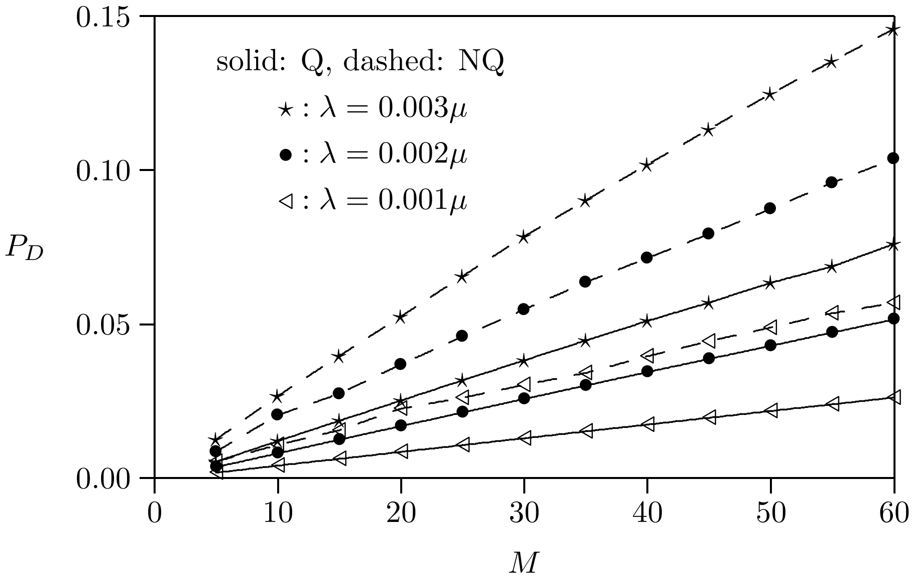


Figure 3.6: Effects of M and λ on P_D ($\omega = \mu$, $V_p = 1/\omega^2$, $V_s = 1/\mu^2$, $T_R = 3/\mu$)

Effects of M and λ on P_D : Under the conditions that $\omega = \mu$, $V_p = 1/\omega^2$, $V_s = 1/\mu^2$, $T_R = 3/\mu$, Figure 3.6 indicates that for the same P_D performance, Approach Q can support twice as many clients as Approach NQ does. For example, when $\lambda = 0.003\mu$, to maintain $P_D = 0.05$, $M = 38$ can be supported in Approach Q and $M = 19$ in Approach NQ (therefore, the maximum number M of group members should be set to 38 for Approach Q and 19 for Approach NQ, respectively). For both Q and NQ approaches, when λ increases, the number M supported in the PoC service decreases. For example, to maintain $P_D = 0.05$ in Approach NQ, $M = 51$ can be supported when $\lambda = 0.001\mu$, $M = 27$ when $\lambda = 0.002\mu$, and $M = 19$ when $\lambda = 0.003\mu$. We observe that the discrepancy of the P_D performance between Approach Q and Approach NQ decreases as ω increases. Approach Q is better than Approach NQ for all ω values (not shown in this dissertation). When $\omega = 0.5\mu$, Approach Q can support 6 times as many clients as Approach NQ does. When $\omega = 100\mu$, Approach

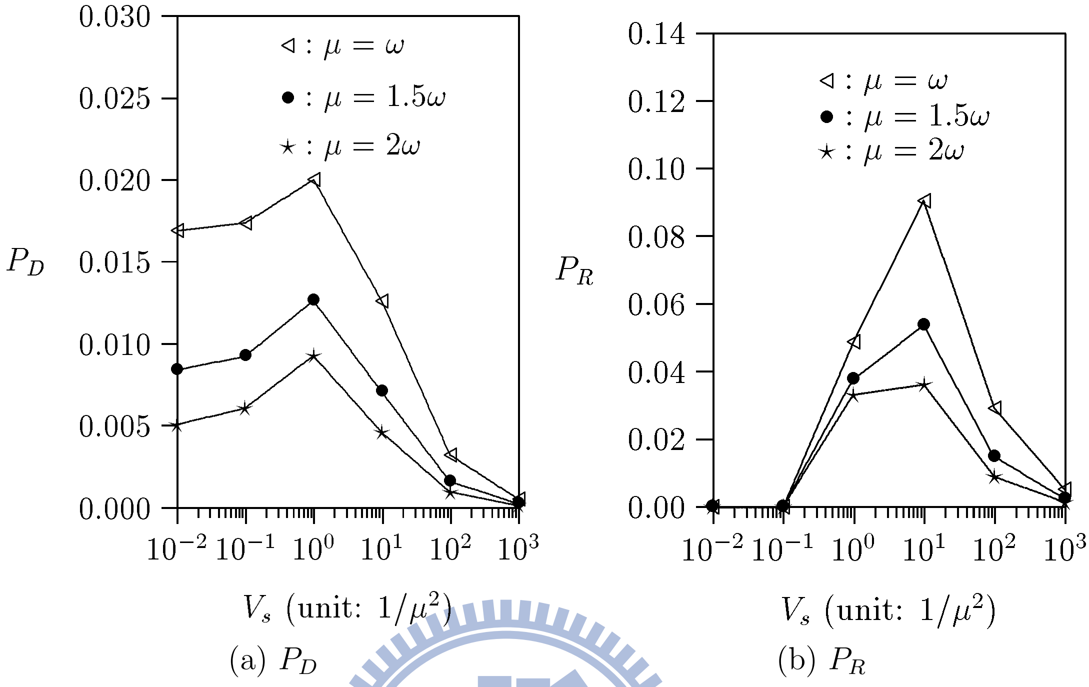
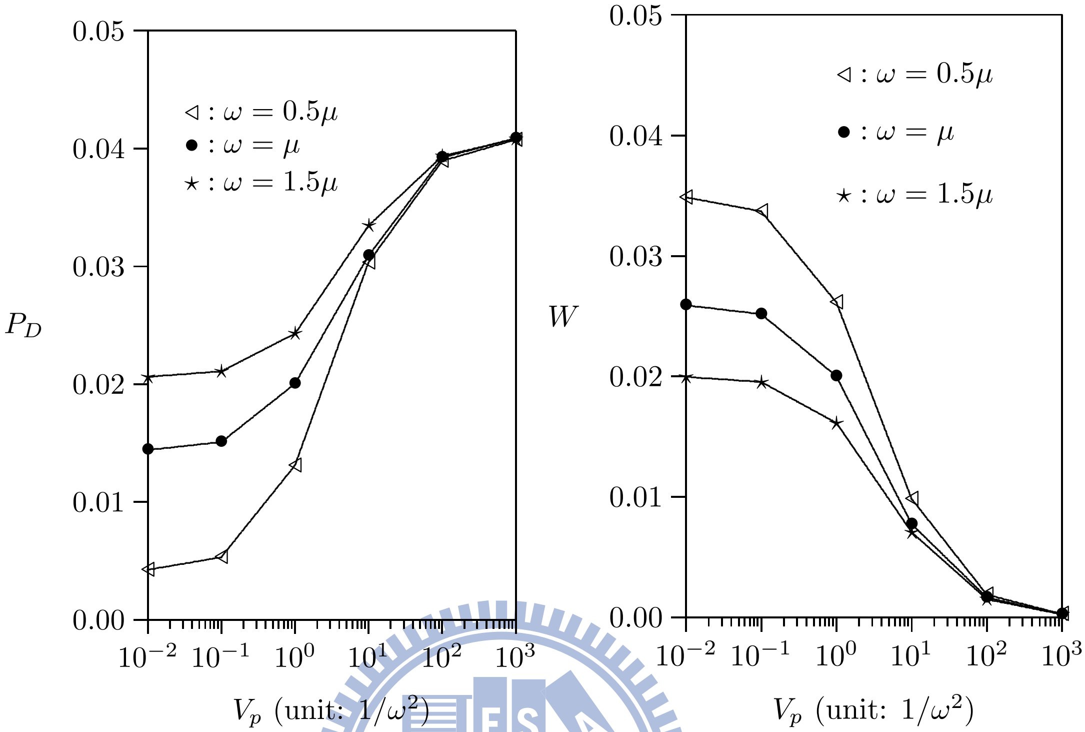


Figure 3.7: Effects of μ and V_s on P_D and P_R for Approach Q ($\lambda = 0.005\omega$, $M = 10$, $V_p = 1/\omega^2$, $T_R = 3/\mu$)

Q can only support 1.1 times as many clients as Approach NQ does. Note that $\omega < \mu$ in most cases (i.e., a queued PoC client is usually patient enough to wait for one permitted PoC client to finish the talk). Therefore, for the same P_D performance, Approach Q can support more than twice clients as Approach NQ does.

Effects of μ and V_s on P_D and P_R : In the remainder of this section, we only consider Approach Q. Similar results are also observed in Approach NQ. Figure 3.7 indicates that both P_D and P_R increase and then decrease as V_s increases (where $\lambda = 0.005\omega$, $M = 10$, $V_p = 1/\omega^2$, and $T_R = 3/\mu$). This phenomenon is explained as follows. When the variance V_s is small (e.g., $V_s \leq 1/100\mu^2$), all speak times are about the same length. As V_s increases, two effects are observed: (I) more speak times longer



(a) P_D (b) W

Figure 3.8: Effects of ω and V_p on P_D and W for Approach Q ($\lambda = 0.005\mu$, $M = 10$, $V_s = 1/\mu^2$, $T_R = 3/\mu$)

than T_R are observed. These speak times are revoked and result in larger P_D and P_R, (II) more short speak times are also observed, which result in smaller P_D and P_R. As V_s increases, Effect (I) is more significant when V_s is small, and Effect (II) is more significant when V_s is large. Therefore, both P_D and P_R increase and then decrease as V_s increases.

Effects of ω and V_p on P_D and W : Figure 3.8 shows that P_D is an increasing function of V_p, and W is a decreasing function of V_p (where $\lambda = 0.005\mu$, $M = 10$, $V_s = 1/\mu^2$, and $T_R = 3/\mu$). As V_p increases, more short and long patient times are observed.

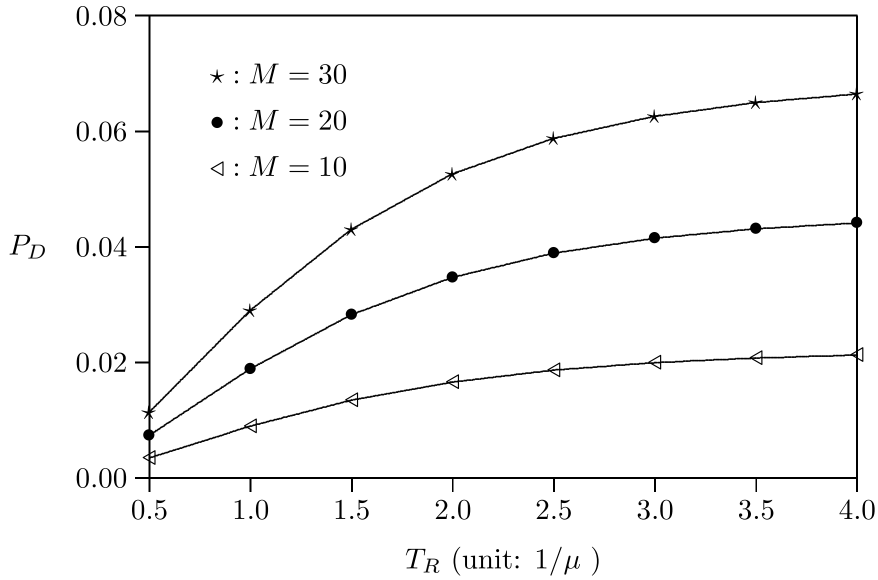


Figure 3.9: Effect of T_R on P_D for Approach Q ($\lambda = 0.005\mu$, $\omega = \mu$, $V_s = 1/\mu^2$, $V_p = 1/\omega^2$)

Short patient times result in larger P_D and shorter W, while long patient times result in smaller P_D and longer W. It is observed that the effect of short patient times is more significant than long patient times. Therefore, the net effect is that as V_p increases, P_D increases and W decreases. The figure also indicates that when V_p is very large, P_D is not sensitive to the ω values. That is, the effect of ω becomes less significant as V_p increases.

Effect of T_R on P_D : Figure 3.9 shows that P_D increases as T_R increases (where $\lambda = 0.005\mu$, $\omega = \mu$, $V_s = 1/\mu^2$, and $V_p = 1/\omega^2$). For a fixed P_D, if more clients are supported, a very short revoking time must be enforced to limit the speak times. For example, to maintain $P_D = 0.021$, the system can support $M = 10$ for $T_R = 4/\mu$ and $M = 30$ for $T_R = 0.75/\mu$.

3.5 Summary

In PoC service, the speak permission is arbitrated through the talk burst control mechanism. This chapter investigated the performance of the talk burst control mechanism for PoC service with queueing (Approach Q) and without queueing (Approach NQ). Our study indicates the following:

- For the same P_D performance, Approach Q can support more than twice clients as Approach NQ does.
- As the request rate λ increases, the maximum number M of group members can be supported in the PoC service decreases.
- Both the ungranting probability P_D and the revoking probability P_R increase and then decrease as the variance V_s of speak times increases.
- P_D is an increasing function of the variance V_p of the patient times. The expected waiting time W of a request is a decreasing function of V_p. The effect of the patient rate ω on P_D and W becomes less significant as V_p increases.
- For a fixed P_D, if more clients are supported, a very short revoking time must be enforced to limit the speak times.

Chapter 4

Bearer Reservation with Preemption for Voice Call Continuity

Voice Call Continuity (VCC) allows a user to switch from one domain to another during a call. A UMTS handset can initiate or receive a call in either the CS or the PS domain. During the call, the user may switch from one domain to another. In order to maintain call continuation, the connection in the old domain is released, and a connection is established in the new domain. For example, when a UE decides to transfer its VCC call from the PS domain to the CS domain, a new call is initiated in the CS domain, and the IP bearer in the PS domain is released.

After the call has been successfully switched to the CS domain, the UE may decide to switch the call back to the PS domain again. In this case, the released IP bearer must be re-established. Such bearer re-establishment contributes extra overload to the domain transfer. To resolve this issue, we propose the Bearer Reservation with Preemption (BRP) scheme, and present both analytic model and simulation experiments to investigate the BRP performance. In the BRP scheme, we do not release the IP bearer at the PS-to-CS domain transfer, and postpone the bearer release until the VCC call is complete. If the user moves back from the CS domain to the PS domain, the bearer re-establishment is eliminated. Our study indicates that when user behavior is irregular (i.e., either the variance of the domain residence times or the variance of the call holding times is large), the advantage of the BRP scheme becomes significant.

4.1 Introduction to VCC

In UMTS, the core network consists of two service domains: the *circuit-switched* (CS) and the *packet-switched* (PS) domains [1]. IMS is developed in the PS domain to provide Internet services. In existing commercial operation, the CS domain provides full service coverage with limited bandwidth. On the other hand, the PS domain typically provides zonal coverage with larger bandwidth and cheaper services. Therefore, when the PS connection is available, the UE will switch to the PS domain. When the PS connection is not available, the UE switches to the CS domain. A UMTS handset can attach to the CS and the PS domains individually or simultaneously, and initiate or receive a call in either domain. The user may switch from one domain to another during the call. In order to maintain call continuation, the connection in the old domain is released, and a connection is established in the new domain. This process is called *domain transfer*. The technique to transfer a voice call between the CS and the PS domains is called *Voice Call Continuity* (VCC) [12].

Figure 4.1 illustrates a simplified UMTS network architecture that accommodates VCC [12, 4]. This architecture consists of the UMTS Terrestrial Radio Access Network (UTRAN) (Figure 4.1 (a)), the Worldwide Interoperability for Microwave Access (WiMAX) or the Wireless Fidelity (WiFi) network (Figure 4.1 (b)), the CS domain (Fig-

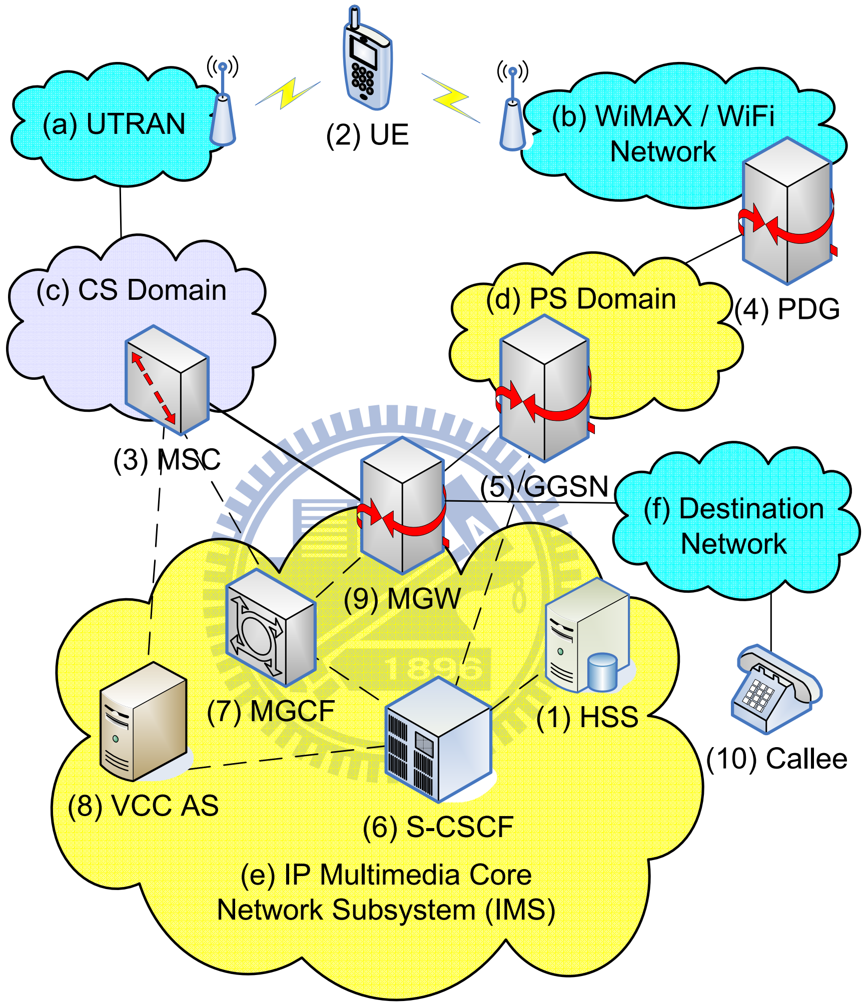


Figure 4.1: The UMTS Network Architecture for VCC (dashed lines: signaling; solid lines: signaling/data)

ure 4.1 (c)), the PS domain, i.e., the *General Packet Radio Service* (GPRS) network (Figure 4.1 (d)), and the IMS network (Figure 4.1 (e)). In this architecture, *Home Subscriber Server* (HSS; Figure 4.1 (1)) is the master database containing all user-related subscription information, which supports mobility management of the users. A mobile user utilizes a *User Equipment* (UE; Figure 4.1 (2)) to access CS and PS services. In the CS domain, the *Mobile Switching Center* (MSC; Figure 4.1 (3)) is responsible for call control, including the processing of user data and control signals. In the PS domain, the WiMAX/WiFi network connects to the GPRS network through the *Packet Data Gateways* (PDGs; Figure 4.1 (4)); the GPRS network connects to the IMS network through the *Gateway GPRS Support Nodes* (GGSNs; Figure 4.1 (5)). In the IMS, the transport of user data is separated from that for control signals. The IMS signaling is carried out by the *Serving Call Session Control Function* (S-CSCF; Figure 4.1 (6)), the *Media Gateway Control Function* (MGCF; Figure 4.1 (7)) and the *VCC Application Server* (VCC AS; Figure 4.1 (8)). When the UE switches from one domain to another during a voice call, the call is domain-transferred by the VCC AS. The IMS user data traffic is transported through the *Media Gateways* (MGWs; Figure 4.1 (9)) controlled by the MGCF. When the UE makes a call to a call party (Figure 4.1 (10)) in a different network (Figure 4.1 (f)), the call is first routed to the MGW. Then the MGW routes the call to the callee through the destination network.

4.2 VCC Call Setup and Domain Transfer

This section describes the VCC call setup and domain transfer procedures defined in 3GPP [28].

4.2.1 VCC Call Setup

To support domain transfer, the VCC AS is inserted into the signal path of the call. This is achieved by adding some VCC service triggering criteria (called initial filter criteria or iFC [29]) into the UE's profile in the HSS. When a UE registers to the IMS, the S-CSCF downloads these iFC of the UE from the HSS. When a call arrives at the S-CSCF, the call is evaluated against the iFC. If the VCC service criteria are matched, the call is routed to the VCC AS for further processing. The VCC call path is partitioned into two segments: the UE-MGW segment and the MGW-Callee segment. When the UE moves from one domain to another during a call, the UE-MGW segment is switched, and the MGW-Callee segment remains unchanged.

Suppose that a UE is attached to both the CS and the PS domains, and has performed the IMS registration. This UE can initiate or receive a call in either domain. Without loss of generality, we only describe the VCC call origination in the PS domain in this dissertation. The reader is referred to [28] for VCC call origination in the CS domain and VCC call termination in both domains. Figure 4.2 illustrates the message flow for VCC call origination in the PS domain with the following steps:

Step A.1 The UE sends the Session Initiation Protocol (SIP) [4, 3] INVITE message to the S-CSCF through the PS domain. This message contains the media information (e.g., IP address, port number and codec) for user data connection.

Step A.2 The S-CSCF evaluates the SIP INVITE message against the iFC of the UE. If the VCC service criteria are matched, the S-CSCF forwards the message to the VCC AS.

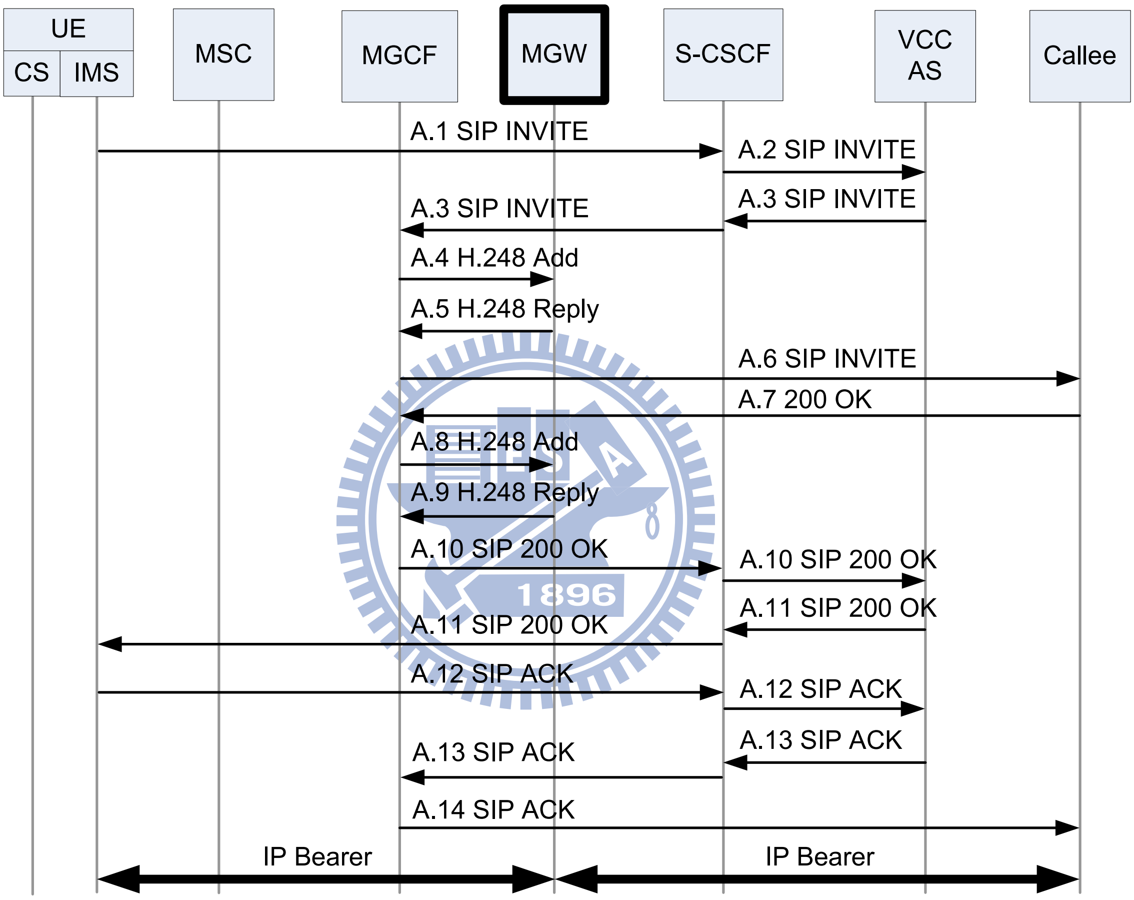


Figure 4.2: VCC Call Origination in the PS Domain

Step A.3 Based on the received SIP INVITE message, the VCC AS records the call information (e.g., From, To and Call-ID headers), and then forwards the SIP INVITE message to the MGCF through the S-CSCF.

Steps A.4 and A.5 Based on the media information retrieved from the SIP INVITE message, the MGCF exchanges the H.248 [30] Add and Reply messages with the MGW to allocate media resources for this call.

Steps A.6 and A.7 The MGCF modifies the media information contained in the SIP INVITE message and forwards the modified message to the callee. Then the callee replies a SIP 200 OK with its media information to the MGCF.

Steps A.8 and A.9 The MGCF retrieves media information from the SIP 200 OK message, and finalizes the MGW media resources for this call by exchanging the H.248 Add and Reply messages with the MGW.

Steps A.10 and A.11 The MGCF provides the final media information and forwards the SIP 200 OK message to the VCC AS. Then the VCC AS forwards this message to the UE. The UE retrieves media information from this message, and the call path in the UE-MGW segment is established.

Steps A.12-A.14 The UE sends a SIP ACK message to the callee through the S-CSCF, the VCC AS and the MGCF. After the callee has received the acknowledgment, the VCC call is established.

4.2.2 3GPP Domain Transfer

In the CS domain, VCC service control is provided through the *Customized Applications for Mobile Network Enhanced Logic* (CAMEL) [31], where the VCC service logic is implemented in the VCC AS. When a UE decides to transfer its VCC call from the PS domain to the CS domain, a new call is initiated in the CS domain with a specific called number *VCC Domain Transfer Number* (VDN). This number is then translated into a routable number *IP Multimedia Routing Number* (IMRN) through the VCC service logic. The IMRN is used to route the call from the CS domain to the VCC AS in the PS domain. When the call setup signal arrives at the VCC AS, the VCC AS updates the UE-MGW segment. The message flow for PS-to-CS domain transfer is illustrated in Figure 4.3 with the following steps:

Step B.1 Through the CS domain, the UE sends a Call Control (CC) SETUP message with the specific called VDN to the MSC.

Step B.2 The MSC sends a CAMEL Application Part (CAP) Initial Detection Point (IDP) message to the VCC AS. This message contains the calling number of the UE and the called VDN.

Step B.3 Based on the calling number in the CAP IDP message, the VCC AS identifies the ongoing call of the UE and allocates an IMRN for this call. Then the VCC AS replies a CAP CONNECT message with the IMRN to the MSC.

Step B.4 The MSC sends an ISDN User Part (ISUP) Initial Address Message (IAM) to the MGCF to set up the CS bearer. This message includes the IMRN received at Step B.3 as the called party number.

Steps B.5 and B.6 Upon receipt of the ISUP IAM message, the MGCF retrieves media information, and exchanges the H.248 Add and Reply messages with the MGW to allocate media resources for CS bearer between the UE and the MGW.

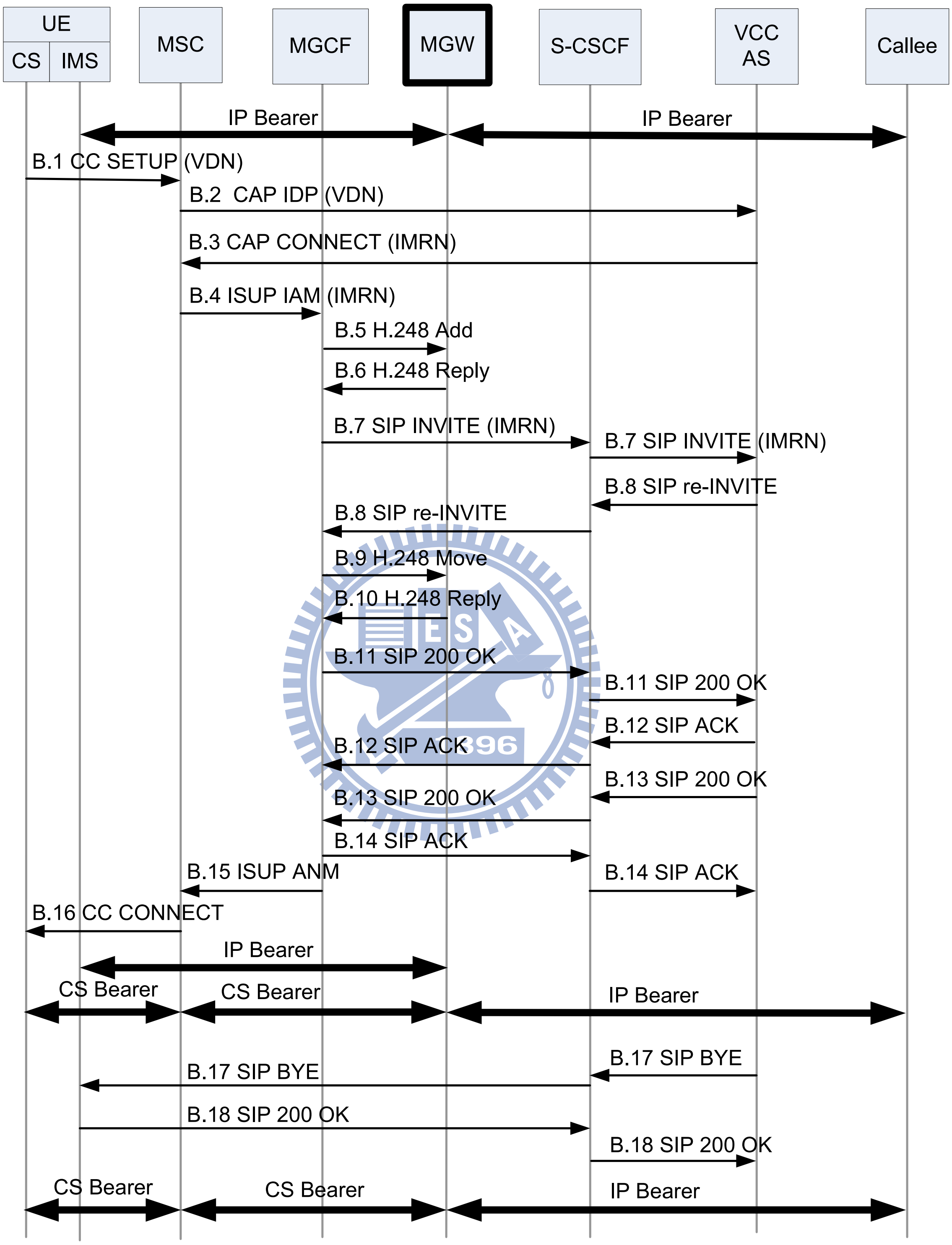


Figure 4.3: PS-to-CS Domain Transfer (3GPP TS 24.206)

Step B.7 The MGCF sends a SIP INVITE message with the called IMRN to the VCC AS through the S-CSCF.

Step B.8 Based on the calling party's identity in the received SIP INVITE message, the VCC AS retrieves the ongoing call information (i.e., the call information recorded at Step A.3) of the UE, and then sends a SIP re-INVITE message to the MGCF to modify the call path in the UE-MGW segment.

Steps B.9 and B.10 Upon receipt of the SIP re-INVITE message, the MGCF retrieves media information, and exchanges the H.248 Move and Reply messages with the MGW to switch the ongoing call in the PS domain to the new call in the CS domain.

Steps B.11 and B.12 The MGCF exchanges the SIP 200 OK and the SIP ACK messages with the VCC AS to indicate successful switching of the call path in the UE-MGW segment (corresponding to the re-INVITE message at Step B.8).

Steps B.13 and B.14 To complete the establishment of the CS bearer, the VCC AS exchanges the SIP 200 OK and the SIP ACK messages with the MGCF (corresponding to the INVITE message at Step B.7).

Steps B.15 and B.16 The MGCF sends an ISUP Answer Message (ANM) to the MSC. Then the MSC sends the CC CONNECT message to the UE. At this moment, the CS bearer for the UE-MGW segment is established.

Steps B.17 and B.18 When the SIP ACK message arrives, the VCC AS exchanges the SIP BYE and the SIP 200 OK messages with the UE to release the previously-established IP bearer in the UE-MGW segment.

After the call has been successfully switched to the CS domain, the UE may decide to switch the call back to the PS domain again. To trigger CS-to-PS domain transfer, the UE initiates a new call in the PS domain with a specific called identity *VCC Domain Transfer Uniform Resource Identifier* (VDI). When the new call arrives at the VCC AS, the VCC AS updates the UE-MGW segment for the UE. Figure 4.4 illustrates the message flow for CS-to-PS domain transfer with the following steps:

Step C.1 The UE sends a SIP INVITE message with the called VDI to the S-CSCF.

Step C.2 The S-CSCF evaluates the SIP INVITE message against the iFC of the UE. If the VCC service criteria are matched, the S-CSCF routes the call to the VCC AS.

Step C.3 Based on the calling party's identity in the received SIP INVITE message, the VCC AS retrieves the ongoing call information (i.e., the call information recorded at Step A.3) of the UE, and then sends a SIP re-INVITE message to the MGCF through the S-CSCF to switch the call path in the UE-MGW segment.

Steps C.4 and C.5 Upon receipt of the SIP re-INVITE message, the MGCF retrieves media information, and exchanges the H.248 Move and Reply messages with the MGW to switch the ongoing call in the CS domain to the new call in the PS domain.

Steps C.6 and C.7 The MGCF exchanges the SIP 200 OK and the SIP ACK messages with the VCC AS to indicate successful switching of the bearer in the UE-MGW segment (corresponding to the re-INVITE message at Step C.3).

Steps C.8 and C.9 To complete the IP bearer establishment, the VCC AS exchanges the SIP 200 OK and the SIP ACK messages with the UE (corresponding to the

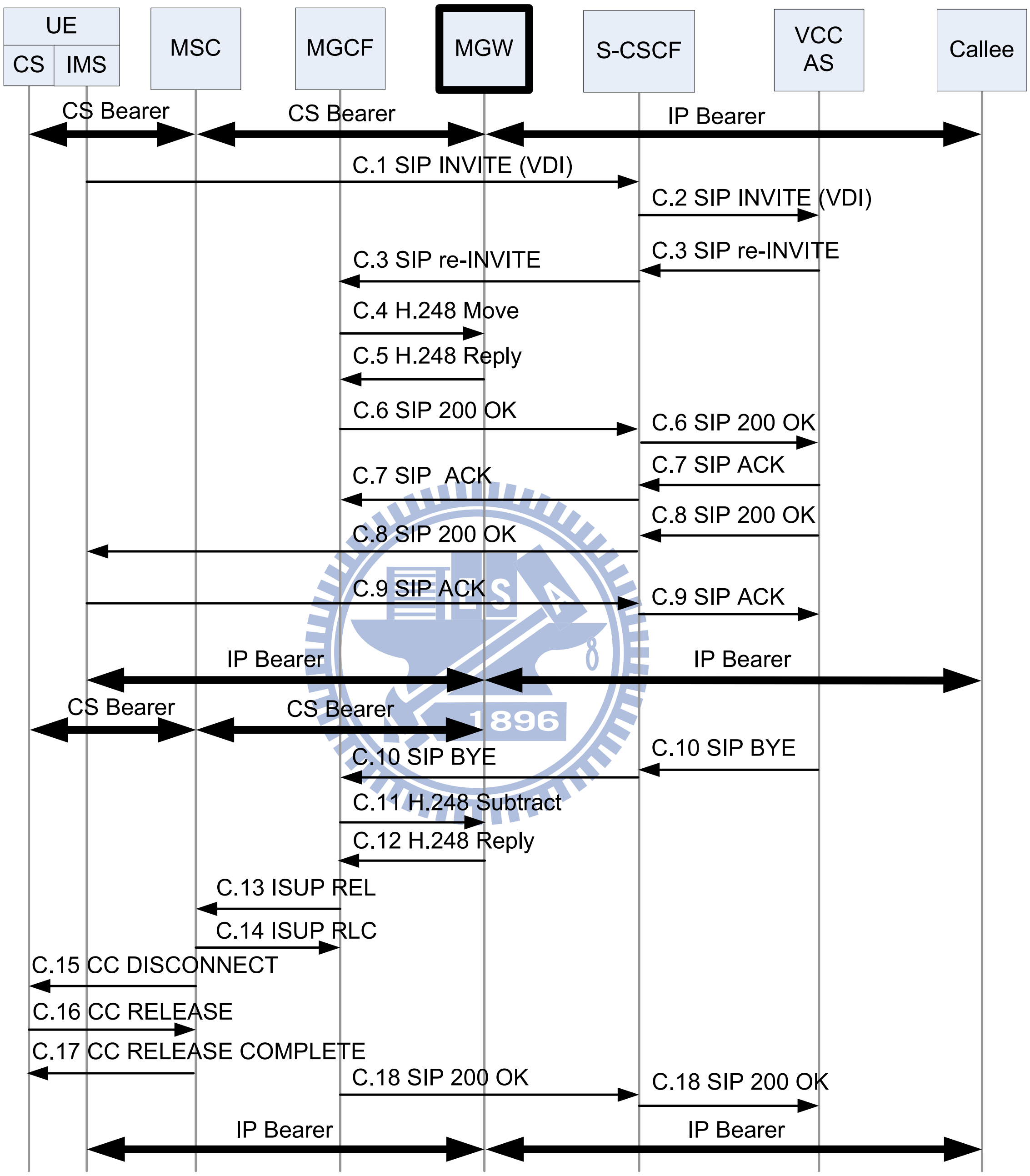


Figure 4.4: CS-to-PS Domain Transfer (3GPP TS 24.206)

INVITE message at Steps C.1 and C.2). At this point, the IP bearer for the UE-MGW segment is established.

Step C.10 To release the previously-established CS bearer, the VCC AS sends the SIP BYE message to the MGCF.

Steps C.11 and C.12 The MGCF exchanges the H.248 Subtract and Reply messages with the MGW to release the CS bearer between the MSC and the MGW.

Steps C.13 and C.14 To complete the CS bearer release between the MSC and the MGW, the MGCF exchanges the ISUP RELEASE (REL) and the RELEASE COMPLETE (RLC) messages with the MSC.

Steps C.15-C.17 The MSC exchanges the CC DISCONNECT, the CC RELEASE and the CC RELEASE COMPLETE messages with the UE to disconnect the CS bearer between the MSC and the UE.

Step C.18 Upon receipt of the ISUP RLC message at Step C.14, the MGCF sends a SIP 200 OK message to the VCC AS to indicate successful release of the CS bearer.

4.3 BRP Domain Transfer

In the 3GPP CS-to-PS domain transfer procedure, the CS bearer of the UE-MGW segment is released after the IP bearer is established. If the UE moves back to the CS domain again, the released CS bearer must be re-established. Such bearer re-establishment contributes extra overload to the domain transfer. To speed up the subsequent switchings, we may not release the CS bearer at the CS-to-PS domain transfer, and postpone the bearer release until the VCC call is complete. If the user moves back from the PS domain to the

CS domain, the bearer re-establishment is eliminated. Same argument applies to the IP bearer re-establishment.

Based on the above intuition, we propose the Bearer Reservation with Preemption (BRP) scheme that speeds up the domain transfer process. The BRP scheme utilizes *enhanced Multi-Level Precedence and Pre-emption* (eMLPP) service [32] and *Multimedia Priority Service* (MPS) [33] to provide reservation and preemption of CS and IP bearers. In BRP, two eMLPP priority levels are defined: the *high* priority and the *low* priority. When there is no available channel at the MSC, a call arrival with high priority can preempt a call with low priority, i.e., the high priority call is established, and the low priority call is force-terminated.

In BRP, a VCC call before domain transfer is set up with high priority. When the UE switches this call from the CS domain to the PS domain, instead of releasing the CS bearer in the UE-MGW segment, this CS bearer is reserved with low priority. When the UE switches the call back to the CS domain, the domain transfer process simply raises the priority level of the reserved CS bearer to high priority. If the reserved CS bearer with low priority is preempted (and the preempted channel is used by an incoming high-priority call), the CS bearer is released. In this case, the VCC call is not terminated because the IP bearer is used. When the call is switched back to the CS domain, the CS bearer needs to be re-established.

4.3.1 CS-to-PS Domain Transfer in the BRP Scheme

Figure 4.5 illustrates the BRP message flow for CS-to-PS domain transfer with IP bearer establishment with the following steps:

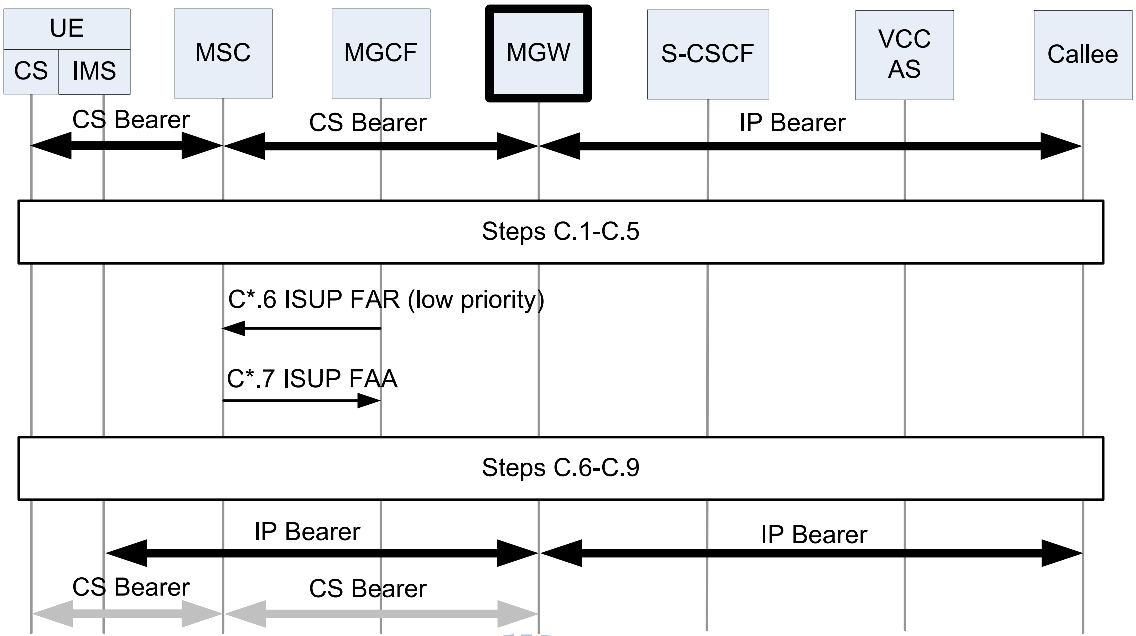


Figure 4.5: CS-to-PS Domain Transfer with IP Bearer Establishment (BRP)

Steps C.1-C.5 Same steps as in Figure 4.4 initiate the establishment of the IP bearer in the UE-MGW segment.

Step C*.6 The MGCF lowers the priority level for the CS bearer, and sends an ISDN User Part (ISUP) Facility Request (FAR) message with the parameter "low priority" to the MSC.

Step C*.7 According to the priority level indicated in the received ISUP FAR message, the MSC lowers the priority level for the CS bearer, and sends an ISUP Facility Accepted (FAA) message to the MGCF.

Steps C.6-C.9 Same steps as in Figure 4.4 exchange the Session Initiation Protocol (SIP) 200 OK and ACK messages to complete the IP bearer establishment in the UE-MGW segment.

By adding two messages (Steps C*.6 and C*.7), the BRP scheme eliminates eleven messages (Steps C.10-C.18) in Figure 4.4. Therefore, the message exchange cost is reduced by 36%. If the IP bearer has been reserved and not preempted before the domain transfer occurs (not shown in this dissertation), the message exchange cost is reduced by 68%. Also note that after the transfer, the CS radio link to the UE may be disconnected, but the CS bearer at the MSC is still maintained. This idea is similar to the "always on" concept of GPRS [4].

4.3.2 PS-to-CS Domain Transfer in the BRP Scheme

After the call has been successfully switched to the PS domain, the UE may decide to switch the call back to the CS domain again. If the reserved CS bearer has not been preempted, the UE does not need to initiate a new call for establishing the CS bearer in the UE-MGW segment. Instead, the UE only needs to raise the priority level of the reserved CS bearer to high priority. Also, unlike the procedure in Figure 4.3, the IP bearer is not released. Therefore, IP bearer needs not be re-established when the call switches back to the PS domain. Figure 4.6 illustrates the BRP message flow for PS-to-CS domain transfer without CS bearer establishment with the following steps:

Step B*.1 The UE sends a Call Management (CM) SERVICE REQUEST message to the MSC to raise the priority level of the CS bearer in the UE-MGW segment.

Step B*.2 The MSC raises the priority level for the CS bearer. Then the MSC sends an ISUP FAR message with the parameter "high priority" to the MGCF.

Steps B*.3 and B*.4 The MGCF raises the CS bearer's priority, and lowers the IP bearer's priority. Then the MGCF exchanges H.248 Move and Reply messages with

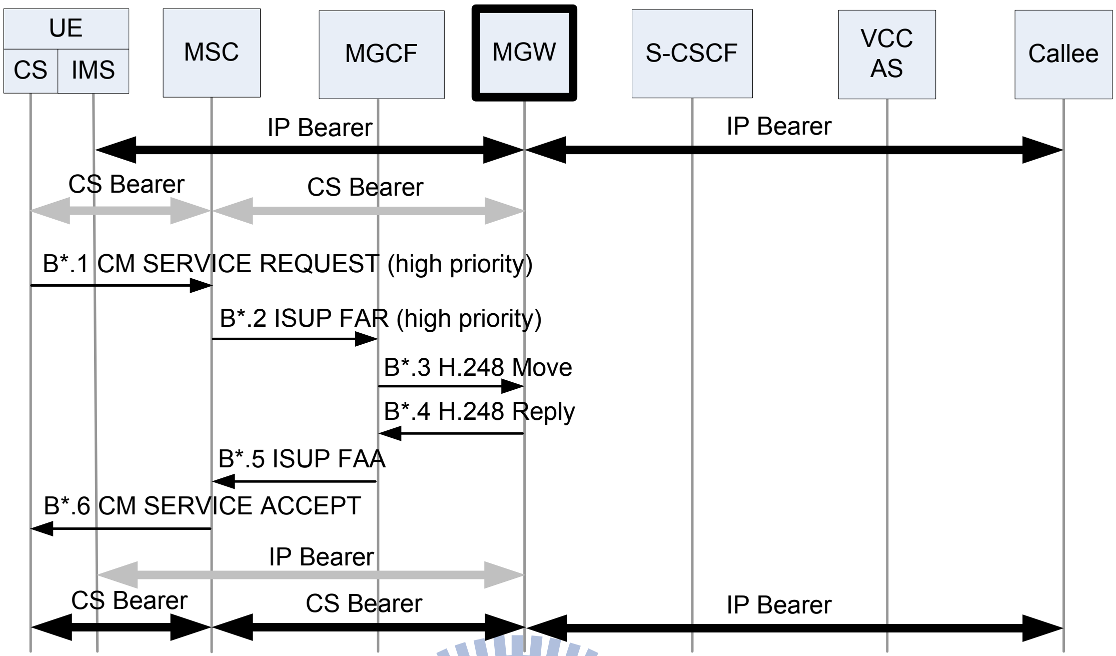


Figure 4.6: PS-to-CS Domain Transfer without CS Bearer Establishment (BRP)

the MGW to switch the UE-MGW segment from the PS bearer to the reserved CS bearer.

Steps B*.5 and B*.6 To complete this priority update, the MGCF sends an ISUP FAA message to the MSC. Then the MSC sends a CM SERVICE ACCEPT message to the UE to indicate successful priority update of the CS bearer. At this point, the UE-MGW segment is switched from the IP bearer to the CS bearer.

In the BRP scheme, six messages (Steps B*.1-B*.6) modify the priorities of the CS and the PS bearers. On the other hand, the 3GPP procedure in Figure 4.3 exchanges twenty-six messages (Steps B.1-B.18) to establish a new CS bearer, and release the old IP bearer. Therefore, the message exchange overhead is reduced by 77%. If the CS bearer has been preempted before the call is switched back to the CS domain (not shown in this dissertation), the message exchange cost is reduced by 15.4%.

4.4 Analytic Modeling of BRP

This section proposes an analytic model to study the performance of the BRP scheme. Without loss of generality, we investigate the BRP performance in the CS domain when the new calls arrive at the MSC are VCC calls (i.e., we do not consider non-VCC calls). Similar conclusions also apply to the PS domain, and the details are omitted. Suppose that a UE has switched its VCC call from the CS to the PS domain. In the BRP scheme, the CS bearer is reserved with low priority. When the UE switches from the PS domain back to the CS domain at time τ, there are three possibilities:

Case I. Before the UE switches back to the CS domain, the reserved CS bearer has been preempted, and there is no available resource (i.e., no channel in the MSC) at time τ. The call is force-terminated. Let p_f be the probability that this case occurs.

Case II. The reserved CS bearer has been preempted before τ, but the resource becomes available when the call is switched back to the CS domain at τ. The CS bearer is re-established at the PS-to-CS domain transfer. In this case, only 15.4% message overhead is saved by our approach. Let p_r be the probability of this case.

Case III. The reserved CS bearer is not preempted. The UE only needs to raise the priority level of the reserved CS bearer to high priority by executing the procedure in Figure 4.6. In this case, our approach saves 77% message overhead as compared with the 3GPP approach. The probability of this case is p_n.

It is clear that p_f is the same for both the 3GPP and the BRP schemes. Probabilities p_r and p_n are used to actually compute the overhead saved by our approach. We use θ_{P2C} (or θ_{C2P}) to represent the percentage of re-connection overhead saved by our approach as compared with the 3GPP approach for PS-to-CS (or CS-to-PS) domain transfer. Specifically,

$$\begin{aligned}
\theta_{P2C} &= 15.4\% \times \frac{p_r}{p_n + p_r} + 77\% \times \frac{p_n}{p_n + p_r}, \\
\theta_{C2P} &= 36\% \times \frac{p_r}{p_n + p_r} + 68\% \times \frac{p_n}{p_n + p_r}
\end{aligned} \tag{4.1}$$

For the illustration purpose, we only consider θ_{P2C} in this dissertation. Similar conclusions also apply to θ_{C2P}. The following input parameters are considered in this study:

- The arrivals of new VCC CS calls are a Poisson stream with rate λ.
- The call holding time is a random variable t_c with mean $1/\mu$ and variance V_c.
- A VCC call resides at a domain for a period t_d before it is switched to another domain. Let t_d be a random variable with mean $1/\delta$ and variance V_d.
- A high-priority (low-priority) call utilizes (reserves) a channel at the MSC for a sojourn time t_s before it is switched to another domain or is completed. It is clear that $t_s = \min(t_c, t_d)$. Let t_s be a random variable with mean $1/\eta$.

In the analytic model, we assume that t_c and t_d are exponentially distributed. Therefore, t_s is also exponentially distributed. The above exponential assumptions result in mean value analysis [25] (this exponential assumption will be relaxed in simulation experiments). We conduct the mean value analysis to provide understanding on the "trend" of performance. Furthermore, this exponential-based analytic model is used to validate the simulation model. Then the validated simulation model will relax the exponential assumptions to accommodate more general (and therefore more practical) scenarios.

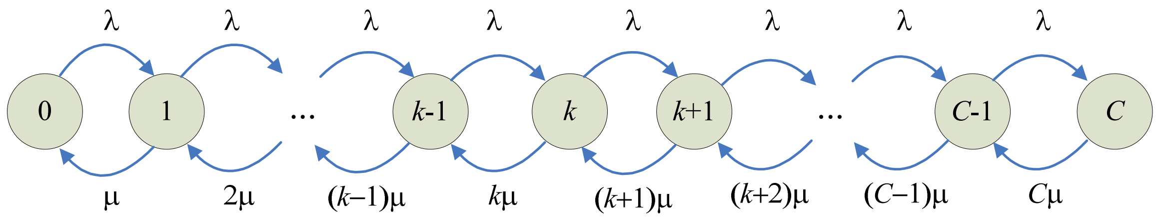


Figure 4.7: State Transition Rate Diagram for the BRP Scheme

The BRP scheme is modeled by a stochastic process. We first derive the number of channels occupied (either used or reserved) by the calls at the MSC. Let C be the capacity (number of channels) of the MSC. Figure 4.7 illustrates the state transition rate diagram of the stochastic process where state k denotes that there are k calls (either high-priority or low-priority) in the MSC. We note that during the call holding time t_c of a VCC call, the call may be switched between the CS and the PS domains, and the channel at the MSC is always occupied by the call. Therefore, the stochastic process can be modeled by a simple M/M/C/C queue with the parameters λ and μ. Let π_k denote the steady-state probability that there are k calls in the MSC. From the standard technique [34], we have

$$\pi_k = \pi_0 \left[\frac{\lambda^k}{(k!)\mu^k}\right], \pi_0 = \left[1 + \sum_{j=1}^{C} \frac{\lambda^j}{(j!)\mu^j}\right]^{-1} \text{ for } 0 \le k \le C \tag{4.2}$$

After a VCC call L switches from the CS to the PS domain, it becomes a low-priority call at the MSC. Figure 4.8 illustrates the timing diagram during L's sojourn time t_s, where L arrives at the MSC at τ_0 (i.e., it transfers from the high to the low priority at τ_0), and leaves the MSC at τ_7 (i.e., it completes or transfers back with high priority). There are two high-priority call arrivals at the MSC at τ_2 and τ_5, and there are four high-priority call departures at τ_1, τ_3, τ_4 and τ_6. When $k = C$, a high-priority call arrival will preempt an existing low-priority call. The order of preemption is based on the *Last-Come-First-Preempted* scheme (i.e., the last call arrival will be preempted first) [35]. Let $\bar{p}_n = 1 - p_n$

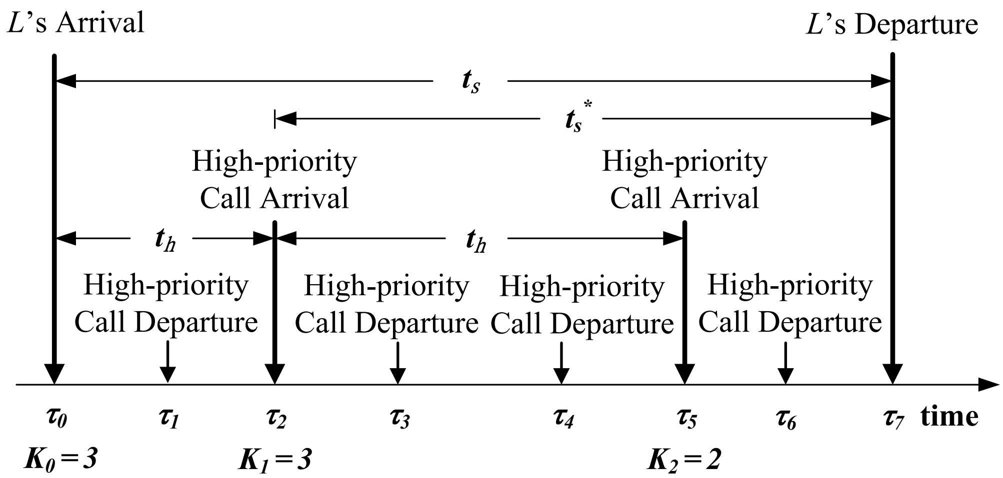


Figure 4.8: Events That May Occur in L's Sojourn Time

be the probability that a low-priority call L is preempted during its sojourn time. Let K_0 be the number of calls in the MSC seen by L at domain transfer (i.e., at τ_0), where L is not included in K_0. Note that from L's viewpoint, these K_0 calls are "high-priority" (i.e., none of them will be preempted before L is preempted). Since t_d is exponentially distributed, $\Pr[K_0 = m]$ can be derived based on the "flow rate" concept [26]. Under this concept, $(m+1)\delta\pi_{m+1}$ represents the number of calls that leave the MSC through domain transfers in a time unit when the system is at state $m+1$ (where $\delta = 1/E[t_d]$ is the domain transfer rate for a call; see page 59 in [26]), and $\sum_{j=0}^{C-1}(j+1)\delta\pi_{j+1}$ represents the total number of domain-transferred calls that leave the MSC in a time unit. From (4.2) and the "flow rate" concept, $\Pr[K_0 = m]$ is derived as

$$\Pr[K_0 = m] = \frac{(m+1)\delta\pi_{m+1}}{\sum_{j=0}^{C-1}(j+1)\delta\pi_{j+1}} = \left[\frac{\lambda^{m+1}}{(m!)\mu^{m+1}}\right]\left[\sum_{j=0}^{C-1}\frac{\lambda^{j+1}}{(j!)\mu^{j+1}}\right]^{-1} \quad (4.3)$$

After τ_0, L can only be preempted by a high-priority call arrival or by a PS-to-CS domain transfer. For simplicity, we do not consider PS-to-CS domain transfers, and we simply observe the moments when a high-priority call arrives. After τ_0, for $i \geq 1$, let K_i

be the number of high-priority calls in the MSC (from L's viewpoint, the low-priority calls counted in K_0 are also included in these "high-priority" calls) when the i-th high-priority call arrives, where this high-priority call is included in K_i. In Figure 4.8, if $K_0 = 3$ at τ_0, then $K_1 = 3$ at τ_2 (because there is one high-priority call departure in $[\tau_0, \tau_2]$), and $K_2 = 2$ at τ_5 (because there are two high-priority call departures in $[\tau_2, \tau_5]$).

For the i-th high-priority call arrival ($i \geq 0$; by convention, L represents the 0-th call arrival), let $p_{(m,n)}$ be the one-step transition probability from state $K_i = m$ to state $K_{i+1} = n$. That is, $p_{(m,n)}$ is the probability that there are $m - n + 1$ high-priority call departures during the inter-arrival time t_h between the i-th and the $(i+1)$-th high-priority call arrivals. Therefore, $p_{(C-1,C)}$ is the probability that when $K_i = C-1$, L will be preempted by the $(i+1)$-th high-priority call arrival. Note that for $0 \leq n \leq C$, $p_{(C,n)} = 0$ because L has already been preempted by the i-th high-priority call arrival. In addition, for $0 \leq m \leq C$, $p_{(m,0)} = 0$ because the $(i+1)$-th high-priority call arrival is included in K_{i+1}, and K_{i+1} is always larger than 0. Also, $p_{(m,n)} = 0$ if $n > m+1$ (because the $(i+1)$-th high-priority call is the only new call that contributes to K_{i+1}). In Figure 4.8, let t_s^* be the excess life (residual life) of t_s upon a high-priority call arrival, which has the density function $f^*(t_s^*)$ and the distribution function $F^*(t_s^*)$. Since t_s is exponentially distributed, t_s^* has the same distribution as t_s due to the memoryless property. Therefore, when $m \neq C$, $n \neq 0$ and $n \leq m+1$, $p_{(m,n)}$ can be derived by considering the relationship between t_s^* (for L and the m existing high-priority calls) and t_h (the inter-arrival times of two high-priority calls):

$$\begin{aligned}
p_{(m,n)} &= \int_{t_s^*=0}^{\infty}\int_{t_h=0}^{t_s^*}\binom{m}{n-1}F^*(t_h)^{m-n+1}\left[1-F^*(t_h)\right]^{n-1}\lambda e^{-\lambda t_h}f^*(t_s^*)dt_h dt_s^* \qquad (4.4)\\
&= \int_{t_s^*=0}^{\infty}\int_{t_h=0}^{t_s^*}\binom{m}{n-1}\sum_{j=0}^{m-n+1}\binom{m-n+1}{j}(-1)^j\lambda e^{[-(n+j-1)\eta-\lambda]t_h}\eta e^{-\eta t_s^*}dt_h dt_s^*\\
&= \sum_{j=0}^{m-n+1}\binom{m}{n-1,j}\left[\frac{(-1)^j\lambda}{\lambda+(n+j)\eta}\right] \qquad (4.5)
\end{aligned}$$

Equation (4.4) says that if $K_i = m$ and $K_{i+1} = n$, then among these m calls, the residual sojourn times of $n-1$ calls are larger than t_h (and therefore remain in the MSC at the end of t_h). The other $m-n+1$ calls have shorter residual sojourn times than t_h (and leave the MSC before the end of t_h). For $l \geq 2$, let

$$p_{(m,n)}^{(l)} = \sum_{j=0}^{C} p_{(m,j)}^{(l-1)} p_{(j,n)} \qquad (4.6)$$

In (4.6), $p_{(m,n)}^{(l)}$ is the probability that the stochastic process moves from state m to state n with exact l steps (i.e., there are l subsequent high-priority call arrivals). By convention, $p_{(m,n)}^{(1)} = p_{(m,n)}$. Then for $i \geq 1$, $\Pr[K_i = n]$ is expressed as

$$\Pr[K_i = n] = \sum_{m=0}^{C-1}\Pr[K_0 = m]p_{(m,n)}^{(i)} \qquad (4.7)$$

For $i \geq 2$, (4.7) can be recursively computed by using (4.6), and we have

$$\Pr[K_i = n] = \sum_{m=0}^{C-1}\Pr[K_0 = m]\left[\sum_{j=0}^{C} p_{(m,j)}^{(i-1)} p_{(j,n)}\right]$$

From (4.3), (4.5) and (4.7), the preemption probability $\bar{p}_n$ is derived as

$$\bar{p}_n = \sum_{i=0}^{\infty}\Pr[K_i = C-1]p_{(C-1,C)} \qquad (4.8)$$

Note that we typically do not see infinite high-priority call arrivals during L's sojourn time. From (4.7), it is clear that $\lim_{i\to\infty}\Pr[K_i = C-1] = 0$. Therefore, it suffices to consider $i \leq 50$ in (4.8). In this analytic model, p_f can be analytically derived using the technique in [36], and p_r is then computed as $p_r = \bar{p}_n - p_f$.

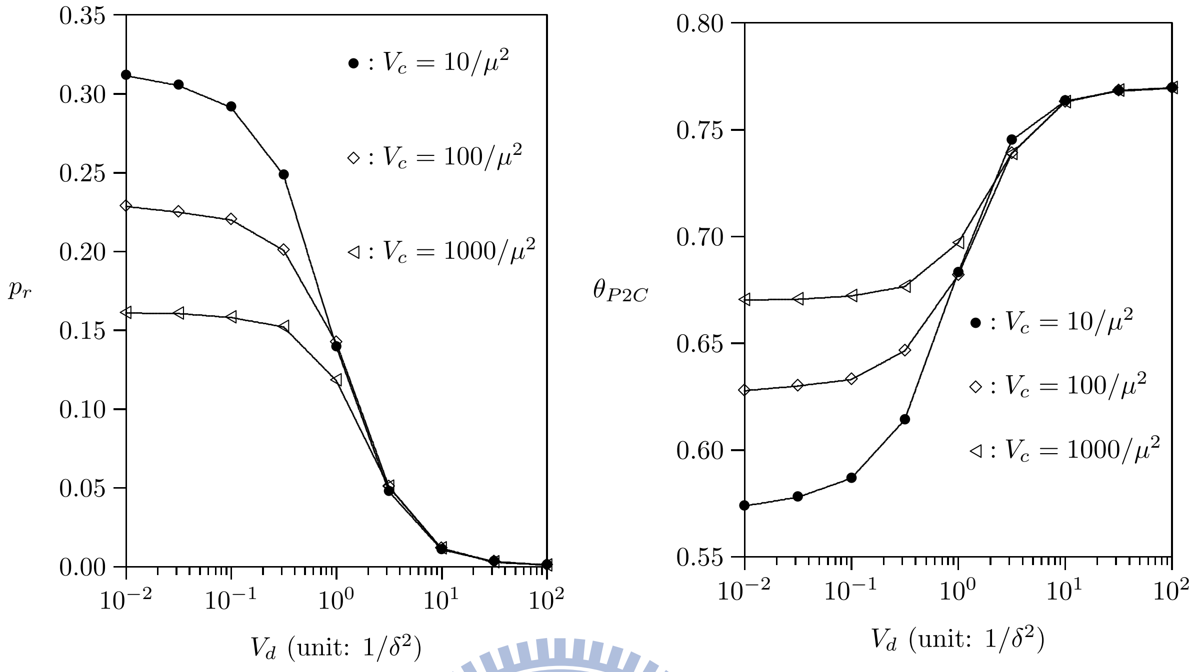
The above analytic model is used to validate against the discrete event simulation experiments described in Appendix C. We check the consistency of the analytic and the simulation models for three measures: π_k, $\Pr[K_0 = m]$ and $\bar{p}_n$. As shown in Table 4.1, the analytic results are consistent with the simulation results in terms of π_k (the errors are within 0.4%) and $\Pr[K_0 = m]$ (the errors are within 0.6%). On the other hand, inaccuracies are observed in $\bar{p}_n$. As δ increases, the inaccuracies between the analytic results and the simulation results become larger. These inaccuracies are caused by the simplifying assumption that L can only be preempted by a high-priority call arrival during its sojourn time. When δ is small (i.e., the domain transfer rate is low), L is more likely preempted by a high-priority call arrival than by a PS-to-CS domain transfer, and therefore the inaccuracies are insignificant. When δ is large (i.e., the domain transfer rate is high), L is more likely preempted by a PS-to-CS domain transfer, which results in larger inaccuracy. For example, when $\delta = 0.1\mu$, the errors are within 1.7%, while the errors are within 5% when $\delta = 0.2\mu$.

4.5 Numerical Examples

Based on the simulation experiments, this section investigates the performance of the BRP scheme. Suppose that t_c has Lognormal distribution with mean $1/\mu$ and variance V_c. The Lognormal distribution is selected because it has been shown that the call holding time distribution can be accurately approximated by a mix of two or more Lognormal distributions [36]. Similarly, we assume that t_d has the Gamma distribution with mean

Table 4.1: Comparison of Analytic and Simulation Models ($C = 5$)

π_k ($\lambda = \mu$)	π_0	π_1	π_2	π_3	π_4	π_5
Analytic	0.3681	0.3681	0.1840	0.0613	0.0153	0.0031
Simulation	0.3679	0.3682	0.18401	0.0615	0.01528	0.00309
Error	0.06 %	0.02 %	0.01 %	0.37 %	0.1 %	0.35 %
$\Pr[K_0 = m]$ ($\lambda = \mu = 5\delta$)	$\Pr[K_0 = 0]$	$\Pr[K_0 = 1]$	$\Pr[K_0 = 2]$	$\Pr[K_0 = 3]$	$\Pr[K_0 = 4]$	$\Pr[K_0 = 5]$
Analytic	0.3692	0.3692	0.1846	0.0615	0.0154	0
Simulation	0.3688	0.3686	0.1857	0.06135	0.01541	0
Error	0.1 %	0.14 %	0.6 %	0.24 %	0.07 %	0 %
$\bar{p}_n$ ($\mu = 10\delta$)	$\lambda = 0.5\mu$	$\lambda = \mu$	$\lambda = 1.5\mu$	$\lambda = 2\mu$	$\lambda = 2.5\mu$	$\lambda = 3\mu$
Analytic	0.000553	0.0099	0.0417	0.0989	0.1735	0.2548
Simulation	0.000548	0.00998	0.0423	0.1005	0.1764	0.2587
Error	0.95 %	0.8 %	1.56 %	1.6 %	1.7 %	1.53 %
$\bar{p}_n$ ($\mu = 5\delta$)	$\lambda = 0.5\mu$	$\lambda = \mu$	$\lambda = 1.5\mu$	$\lambda = 2\mu$	$\lambda = 2.5\mu$	$\lambda = 3\mu$
Analytic	0.000445	0.0081	0.0349	0.0846	0.1515	0.2265
Simulation	0.000458	0.0085	0.0335	0.0888	0.159	0.2362
Error	2.95 %	4.95 %	3.93 %	4.94 %	4.96 %	4.27 %



(a) p_r (b) θ_{P2C}

Figure 4.9: Effects of V_c and V_d on p_r and θ_{P2C} ($C = 5, \delta = \mu/5, \lambda = 2\mu$)

$1/\delta$ and variance V_d. The Gamma distribution is considered because the distribution of any positive random variable can be approximated by a mixture of Gamma distributions (see Lemma 3.9 in [15]), and is often used to represent the location residence times (inter-moving times) [25, 37, 38]. We can measure VCC call holding times and domain residence times from the commercial operation and then generate the Lognormal and Gamma distributions from the measured data. Experience from commercial operation shows that $\delta = \mu/10 \sim 10\mu$ is reasonable. In our numerical examples, we set $\delta = \mu/5$. The results for other δ values are similar, and are not presented. The effects of the input parameters are investigated as follows:

Effects of V_c and V_d on p_r and θ_{P2C} : Figure 4.9 (a) plots p_r against V_c and V_d, which

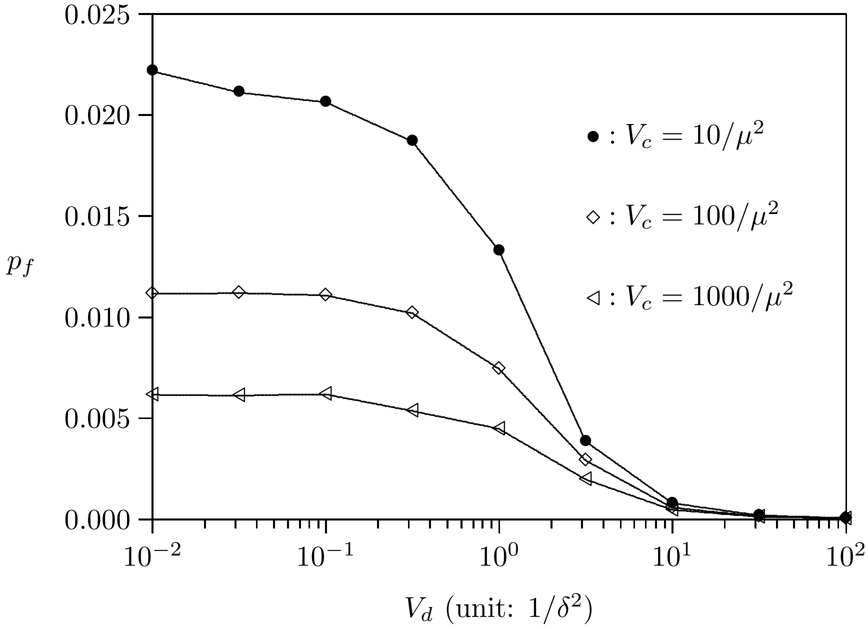
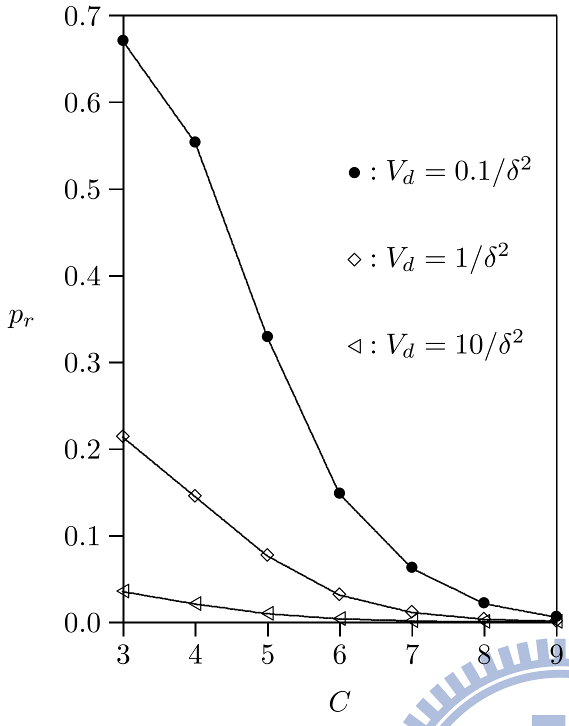


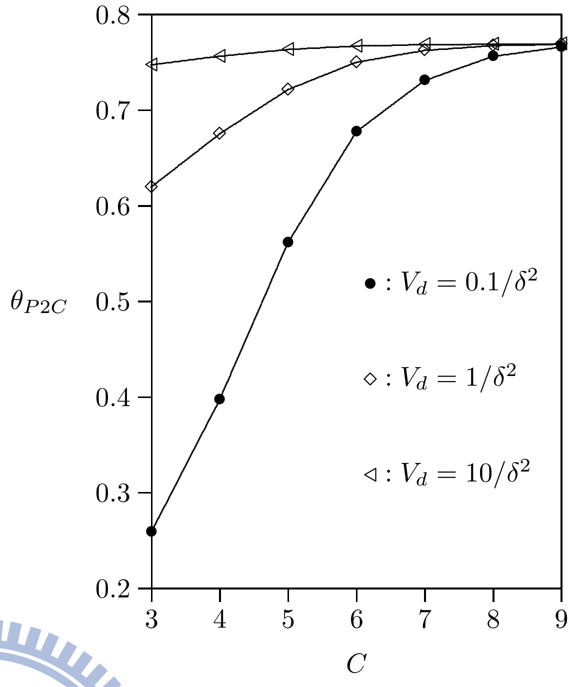
Figure 4.10: Effects of V_c and V_d on p_f $(C = 5, \delta = \mu/5, \lambda = 2\mu)$

indicates that p_r decreases as V_d increases. This phenomenon is explained as follows. When the domain residence times become more irregular (i.e., V_d increases), more short domain residence times are observed. Since t_s = min (t_c, t_d), more short sojourn times t_s are also observed. For a CS-to-PS domain transfer, the reserved CS bearer is less likely to be preempted if the call is more quickly switched back to the CS domain (i.e., t_d is shorter). Therefore, p_r decreases as V_d increases. For the same reason, p_r decreases as V_c increases. The figure also indicates that when $V_d > 10/\delta^2$, p_r is not significantly affected by V_c. Since θ_{P2C} is a decreasing function of p_r in Equation (4.1), θ_{P2C} is an increasing function of V_d and V_c as illustrated in Figure 4.9 (b).

Effects of V_c and V_d on p_f : Figure 4.10 plots p_f against V_c and V_d. This figure shows that p_f decreases as V_c or V_d increases. This phenomenon is similar to that of V_d



(a) p_r



(b) θ_{P2C}

Figure 4.11: Effect of C on p_r and θ_{P2C} ($\delta = \mu/5, \lambda = 2\mu, V_c = 1/\mu^2$)

and V_c on p_r, and is consistent with that observed in [37]. When $V_d > 30/\delta^2$, p_f is small and is not sensitive to the change of V_c.

Effect of C on p_r and θ_{P2C} : Figure 4.11 plots p_r and θ_{P2C} against V_d and C. The figure illustrates the trivial result that p_r is a decreasing function of C, and θ_{P2C} is an increasing function of C. The non-trivial result is that we quantitatively show that when $C < 7$, adding more channels at MSC significantly reduces p_r (and therefore significantly increases θ_{P2C}). When $C \geq 7$, p_r is sufficiently small (θ_{P2C} is sufficiently large), and increasing C simply wastes the resources. The figure also shows that the user behavior (i.e., V_d) significantly affects the resource allocated at the MSC (i.e., C) to achieve the same p_r and θ_{P2C} performances. For example, if the mobile operator wants to limit p_r to 15% (which ensures that $\theta_{P2C} \geq 67\%$) under the condition $\delta = \mu/5, \lambda = 2\mu$ and $V_c = 1/\mu^2$, then only 4 channels are required at the MSC when $V_d = 1/\delta^2$, while 6 channels should be supported when $V_d = 0.1/\delta^2$ (when user behavior is regular). In addition, when $V_d > 10/\delta^2$ (user behavior becomes more irregular), p_r is sufficiently small (and θ_{P2C} is sufficiently large), and there is no need to add extra resources (i.e., to increase C) at the MSC. Note that V_c has the same effect on p_r and θ_{P2C} as V_d does, and the details are omitted.

4.6 Summary

This chapter investigated Voice Call Continuity (VCC) technique that transfers a voice call between the CS and the PS domains. When a UE switches from one domain to another during a VCC call, the bearer in the old domain is released, and a bearer is established in the new domain. This chapter proposed the Bearer Reservation with Preemption (BRP) scheme to support fast and seamless domain transfer. When the UE switches the call from the CS domain to the PS domain, instead of releasing the CS bearer, this CS bearer is reserved with low priority. When the UE switches the call back to the CS domain, the domain transfer process simply raises the priority level of the reserved CS bearer to high priority. Through the preemption mechanism, the reserved bearers in the BRP scheme do not occupy the resources in the MSC for other normal calls. The percentage of re-connection overhead saved by BRP over the 3GPP procedures is denoted by θ_{P2C} for domain transfer from the PS domain to the CS doamin. From the BRP performance study, we observe the following:

- As V_d or V_c increases, θ_{P2C} increases. When V_d is large, θ_{P2C} is not sensitive to the change of V_c.

- V_d and V_c significantly affects the resource (i.e., C) allocated to achieve the same θ_{P2C} performance. When C, V_d or V_c is large, θ_{P2C} is sufficiently large, and increasing C simply wastes the resources at the MSC.

The above observations are also true for domain transfer from the CS domain to the PS domain, which indicate that when the user behavior (either in terms of call holding time or movement pattern) is more irregular, the advantage of the BRP scheme becomes more significant.

Chapter 5

Conclusions and Future Work

Universal Mobile Telecommunications System (UMTS) is one of the major standards for the third generation (3G) mobile telecommunications. In UMTS, the core network consists of two service domains: the *circuit-switched* (CS) and the *packet-switched* (PS) domains. The *IP Multimedia Core Network Subsystem* (IMS) is developed in the PS domain to provide Internet services through *Voice over Internet Protocol* (VoIP) technology. In integration of UMTS and Internet, call control has become one of the most fundamental issues in telecommunications. As IMS evolves, more new call control features are supported. This dissertation investigated how the IMS call control affects three major applications: emergency call, *Push-to-talk over Cellular* (PoC), and *Voice Call Continuity* (VCC). 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 the Active Location Reporting scheme to improve the performance of location tracking for IMS emergency call. In IMS, an emergency call is established by an *Emergency-Call Session Control Function* (E-CSCF). The E-CSCF dispatches the call to the nearest *Public Safety Answering Point* (PSAP) according to the location of the caller. After emergency call setup, the caller's location is tracked by the PSAP through Location Polling. The problems of existing mechanism include: (1) the caller does not change its location between two queries, then the second query is redundant; (2) the caller has moved to several new locations between two queries, then the PSAP may lose track of the caller. For example, if the caller randomly moves between service areas, and the PSAP randomly queries the caller's location, then 40% of service area crossings are mis-tracked, and 20% of queries are redundant. To resolve this issue, we proposed the Active Location Reporting scheme which reports the caller's location in real time without redundant query.

In Chapter 3, we proposed an analytic model to study the performance of the talk burst control mechanism for IMS PoC. PoC provides a "walkie-talkie"-like group communication method in the IMS. The PoC communication is half-duplex, where only one PoC member speaks at a time, and the others listen. The speak permission is arbitrated through the talk burst control mechanism; the design of the talk burst control mechanism significantly affects the performance of PoC. Our study indicated that by including the buffer mechanism for PoC (where an ungranted request can be buffered in the queue, and may be granted later), more than twice clients can be supported while maintaining the same granting probability of a request.

In Chapter 4, we investigated VCC technique that transfers a voice call between the CS and the PS domains. When the user switches from one domain to another during a voice call, the connection in the old domain is released, and a connection is established in the new domain. If the user decides to switch the call back to the old domain again, the released bearer in the old domain must be re-established. Such bearer re-establishment contributes extra overload to the domain transfer. To resolve this issue, we proposed the Bearer Reservation with Preemption (BRP) scheme to support fast domain transfer. When the UE switches the call from the CS domain to the PS domain, instead of releasing the CS bearer, this CS bearer is reserved with low priority. When the UE switches the call back to the CS domain, the domain transfer process simply raises the priority level of the reserved CS bearer to high priority. Through the preemption mechanism, the reserved bearers in the BRP scheme do not occupy the resources in the MSC for other normal calls. Our study indicated that when the user behavior (either in terms of call holding time or movement pattern) is more irregular, the advantage of the BRP scheme becomes more significant.

5.2 Future Work

Based on the research results of this dissertation, we suggest the following topics for further study.

Single Radio VCC: In the next generation 3GPP radio access network (called *Evolved UMTS Terrestrial Radio Access Network* or E-UTRAN), only PS domain (with zonal coverage and larger bandwidth) is supported. The UE may need to switch to the legacy UTRAN when the E-UTRAN connection is not available. However, E-UTRAN and UTRAN may be deployed on the same or neighboring frequency bands. In this case, supporting simultaneous signaling on both E-UTRAN and UTRAN radio channels results in severe technical challenges for the design of the physical layer [39]. Therefore, the domain transfer procedures described in Chapter 4 (called

Dual Radio VCC) may not be exercised to switch from E-UTRAN to UTRAN. This issue is called *Single Radio VCC* (SR-VCC) discussed in 3GPP TS 23.216 [40], and merits further investigation.

VCC for Emergency Call: In the current VCC specifications, emergency call is not supported. A feasibility study on VCC support for emergency calls is described in 3GPP TR 23.826 [41]. This study indicates the following issues:

- Since anchoring contributes extra signaling delay, an emergency call should not be anchored for VCC due to call setup delay limitation for emergency call.
- The PSAP may lose track of the UE if the UE switches to a domain that does not support location service.
- In UMTS, an unauthenticated UE (i.e., a handset without *Subscriber Identity Module* (SIM) card) is allowed to establish an emergency call with a temporary identity. Since the temporary identity can be used only in the current domain, technical challenges exist to switch an emergency call established by an unauthenticated UE to another domain.

Based on the above descriptions, the domain transfer procedures described in Chapter 4 do not work well with emergency calls, and need to be enhanced to address the above issues.

PoC for Emergency Service: In the first version of the *Open Mobile Alliance* (OMA) PoC specifications, emergency service has not been considered. As PoC evolves, many new crisis handling features are provided in OMA PoC 2.0 specifications: ad hoc PoC group, multiple PoC session participation, priority control, integration with presence and location services. We will study how these new features affect the IMS PoC performance.

Bibliography

[1] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS); Service Description; Stage 2. Technical Specification 3GPP TS 23.060 version 7.8.0 (2008-09), 2008.

[2] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; IP Multimedia Subsystem Stage 2. Technical Specification 3GPP TS 23.228 version 7.7.0 (2007-03), 2007.

[3] IETF. SIP: Session Initiation Protocol. IETF RFC 3261, 2002.

[4] Lin, Y.-B. and Pang, A.-C. *Wireless and Mobile All-IP Networks*. John Wiley & Sons, Inc., 2005.

[5] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Open Service Access (OSA); Stage 2. Technical Specification 3GPP TS 23.198 version 8.0.0 (2008-06), 2008.

[6] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Customised Applications for Mobile network Enhanced Logic (CAMEL); Service description, Stage 1. Technical Specification 3GPP TR 22.078 version 8.0.0 (2008-12), 2008.

[7] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Stage 2 functional specification of User Equipment (UE) positioning in UTRAN. Technical Specification 3GPP TS 25.305 version 7.3.0 (2007-10), 2007.

[8] Zhao, Y. Standardization of Mobile Phone Positioning for 3G Systems. *IEEE Communications Magazine*, 40(7):108–116, 2002.

[9] OMA. Push to Talk over Cellular (PoC) - Architecture. OMA-AD-PoC-V2_0_1-20080226-C Candidate Version 2.0 - 26 Feb 2008.

[10] OMA. Push to Talk over Cellular (PoC) - User Plane. OMA-TS-PoC-V2_0_1-20080226-C Candidate Version 2.0 - 26 Feb 2008.

[11] IETF. RTP: A Transport Protocol for Real-Time Applications. IETF RFC 3550, 2003.

[12] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Voice Call Continuity (VCC) between Circuit Switched (CS) and IP Multimedia Subsystem (IMS); Stage 2. Technical Specification 3GPP TS 23.206 version 7.5.0 (2007-12), 2007.

[13] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; IP Multimedia Subsystem (IMS) emergency sessions. Technical Specification 3GPP TS 23.167 version 7.11.0 (2008-12), 2008.

[14] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Functional stage 2 description of Location Services (LCS). Technical Specification 3GPP TS 23.271 version 7.9.0 (2007-09), 2007.

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

[16] FarEasTone Telecom. Private communication. 2003.

[17] Pang, A.-C. and Chen, Y.-K. A Multicast Mechanism for Mobile Multimedia Messaging Service. *IEEE Transactions on Vehicular Technology*, 53(6):1891–1902, 2004.

[18] Yang, S.-R. Dynamic Power Saving Mechanism for 3G UMTS System. *ACM/Springer Mobile Networks and Applications*. published online, 2006.

[19] IETF. SDP: Session Description Protocol. IETF RFC 4566, 2006.

[20] Hsu, S.-F., Lin, Y.-C., Lin, Y.-B. and Yang, J.-S. An OSA Application Server for Mobile Services. *International Journal of Pervasive Computing and Communications*, 3(1):102–113, 2007.

[21] Wu, L.-Y., Tsai, M.-H., Lin, Y.-B., and Yang, J.-S. A Client-Side Design and Implementation for Push to Talk over Cellular Service. *Wireless Communications & Mobile Computing*, 7(5):539–552, 2007.

[22] Brandt, A., Brandt, M., Rugel, S., and Weber, D. Admission Control for Real-time Traffic: Improving Performance of Mobile Networks by Operating on Actual Throughput. *IEEE Wireless Communications and Networking Conference (WCNC), New Orleans*, March 2005.

[23] Kim, P., Balazs, A., van den Brock, E., Kieselinann, G., and Bohm, W. IMS-based push-to-talk over GPRS/UMTS. *IEEE Wireless Communications and Networking Conference (WCNC), New Orleans*, March 2005.

[24] O'Regan, E. and Pesch, D. Performance Estimation of a SIP based Push-to-Talk Service for 3G Networks. *European Wireless Conference (EW), Barcelona*, February 2004.

[25] Lin, Y.-B. Performance Modeling for Mobile Telephone Networks. *IEEE Network Magazine*, 11(6):63–68, November/December 1997.

[26] Kleinrock, L. *Queueing Systems: Volume I – Theory*. New York: Wiley, 1976.

[27] Rey, R.F. *Engineering and Operations in the Bell System*. AT&T Bell Laboratories, 1989.

[28] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Voice Call Continuity (VCC) between Circuit Switched (CS) and IP Multimedia Subsystem (IMS); Stage 3. Technical Specification 3GPP TS 24.206 version 7.5.0 (2008-12), 2008.

[29] 3GPP. 3rd Generation Partnership Project; Technical Specification Core Network; IP Multimedia Subsystem Cx and Dx Interfaces; Signaling Flows and Message Contents (Release 7). Technical Specification 3GPP TS 29.228 version 7.11.0 (2009-03), 2009.

[30] ITU-T. Gateway Control Protocol: Version 3. Technical Report Recommendation H.248.1, ITU-T, 2005.

[31] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Customised Applications for Mobile network Enhanced Logic (CAMEL) Phase 4; Stage 2. Technical Specification 3GPP TS 23.078 version 7.9.0 (2007-09), 2007.

[32] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; enhanced Multi-Level Precedence and Pre-emption service (eMLPP); Stage 1. Technical Specification 3GPP TS 22.067 version 7.0.0 (2006-03), 2006.

[33] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Multimedia priority service. Technical Specification 3GPP TS 22.153 version 8.2.0 (2008-09), 2008.

[34] Gross, D. and Harris, C.M. *Fundamentals of Queueing Theory, 3rd Ed.* John Wiley & Sons, 1998.

[35] Shen, W. and Zeng, Q.-A. Two Novel Resource Management Schemes for Integrated Wireless Networks. *International Conference on Information Technology, Las Vegas*, April 2007.

[36] Chen, W.-E., Hung, H.-N. and Lin, Y.-B. Modeling VoIP Call Holding Times for Telecommunications. *IEEE Network Magazine*, 21(6):22–28, December 2007.

[37] Lin, Y.-B., Mohan, S., and Noerpel, A. Queueing Priority Channel Assignment Strategies for Handoff and Initial Access for a PCS Network. *IEEE Trans. Veh. Technol.*, 43(3):704–712, 1994.

[38] Yang, S.-R. Dynamic Power Saving Mechanism for 3G UMTS System. *ACM/Springer Mobile Networks and Applications*, 12(1):5–14, 2007.

[39] Salkintzis, A. and Hammer, M. and Tanaka, I. and Wong, C. Voice Call Handover Mechanisms in Next-generation 3GPP Systems. *IEEE Communications Magazine*, 47(2):46–56, 2009.

[40] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Single Radio Voice Call Continuity (SRVCC); Stage 2. Technical Specification 3GPP TS 23.216 version 8.4.0 (2009-06), 2009.

[41] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Feasibility Study on Voice Call Continuity Support for Emergency Calls. Technical Report 3GPP TR 23.826 version 9.0.0 (2009-03), 2009.

Appendix A

Simulation Model for Location Polling

This appendix describes the discrete event simulation model for Location Polling. The following attributes are defined for an event **e**.

- The **type** attribute indicates the event type. A **Query** event represents a query from the PASP. An **SA Crossing** event represents that the user moves from one SA to another.
- The **time** attribute indicates the time when the event occurs.

In the simulation model, the N_Q counter represents the number of queries occurred in the current SA residence time interval, τ is the invalid period, and t_{SA} is the time when the user enters this SA. The output measures of the simulation are the total number N_{SA} of SA crossings, the total number N_Q^* of queries in SA residence time intervals with more than one query, the number N_0 of SA residence time intervals without any query, the number N_1 of SA residence time intervals with more than one query. Let τ^* be the summation of τ's, and τ_{sq}^* be the summation of τ^2's. From the above output measures, we compute

$$\alpha = N_0/N_{SA},\ E[\tau_q|N \geq 1] = \tau^*/(N_{SA} - N_0),$$

$$V_q = \tau^*_{sq}/(N_{SA} - N_0) - T_i^2,$$

$$\beta = N_1/N_{SA},\ \text{and } E[N|N > 1] = N^*_Q/N_1 \tag{A.1}$$

A simulation clock is maintained to indicate the simulation progress, which is the timestamp of the event being processed. All events are inserted into the event list, and are deleted/processed from the event list in the non-decreasing timestamp order. Figure C.1 illustrates the simulation flow chart for Location Polling. In this flow chart, Step 1 initializes the output parameters. Step 2 generates the first **Query** and **SA Crossing** events and inserts them into the event list. N_{SA} is incremented by one. At Steps 3 and 4, the first event **e** in the event list is processed based on its type described as follows:

Query: N_Q is incremented by one at Step 5. Step 6 checks if this **Query** event is the first **Query** event occurred in the current SA residence time interval (i.e., $N_Q = 1$). If so, Step 7 computes τ, τ^* and τ^*_{sq}. Step 8 generates the next **Query** event and inserts it into the event list.

SA Crossing: Step 9 checks if there is any **Query** event occurred between this **SA Crossing** event and the previous **SA Crossing** event (i.e., $N_Q > 0$). If not, N_0 is incremented by one at Step 10, and the simulation proceeds to Step 13. If $N_Q > 1$ at Step 11, then N_1 is incremented by one, and N^*_Q is incremented by N_Q at Step 12. Step 13 sets t_{SA} to the timestamp of the current event (i.e., **e.time**), resets N_Q and τ to 0, generates the next **SA Crossing** event and inserts it into the event list. N_{SA} is incremented by one. Step 14 checks if $N_{SA} > 1,000,000$. If so, Step 15 computes the output measures (C.1), and the simulation terminates.

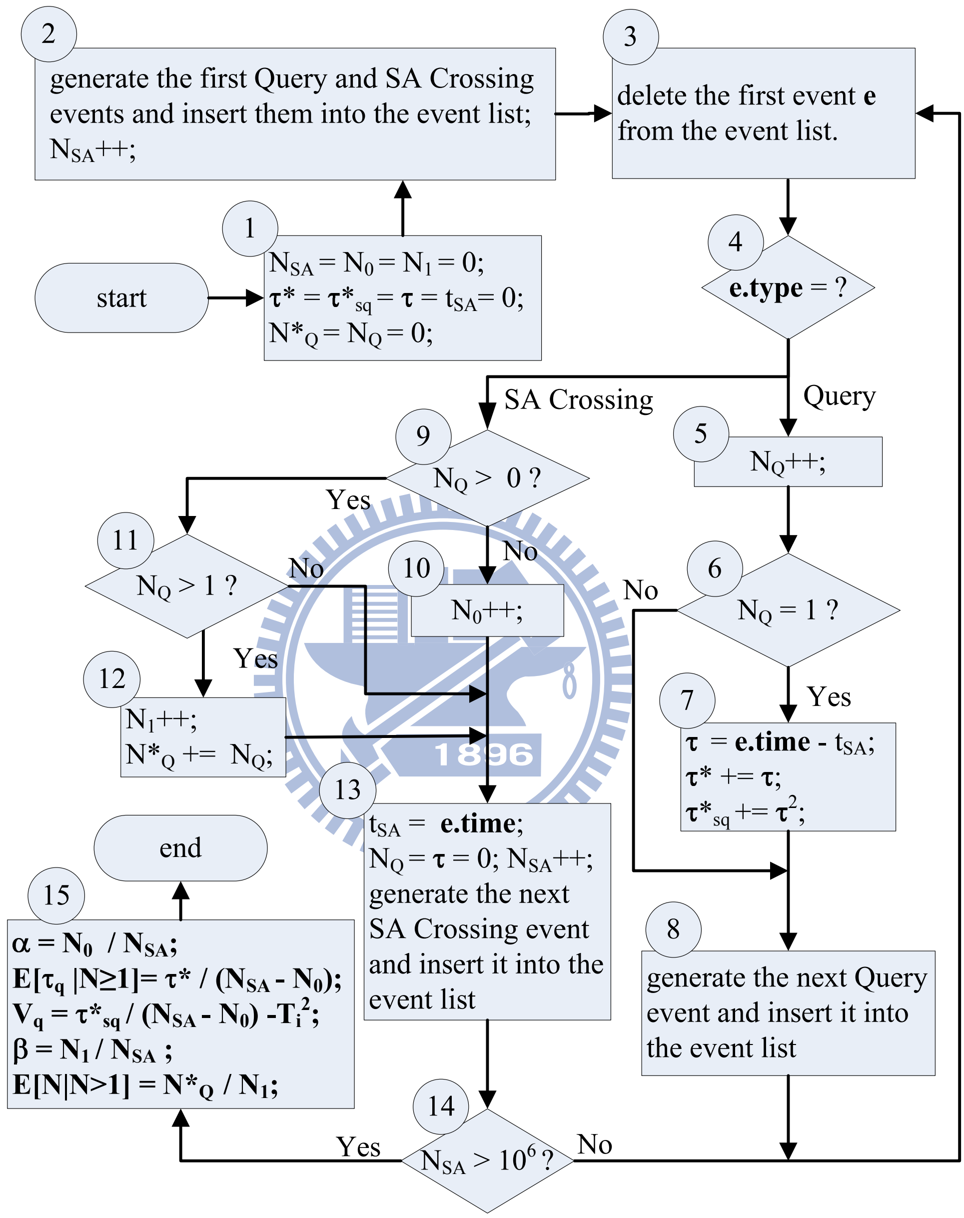


Figure A.1: Simulation Flow Chart for Location Polling

Appendix B

Simulation Model for Approach Q

This appendix describes the discrete event simulation model for Approach Q. In this model, four types of events are defined to represent request arrival, release, revoke and impatience. The following attributes are defined for an event e.

- The **type** attribute indicates the event type. An **Arrival** event represents a request arrival. A **Release** event represents a talk burst release. A **Revoke** event represents a timeout of the revoking timer T_R. An **Impatience** event represents that a queued PoC client leaves the queue due to impatience.
- The **ts** attribute indicates the time when the event occurs.
- The **id** attribute uniquely identifies a PoC client.
- The **seq** attribute indicates the sequence number of request arrival. This attribute is used for **Impatience** event to recognize its corresponding **Arrival** event. An **Impatience** event will cause a queued **Arrival** event with the same **seq** value to leave the queue.

In the simulation model, Q_l is used to represent the queue length, T_R is the revoking timer described in Section 3.3, and S_t is used to represent the talk burst status (taken or idle). The output measures of the simulation are the total number N_a of request arrivals, the number N_d of dropped requests due to impatience of the member, the number N_r of revoked talk bursts, and the total waiting time W_t. From the above output measures, we can compute

$$P_D = N_d/N_a, P_R = N_r/N_a, \text{and } W = W_t/N_a. \tag{B.1}$$

A simulation clock is maintained to indicate the simulation progress, which is the timestamp of the event being processed. All events are inserted into the event list, and are deleted/processed from the event list in the non-decreasing timestamp order. An **Arrival** event is inserted into the queue if the talk burst status is taken (it means that some PoC client is speaking, and the request is queued in the PoC server). Figure B.1 illustrates the simulation flow chart for Approach Q. In this flow chart, Step 1 initializes the input parameters. For each PoC member, the first **Arrival** event is generated and is inserted into the event list at Step 2. At Steps 3 and 4, the first event e in the event list is processed based on its type described as follows:

Arrival: N_a is incremented by one at Step 5. At Step 6, if $N_a > 1,000,000$, then Step 7 computes the output measures using (B.1), and the simulation terminates. Otherwise, Step 8 checks the talk burst status S_t. If $S_t = IDLE$, Step 9 generates speaking time s for event e, and sets S_t to $TAKEN$. Step 10 checks if $s > T_R$. If so, a **Revoke** event f with $f.id = e.id$ and $f.ts = e.ts + T_R$ is generated and is inserted into event list at Step 11 (it means that the permitted PoC client will be revoked later). Otherwise, a **Release** event f with $f.id = e.id$ and $f.ts = e.ts+s$ is generated and is inserted into event list at Step 12 (it means that the permitted PoC client

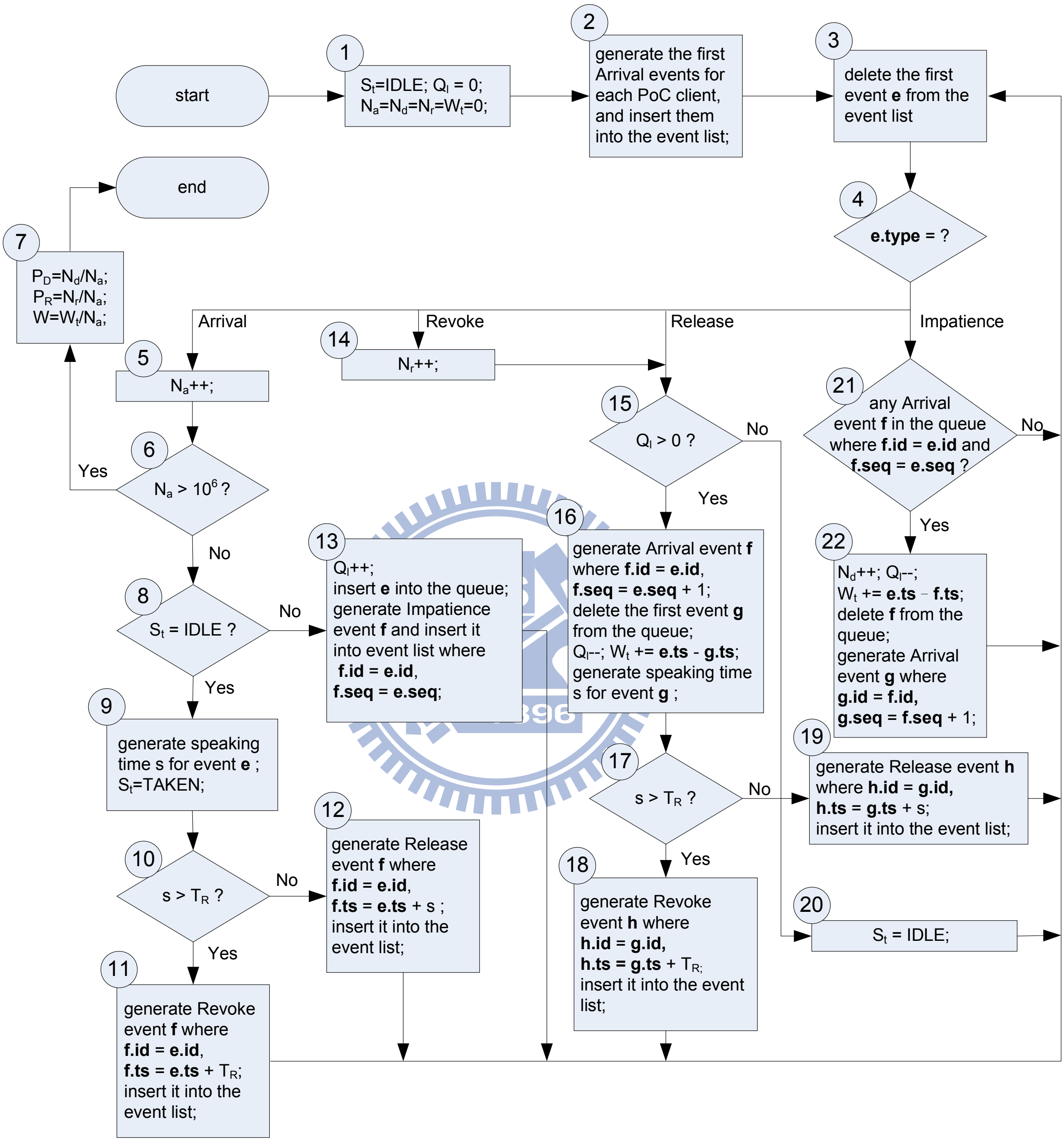


Figure B.1: Simulation Flow Chart for Approach Q will normally finish speaking). If $S_t = TAKEN$ at Step 8, Q_l is incremented by one, event e is inserted into the queue, and an **Impatience** event f with $f.id = e.id$ and $f.seq = e.seq$ is generated and is inserted into the event list at Step 13.

Release: Step 15 checks if there is any queued **Arrival** event. If so, Step 16 generates the next **Arrival** event f with $f.id = e.id$ and $f.seq = e.seq + 1$, deletes the first event g from the queue, decrements Q_l by one, add $e.ts - g.ts$ (waiting time for event g) to W_t, and generates speaking time s for event g. Steps 17-19 are the same as Steps 10-12 except that the corresponding **Arrival** event g is from the queue (not from the event list). If $Q_l \leq 0$ at Step 15, S_t is set to $IDLE$ at Step 20.

Revoke: N_r is incremented by one at Step 14. Steps 15-20 described above are exercised.

Impatience: Step 21 checks if there is any corresponding **Arrival** event f (i.e., $f.id = e.id$ and $f.seq = e.seq$) in the queue. If so, Step 22 increments N_d by one, decrements Q_l by one, add $e.ts - f.ts$ (waiting time for event f) to W_t, deletes event f from the queue, and generates the next **Arrival** event g where $g.id = f.id, g.seq = f.seq + 1$ and inserts this event into the event list.

Appendix C

Simulation Model of Bearer Reservation with Preemption

This appendix describes a discrete event simulation model for the BRP scheme. Without loss of generality, we investigate the BRP performance in the CS domain when the new calls arrive at the MSC are VCC calls (i.e., we do not consider non-VCC calls). Similar conclusions also apply to the PS domain, and the details are omitted. Three types of input parameters are considered in the simulation model.

System Parameter: There are C channels in the MSC.

Traffic Parameters: The CS call arrivals to the MSC form a Poisson stream with rate λ. The expected call holding time is $1/\mu$ with variance V_c.

Mobility Parameter: The expected domain residence time for each domain is $1/\delta$ with variance V_d.

In our simulation model, the following attributes are defined for an event.

- The **type** attribute indicates the event type. An **Arrival** event represents a VCC CS call arrival (i.e., a high-priority call arrival in the MSC). A **CS2PS** event represents that a VCC CS call switches to the PS domain (and the call becomes low-priority at the MSC). A **PS2CS** event represents that a VCC PS call switches back to the CS domain (and the call becomes high-priority at the MSC). A **Completion** event represents a call completion.

- The **callId** attribute specifies the ID of the call.
- The **priority** attribute indicates high-priority (with value 1) or low-priority (with value 0). This attribute is used in a **Completion** event.
- The **tc** attribute indicates the remaining call holding time.
- The **td** attribute indicates the domain residence time in the current domain.

All events are inserted into the event list, and are deleted/processed from the event list in the non-decreasing timestamp order. A simulation clock is maintained to indicate the simulation progress, which is the timestamp of the event being processed.

In this simulation model, the N_l counter records the number of low-priority calls in the MSC, and the N_c counter records the number of available channels in the MSC. The output measures of the simulation are the total number N of **PS2CS** events, the total number N_r of CS bearer re-establishments and the total number N_f of force-terminations. From the above output measures, we compute

$$p_f = N_f/N,\ p_r = N_r/N,\ \text{and } \bar{p}_n = p_f + p_r \tag{C.1}$$

The simulation model uses a queue `LQueue` to record the low-priority calls in the MSC. When a **CS2PS** event occurs (i.e., the VCC CS call becomes low-priority at the MSC), the **callId** attribute of this event is inserted at the tail of LQueue. If an **Arrival** event occurs when $N_c = 0$ and $N_l > 0$, then according to the First-Come-Last-Preempted rule, the **callId** at the tail of LQueue is deleted (i.e., the high-priority call arrival preempts a low-priority call).

Figure C.1 illustrates the simulation flow chart for the BRP scheme. In this flow chart, Step 1 initializes the parameters. Step 2 generates the first **Arrival** event **h** where **h.priority = 1** and this event is inserted into the event list. At Steps 3 and 4, the first event **e** in the event list is processed based on its type described as follows:

Arrival: Step 5 generates the next **Arrival** event **f** where **f.priority = 1** and inserts it into the event list. Step 6 checks whether $N_c > 0$ (i.e., there is available channel in the MSC). If so, N_c is decremented by one at Step 7 (i.e., this CS call occupies a channel in the MSC). Otherwise, Step 8 checks whether $N_l > 0$ (i.e., there exists a low-priority call for preemption in the MSC). If so, the **callId** at the tail of LQueue is deleted, and N_l is decremented by one at Step 9 (i.e., the last-coming low-priority call is preempted). If $N_c > 0$ at Step 6 or $N_l > 0$ at Step 8 (i.e., a channel is available), Step 10 compares the remaining call holding time **e.tc** and the domain residence time **e.td**. If **e.tc** < **e.td** (this call completes before it switches to the PS domain), Step 11 generates a **Completion** event **g** where **g.priority = 1** for this high-priority call completion, and inserts it into the event list. Otherwise, Step 12 generates a **CS2PS** event **h** where **h.tc = e.tc - e.td** and **h.priority = 0**, and inserts it into the event list.

CS2PS: Step 13 inserts **e.callId** at the tail of LQueue, and increases N_l by one. If **e.tc** < **e.td** at Step 14, then Step 15 generates a **Completion** event **g** where **g.priority** =

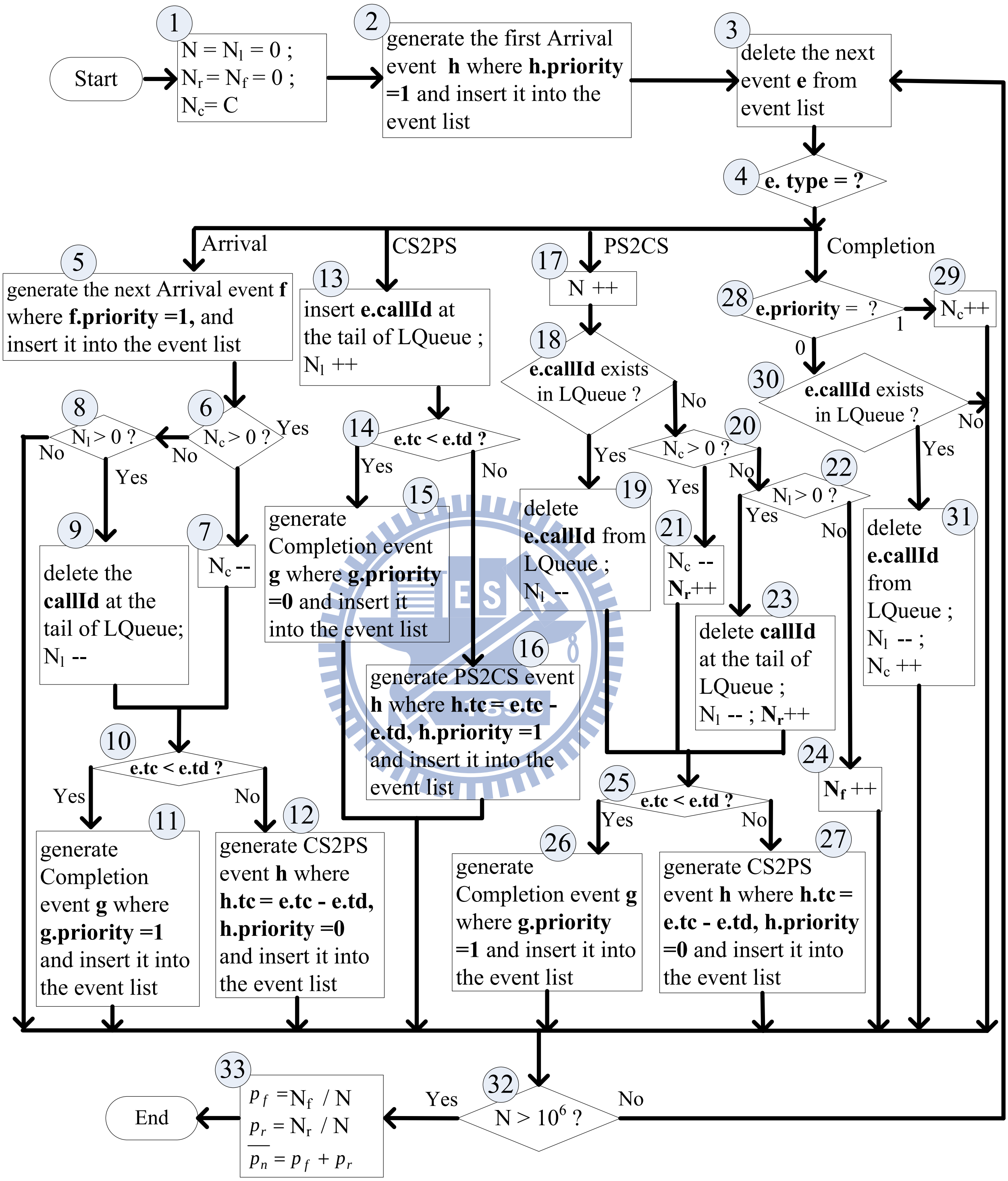


Figure C.1: Simulation Flow Chart for the BRP Scheme

0 for this low-priority call completion, and inserts it into the event list. Otherwise, Step 16 generates a **PS2CS** event **h** where **h.tc** = **e.tc - e.td** and **h.priority** = **1**, and inserts it into the event list.

PS2CS: N is incremented by one at Step 17. If **e.callId** exists in LQueue at Step 18 (i.e., the reserved CS bearer is still available), then Step 19 deletes **e.callId** from LQueue, and N_l is decremented by one. Otherwise, Step 20 checks whether $N_c > 0$. If so, N_r is incremented by one, and N_c is decremented by one at Step 21 (i.e., the CS bearer is re-established). If $N_c = 0$ at Step 20, Step 22 checks whether $N_l > 0$. If so, the **callId** at the tail of LQueue is deleted, N_r is incremented by one, and N_l is decremented by one at Step 23 (i.e., the last-coming low priority call is preempted). If $N_c = 0$ and $N_l = 0$ at Step 22, N_f is incremented by one at Step 24 (i.e., this call is force-terminated). If **e.callId** exists in LQueue, $N_c > 0$ or $N_l > 0$, Step 25 compares the remaining call holding time **e.tc** and the domain residence time **e.td**. If **e.tc** < **e.td**, Step 26 generates a **Completion** event **g** where **g.priority** = **1** and inserts it into the event list. Otherwise, Step 27 generates a **CS2PS** event **h** where **h.tc** = **e.tc - e.td** and **h.priority** = **0**, and inserts it into the event list.

Completion: There are two cases at Step 28:

e.priority = 1. N_c is incremented by one at Step 29 (i.e., a channel in the MSC is released).

e.priority = 0. If **e.callId** exists in LQueue at Step 30, then Step 31 deletes the **e.callId** from LQueue, decreases N_l by one, and increases N_c by one (i.e., the reserved channel for this VCC PS call is released).

At the end of each iteration, Step 32 checks if $N > 1,000,000$ (which is large enough to produce stable simulation results). If so, Step 33 computes the output measures by (C.1), and the simulation terminates.

Curriculum Vitae

Meng-Hsun Tsai received the B.S. and the M.S. degrees from National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C., in 2002 and 2004, respectively. His current research interests include design and analysis of personal communications services networks, mobile computing and performance modeling.

Publication List

- Journal Papers

    1. Tsai, M.-H., Lin, Y.-B. and Wang, H.-H. Active Location Reporting for Emergency Call in UMTS IP Multimedia Subsystem. Accepted and to appear in *IEEE Transactions on Wireless Communications*.
    2. Gan, C.-H., Tsai, M.-H. and Lin, Y.-B. Efficient Routing for International Mobile Call Setup. Accepted and to appear in *IEEE Wireless Communications Magazine*.
    3. Lin, Y.-B., Tsai, M.-H., Dai, H.-W. and Chen, Y.-K. Bearer Reservation with Preemption for Voice Call Continuity. *IEEE Transactions on Wireless Communications*, 8(5): 2716-2725, 2009.
    4. Tsai, M.-H. and Lin, Y.-B. Talk Burst Control for Push-to-talk over Cellular. *IEEE Transactions on Wireless Communications*, 7(7): 2612-2618, 2008.
    5. Chlamtac, I., Lee, H.-Y., Lin, Y.-B. and Tsai, M.-H. An OSA Service Capability Server for Mobile Services. *International Journal of Pervasive Computing and Communications*, 4(3):268-278, 2008.
    6. Lin, Y.-B. and Tsai, M.-H. Eavesdropping through Mobile Phone. *IEEE Transactions on Vehicular Technology*, 56(6): 3596-3600, 2007.
    7. Wu, L.-Y., Tsai, M.-H., Lin, Y.-B. and Yang, J.-S. A Client-Side Design and Implementation for Push to Talk over Cellular Service. *Wireless Communications and Mobile Computing*, 7(5): 539-552, 2007.
    8. Lin, Y.-B. and Tsai, M.-H. Caching in I-CSCF of UMTS IP Multimedia Subsystem. *IEEE Transactions on Wireless Communications*, 5(1):186-192, 2006.

- Conference Papers

    1. Tsai, M.-H. and Dai, H.-W. Bearer Reservation with Preemption for Voice Call Continuity. *In Proceeding of the 10th International Conference on Mobile Data Management: Systems, Services and Middleware (MDM 2009)*, 2009.

    2. Tsai, M.-H., Lin, Y.-B. and Wang, H.-H. Active Location Reporting for Emergency Call in UMTS IP Multimedia Subsystem. *In Proceeding of the 10th International Conference on Mobile Data Management: Systems, Services and Middleware (MDM 2009)*, 2009.

    3. Lee, H.-Y., Lin, Y.-B. and Tsai, M.-H. An OSA Service Capability Server for Mobile Services. *12th Mobile Computing Workshop*, Taichung, March 2006.

- Patents

    1. Gan, C.-H., Tsai, M.-H., Lin, Y.-B. and Liang, C.-F. METHOD AND GATEWAY FOR ROUTING INTERNATIONAL MOBILE TELEPHONE CALLS. Submitted to the U.S., R.O.C. and P.R.O.C. patent offices. (pending)

    2. Lin, Y.-B., Tsai, M.-H. and Yang, J.-S. SYSTEM AND METHOD FOR ACCELERATING CALL SETUP BY CACHING. R.O.C. Patent, Patent Number: I 252027, Assignee: Computer and Communications Research Lab/ Industrial Technology Research Institute (CCL/ITRI), R.O.C. (R.O.C. granted; U.S. pending)