國 立 交 通 大 學

資訊工程學系

博士論文

在無線網路中針對優先權傳送之排程技術研究

Scheduling Techniques for Priority Transmission in Wireless Networks

研究生：賈仲雍

指導教授：張明峰 博士

中華民國九十八年七月

在無線網路中針對優先權傳送之排程技術研究

Scheduling Techniques for Priority Transmission in Wireless Networks

研 究 生：賈仲雍　　Student： Chung-Yung Chia

指導教授：張明峰 博士　　Advisor： Dr. Ming-Feng Chang

國 立 交 通 大 學

資 訊 工 程 學 系

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 and Engineering

July 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年七月

在無線網路中針對優先權傳送之排程技術研究

學生：賈仲雍　　指導教授：張明峰 博士

國立交通大學資訊科學與工程研究所

摘要

在越來越競爭的電信市場上，電信業者如何能同時滿足用戶服務的滿意度和保持各項業務的盈收是相形重要。論文中針對幾種不同無線網路型態(如 GPRS, 3G，HSDPA)的用戶在使用無線資源時，除考量傳送或接收封包優先權排程方式之外，另針對吃到飽(flat-rate)用戶，考量其對行動業者的最有效率優先權排程機制，做了廣泛且深入之研究。

首先在 GSM 行動網路中，我們研究和比較 3 種不同排程方式，如何在用戶傳送上傳(uplink)封包時，控制和調整用戶的上傳封包的傳送優先權，透過建立數學分析模式和幾種不同分析參數，比較何種優先權考量可滿足用戶服務滿意度。研究發現利用上鏈路狀態旗標(USF)配置的排程方式可以讓有較高優先權封包(如 VoIP)，其傳輸延遲可以比較少。

接著，在 UMTS 行動網路中，我們研究和比較 4 種不同排程方式，如何在用戶啟動上傳連線(uplink connection transmission)時，控制調整用戶的上傳連線的傳送優先權，透過建立數學分析模式、幾種不同分析參數和考量吃到飽用戶所提出的代價函數(cost function)，來比較何種優先權考量可同時滿足用戶服務滿意度和保持電信業者盈收。研究發現單憑降低用戶的上傳連線速度，並不會讓整體用戶的上傳連線傳送效能最高和讓電信業者擁有最大盈收，若利用等待隊列(waiting queue)和強制取代(preemption)的排程方式，可以同時兼顧用戶服務滿意度和讓電信業者擁有最大盈收。

最後，在 HSDPA 行動網路中，我們研究和比較 4 種不同排程方式，如何在用戶傳送下傳(downlink)封包時，控制調整用戶的下傳封包的傳送優先權，透過建立數學分析模式、幾種不同分析參數和考量吃到飽用戶所提出的代價函數(cost function)，比較何種優先權考量可同時滿足用戶服務滿意度和保持電信業者盈收。研究發現若利用等待隊列(waiting queue)和固定強制取代(preemption)的排程方式，並不會讓整體用戶的下傳封包傳送效能最高和讓電信業者擁有最大盈收，若考量加入動態丟棄計時器(Drop Timer)和動態防護通道(Guard Slot)的排程方式，來動態調整用戶下傳封包的傳送優先權，可以同時兼顧用戶服務滿意度和讓電信業者擁有最大盈收。

在無線網路發展趨勢上，電信業者必須持續不斷的研究如何能同時滿足用戶服務的滿意度和保持各項業務的盈收。本篇論文研究的結果可被當成繼續研究在無線網路中如何考量優先權排程方式的基礎。

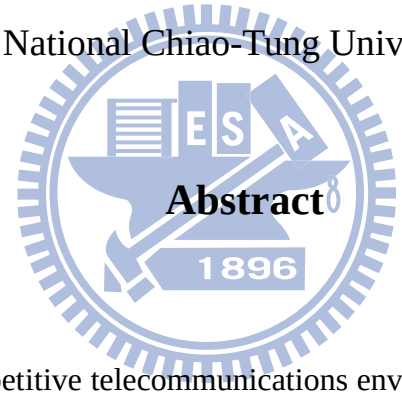
關鍵字：優先權傳送、點陣圖通道配置、上鏈路狀態旗標、吃到飽服務、連線排程、動態丟棄計時器和動態防護通道。

Scheduling Techniques for Priority Transmission in Wireless Networks

Student: Chung-Yung Chia Advisor: Dr. Ming-Feng Chang

Institute of Computer Science and Engineering

National Chiao-Tung University



Abstract

In today's highly competitive telecommunications environment, the emphasis has shifted to delivering innovative services to satisfy increasingly sophisticated customers' need and to improve the revenues of wireless operators. This dissertation has studied different scheduling mechanisms for priority transmission in public land mobile networks (PLMNs). We considered not only the priority in packet transmission/reception using various scheduling techniques, but also the most efficient mechanisms for serving both normal and flat-rate customers.

First, we study and compare three different scheduling mechanisms in the GSM network. As the General Packet Radio Service (GPRS) network begins to provide such as "push-to-talk" (PTT) service, delay-sensitive packets should be given higher priority in transmission. In this paper, we study two channel allocation schemes that implement priority

queues for priority packets in the GPRS network: Bitmap Channel Allocation (BCA) and Uplink State Flag Channel Allocation (USFCA). Our study shows that the transmission delay of priority packets in the GPRS network can be better guaranteed using USFCA.

Second, we study and compare four different scheduling mechanisms in the UMTS network. To attract more users to mobile packet services, the Universal Mobile Telecommunication System (UMTS) operators have been prompting flat-rate packet services. Since usage does not incur cost, flat-rate users tend to stay on line longer and occupy most of the radio channel resources. We consider a UMTS network serving two types of user connections: Normal User Connections (NUCs) and Flat-Rate User Connections (FRUCs). Our goal is to maximize the revenue of the operator by giving a priority to NUCs over FRUCs without discontenting the flat-rate users, in order not to lose the flat-rate users to other operators. Uplink FRUCs may be asked to sub-rate or suspend transmission when the radio network is fully utilized. Four combinations of scheduling techniques including queueing, guard channels, preemption and rate-adaptation, have been studied, and analytic models using Markov processes were used to evaluate their performances. We proposed a cost function representing the revenue loss due to both blocked NUCs and lost flat-rate users. The system parameters used in our analysis are based on realistic operation data. Our analytic results indicate that the revenue loss can be minimized by using waiting queues and preemption. Rate-adaptation is ineffective in minimizing the revenue loss because sub-rated connections are less efficient in using radio spectrum. Guard channels for NUCs are unnecessary when waiting queue or preemption is used. Our study may be valuable for UMTS operators in serving flat-rate users.

Third, we study and compare four different scheduling mechanisms in the HSDPA network. We consider a HSDPA network serving two types of user packets: charged packets (CPs) and flat-rate packets (FRPs). Since CPs are charged by usage, they are given a higher priority to receive downlink packets for revenue consideration. However, this priority

preference may lead to poor quality of service for flat rate users. In particular, FRPs may experience longer transmission latency and higher dropped probability. We should consider the balance between serving the FRPs and CPs. Analytic models using Markov process were used to study their performance. Our study shows that DDT-PQ and DGS-PQ methods are more effective to transmit the downlink CPs especially when the downlink FRP traffic is high. Therefore, they are better in guaranteeing the system throughput for CPs, and thus the operator revenue can be better protected.

In the development trend of wireless network, wireless operators need to keep studying how to satisfy the QoS requirements of the customers and have the best revenues at the same time. The research results presented in this dissertation can be viewed as a useful foundation for further study in the scheduling mechanisms for priority transmission in the wireless network.

Key Words: Priority Transmission, Bitmap Channel Allocation (BCA), Uplink State Flag (USF), Flat-Rate Service, Connection Scheduling, Dynamic Discard Timer (DDT) and Dynamic Guard Slot(DGS)

Acknowledgement

I would like to express my sincere thanks to my advisors, Prof. Ming-Feng Chang. Without their supervision and perspicacious advice, I can not complete this dissertation. Special thanks to my committee members, Prof. Chien-Chao Tseng, Prof. Kuo-Chen Wang, Prof. Duan-Shin Lee, Prof. Ai-Chun Pang, Dr. Sheng-Lin Chou and Dr. Kuang-Yao Chang for their valuable comments. Thanks also to the colleagues in Internet Communication Laboratory.

I also express my appreciation to all the faculty, staff and colleagues in the Institute of Computer Science and Engineering, NCTU and Wireless Communications Laboratory, Telecommunication Laboratories of CHT. In particular, I would like to thank Dr. Chen and Dr. Yoau for their friendship and support in various ways.

Finally, I am grateful to my family, my wife, Shiely, my children Arron and Sarah, and friends for their encouragement and support during these years.

Contents

Abstract in Chinese i

Abstract in English iii

Acknowledgement vi

Contents vii

List of Figures x

List of Tables xii

Abbreviation xiiiii

CHAPTER 1 Introduction 1

1.1 Channel Allocation for Priority Packets in the GPRS Network 1

1.2 Uplink Connection Scheduling for Flat-Rate Data Services in the UMTS Network 2

1.3 Flate-Rate Packet Scheduling for the WCDMA Systems with HSDPA 2

1.4 Synopsis of This Dissertation 3

CHAPTER 2 Channel Allocation for Priority Packets in the GPRS Network 4

2.1 Introduction 4

2.2 The Methods of BCA and USFCA 4

2.2.1 BCA Method 5

2.2.2 USFCA Method 5

2.3 The Analytic Models 6

2.3.1 BCA Method 7

2.3.2 USFCA Method 9

2.4 Numeric Results 10

2.5 Conclusions 12

CHAPTER 3 Uplink Connection Scheduling for Flat-Rate Data Services in the UMTS Network 13
3.1 Introduction 13
3.2 System Models and Assumptions 19
3.2.1 CDMA Uplink Capacity Model 22
3.2.2 A Scheduler with all four features 23
3.3 The Analytic Models 25
3.3.1 The Analytic Model of S_{All} 25
3.3.2 The Performance Measures 28
3.3.3 Cost Function Scheme 32
3.3.4 An Iterative Algorithm 34
3.4 Numerical results and Discussions 35
3.5 Conclusions 44
CHAPTER 4 Flate-Rate Packet Scheduling for the WCDMA Systems with HSDPA 46
4.1 Introduction 46
4.2 HSDPA Basic Principles 48
4.3 System Models and Assumptions 51
4.3.1 M-PQ Method 52
4.3.2 P-PQ Method 52
4.3.3 DDT-PQ Method 54
4.3.4 DGS-PQ Method 56
4.4 The Analytic Method 59
4.4.1 M-PQ Method 60
4.4.2 P-PQ Method 61
4.4.3 DDT-PQ Method 63

4.4.4 DGS-PQ Method.....64
4.5 Cost Function Scheme 66
4.6 Numreical Analysis.....68
4.7 Conclusions 77
CHAPTER 5 Conclusions and Future Work 79
5.1 Summary..... 79
5.2 Future Works 81
Bibliography 83
Curriculum Vitae 88
Publication List..... 89

List of Figures

Fig. 2.1: A USFCA example using priority-packet-first scheme 6

Fig. 2.2: The queuing model for BCA and USFCA schemes 7

Fig. 2.3: The mean waiting time and system time of uplink packets 11

Fig. 3.1: The Radio Access Bearer (RAB) assignment procedure 20

Fig. 3.2: The system queueing model for a reference cell 24

Fig. 3.3: An algorithm for the numbers of full-rate and half-rate FRUCs in state (i, j,k)...26

Fig. 3.4: The state transition diagram of S_{All}, and the rates of input/output flows.....27

Fig. 3.5: An iterative algorithm minimizes the cost function.....34

Fig. 3.6.a:The cost function (C) with α =0.504 and ε =0.02, (B=4, Q=10).....36

Fig. 3.6.b:The numbers of guard channels (GC) with α =0.504 and ε =0.02, (B=4, Q=10).37

Fig. 3.7.a: Average NUC blocking probabilities (P_{BN}) with B=4, Q=10.....38

Fig. 3.7.b: Average FRUC blocking probabilities (P_{BF}) with B=4, Q=10.....39

Fig. 3.8.a: Average NUC waiting times (W_{TN}) with B=4, Q=10.....39

Fig. 3.8.b: Average FRUC waiting times (W_{TF}) with B=4, Q=10.....40

Fig. 3.9.a: Average NUC queueing probabilities (P_{QN}) with B=4, Q=10.....41

Fig. 3.9.b: Average FRUC queueing probabilities (P_{QF}) with B=4, Q=10.....41

Fig. 3.10.a: Average preempted probabilities of a serving full-rate FRUCs (P_{FPrm}) with B=4, Q=10.....42

Fig. 3.10.b: Average preempted probabilities of a serving half-rate FRUCs (P_{SPrm}) with B=4, Q=10.....42

Fig. 3.11:Average sub-rated probabilities of a serving full-rate FRUCs (P_{FS}) (B=4, Q=10) 43

Fig. 3.12:Average transmission rate of serving FRUCs (T_F) (B=4, Q=10) 44

Fig. 4.1: The network architecture of the UMTS network 47

Fig. 4.2: Downlink SF codes allocation tree for HS-DSCH and HS-SCCH.....50

Fig. 4.3: An example downlink packet scheduling for MS1-MS4 in a cell.....50

Fig. 4.4: The scheduling parameters sent from an SGSN through a RNC to a Node-B in HSDPA network.....51

Fig. 4.5: A queueing model for M-PQ scheme 53

Fig. 4.6: A queueing model for P-PQ scheme 53

Fig. 4.7: A queueing model for DDT-PQ scheme 54

Fig. 4.8: A peudocode for DDT-PQ scheme 56

Fig. 4.9: A queueing model for DGS-PQ scheme 57

Fig. 4.10: A peudocode for DGS-PQ scheme 59

Fig. 4.11: The state transition diagram of M-PQ scheme 60

Fig. 4.12: The state transition diagram of P-PQ scheme 62

Fig. 4.13: The state transition diagram of DDT-PQ scheme 63

Fig. 4.14: The state transition diagram of DGS-PQ scheme 65

Fig. 4.15: The average dropped probability of CP for four packet scheduling methods 70

Fig. 4.16: The average dropped probability of FRP for four packet scheduling methods 71

Fig. 4.17: The average network utilization of CP for four packet scheduling methods 72

Fig. 4.18: The results of dynamically adjusting the DT value in DDT-PQ scheme 73

Fig. 4.19: The results of dynamically adjusting the GS value in DGS-PQ scheme 74

Fig. 4.20: The cost function (C) when $\alpha > \rho$ 75

Fig. 4.21: The number of GS for CPs and FRPs when $\alpha > \rho$ 76

Fig. 4.22: The value of DT for FRPs and CPs when $\alpha > \rho$ 76

Fig. 4.23: The blocking probabilities of FRPs when $\alpha > \rho$ 77

List of Tables

Table 3.1: System notations in uplink connection scheduling for Flat-Rate Users 16

Abbreviation

The abbreviations used in this dissertation are listed below.

3GPP: 3rd Generation Partnership Project

AP: Access Point

BCA:Bitmap Channel Allocation

BSS: Business Support System

CN:Core Network

CS: Circuited-Switch

CP: Charged Packets

CPICH: Common Pilot Indicator Channel

D_CH: Dedicated radio Channel

DDT: Dynamic Discard Timer

DGS: Dynamic Guard Slots

FDD:Frequency Division Duplex

FCFS:First Come, First Served

FRP: Flat-Rate Packets

FRUC:Flat-Rate User Connections

GC: Guard Channels

GGSN: Gateway GPRS Support Node

GPRS: General Packet Radio Service

GSM: Global System for Mobile Communication

HSDPA: High Speed Downlink Packet Access

HS-DSCH : High Speed Downlink Shared Channel

HS-SCCH : High Speed Shared Control Channel

LB: Lower Bound

Max C/I: Maximum Carrier-to-Interference ratio

MCS: Modulation Coding Scheme

MS: Mobile Station

NUC:Normal User Connections

PASTA:Poisson Arrivals See Time Averages

PDP: Packet Data Protocol

PF: Proportionally Fair

PLMN: Public Land Mobile Network

PS: Packet-Switched

PSTN: Public Switched Telephone Network

PTT: Push To Talk

PQ: Priority Queue

QAM: Quadrature Amplitude Modulation

QoS:Quality of Service

QPSK: Quadrature Phase-Shift Keying

RAB: Radio Access Bearer

RAU:Routing Area Upate

RB:Radio Bearer

RF: Radio Frequency

RNC:Radio Network Controller

RR: Round Robin

RRC:Radio Resource Control

SF: Spreading Factor

SGSN: Serving GPRS Support Node

SIR: Signal to Interference Ratio

SV: Step Value

TDMA:Time Division Multiple Access

TTI: Transmission Time Interval

UB: Upper Bound

UMTS: Universal Mobile Telecommunications System

URI: Universal Resource Identifier

USFCA:Uplink State Flag Channel Allocation

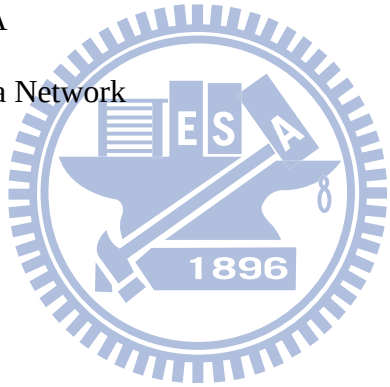
VoIP: Voice over IP

UTRAN:UMTS Terrestrial Radio Access Network

WCDMA: Wideband CDMA

WLAN: Wireless Local Area Network

WQ: Qaiting Queues



CHAPTER 1

Introduction

In recent years, telecommunication industry is growing fast especially in mobile market. Many technologies have been developed and deployed, such as 2G/GPRS/3G/HSDPA, VoIP, and NGN (Next Generation Network). They provide not only the traditional voice communication service but also many advanced data and information services. However, technical advances no longer drive the market trend. In today's highly competitive environment, the emphasis has shifted to delivering innovative service to satisfy increasingly sophisticated customers' need. Customers have changed from being the passive role to active, and operators need to focus on customers' feeling. It is important to satisfy all kinds of customers and make them feel that the service is tailored for them, for their benefits and interests. Of coourse, wireless operators also need to consider the revenues when they provide the services to the customers. To consider the customer's need and the revenues of the operators, they need effective scheduling mechanisms to process the customer uplink/downlink packets. Operators shall make a win-win new telecom business market.

1.1 Channel Allocation for Priority Packets in the GPRS Network

As the General Packet Radio Service (GPRS) network begins to provide such as "push-to-talk" (PTT) service, delay-sensitive packets should be given higher priority in transmission. In this chapter, we study two channel allocation schemes that implement priority queues for priority packets in the GPRS network: Bitmap Channel Allocation (BCA)

and Uplink State Flag Channel Allocation (USFCA). Our study shows that the transmission delay of priority packets in the GPRS network can be better guaranteed using USFCA.

1.2 Uplink Connection Scheduling for Flat-Rate Data Services in the UMTS Network

To attract more users to mobile packet services, the Universal Mobile Telecommunication System (UMTS) operators have been prompting flat-rate packet services. Since usage does not incur cost, flat-rate users tend to stay on line longer and occupy most of the radio channel resources. We consider a UMTS network serving two types of user connections: Normal User Connections (NUCs) and Flat-Rate User Connections (FRUCs). Our goal is to maximize the revenue of the operator by giving a priority to NUCs over FRUCs without discontenting the flat-rate users, in order not to lose the flat-rate users to other operators. Uplink FRUCs may be asked to sub-rate or suspend transmission when the radio network is fully utilized. Four combinations of scheduling techniques including queueing, guard channels, preemption and rate-adaptation, have been studied, and analytic models using Markov processes were used to evaluate their performances. We proposed a cost function representing the revenue loss due to both blocked NUCs and lost flat-rate users. The system parameters used in our analysis are based on realistic operation data. Our analytic results indicate that the revenue loss can be minimized by using waiting queues and preemption. Rate-adaptation is ineffective in minimizing the revenue loss because sub-rated connections are less efficient in using radio spectrum. Guard channels for NUCs are unnecessary when waiting queue or preemption is used. Our study may be valuable for UMTS operators in serving flat-rate users.

1.3 Flate-Rate Packet Scheduling for the WCDMA Systems with HSDPA

To attract more users to use mobile packet services, mobile operators have begun to

provide flat-rate packet services in the WCDMA system with High Speed Downlink Packet Access (HSDPA). Since usage does not incur extra cost, flat-rate users may always stay on line and occupy most of the radio channel resources. In this chapter, we consider a HSDPA network serving two types of user packets: charged packets (CPs) and flat-rate packets (FRPs). Since CPs are charged by usage, they are given a higher priority to receive downlink packets for revenue consideration. However, this priority preference may lead to poor quality of service for flat rate users. In particular, FRPs may experience longer transmission latency and higher dropped probability. We should consider the balance between serving the FRPs and CPs. Four downlink packet scheduling methods are studied in this chapter: (1) Max. C/I first in a priority queue (M-PQ) ; (2) CPs first in a PQ (P-PQ) ; (3) Dynamic discard timer for FRPs in a PQ (DDT-PQ) and (4) Dynamic guard slots for CPs in a PQ (DGS-PQ). Analytic models using Markov process were used to study their performance. Our study shows that DDT-PQ and DGS-PQ methods are more effective to transmit the downlink CPs especially when the downlink FRP traffic is high. Therefore, they are better in guaranteeing the system throughput for CPs, and thus the operator revenue can be better protected.

1.4 Synopsis of This Dissertation

This dissertation is organized as follows. Chapter 2 presents Channel Allocation for Priority Packets in the GPRS Network. Chapter 3 presents Uplink Connection Scheduling for Flat-Rate Data Services in the UMTS Network. Chapter 4 presents Flate-Rate Packet Scheduling for the WCDMA Systems with HSDPA. Chapter 5 concludes this dissertation and describes the future work.

CHAPTER 2

Channel Allocation for Priority Packets in the GPRS Network

2.1 Introduction

General Packet Radio Service (GPRS) has been developed to provide packet data services based on the circuit-switching GSM network. Much research has been done on analyzing the performance of fixed or dynamic channel (i.e., timeslots) allocation to support multiple-slot data transmission [1-2]. However, very few studies considered special treatments to priority packets in the GPRS network. In [3], Chew and Tafazolli give priority to mobility management packets to ensure minimal delay. Their results indicated that the priority queue provides shorter RAU completion time and higher packet throughput than the others. However, the way in which the priority queue is implemented in the GPRS network has not been thoroughly studied.

In addition to the mobility management packets, some data services, such as "push to talk" (PTT) are delay-sensitive; the transmission latency of voice packets is very important to the quality of the communications. In this chapter, we study two channel allocation schemes [4], Bitmap Channel Allocation (BCA) and Uplink State Flag Channel Allocation (USFCA), that implement priority queues to give transmission priority to packets requiring shorter transmission latency. We also present analytic models to analyze their performance in terms of packet transmission delay.

2.2 The Methods of BCA and USFCA

A GSM/GPRS TDMA frame consists of eight timeslots, numbered 0-7, which can be used for data or voice transmission. Channel allocation in the GPRS network can be performed in unit of radio blocks. A radio block consists of four identical timeslots from four successive TDMA frames. Uplink packet requests from a mobile station (MS) can specify different priorities for special treatment by the GPRS network [4]. In this chapter, we assume only two types of packets: priority packets that are sensitive to delay, and non-priority packets that are not.

2.2.1 BCA Method

For an uplink "Packet Channel Request" message from a MS, the GPRS network may return a "Packet Uplink Assignment" message with the allocation_bitmap element indicating the allocated radio blocks to the uplink packet request. To reduce the number of messages exchanged between the MS and the network, the network allocates radio blocks in full amount requested by the MS. As a result, when all timeslots of the network are assigned out, new uplink packet requests need to wait until a transmitting packet completes. The transmitting packets cannot be interrupted during transmission.

2.2.2 USFCA Method

For an uplink "Packet Channel Request" message from a MS, the GPRS network may return a "Packet Uplink Assignment" message with a USF_for_each_timeslot_number element indicating a specific USF value for each timeslot allocated to the uplink packet request. For USFCA, the network broadcasts a USF value at each downlink radio block. In the next uplink radio block, the MS assigned with the same USF value can transmit for one radio block. In this way, the network can schedule an uplink packet to transmit at the next radio block on a radio block by radio block basis. As a result, a transmitting packet can be suspended at the end of a radio

block. The way in which multiple packets shares a timeslot is controlled the network; the network can use various scheduling schemes, such as priority-packet-first. Fig. 2.1 shows a USFCA example using priority-packet-first scheme, a non-priority packet 1 is assigned with USF value 1 and a priority packet 2, which needs m radio blocks to transmit, is assigned with USF value 2 by network. Packet 1 is transmitting when packet 2 arrives at radio block n. The network suspends the transmission of packet 1, and instructs packet 2 to transmit at downlink radio block n+1. Packet 1 can resume transmission after packet 2 completes transmission.

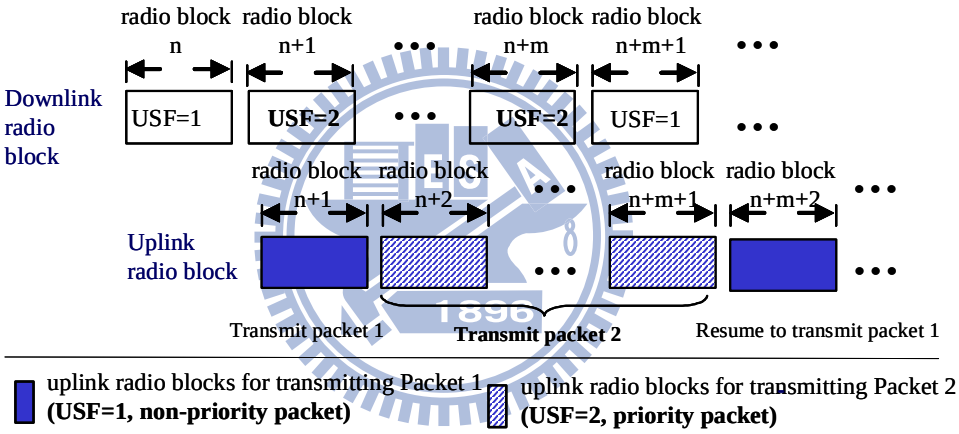


Fig. 2.1: A USFCA example using priority-packet-first scheme

2.3 The Analytic Models

Let *C* denote the number of GPRS timeslots reserved for transmission of data packets. When all the GPRS timeslots are assigned, additional uplink packet requests are put in a priority queue of size *B* maintained by the network. In the priority queue, packets of the same priority will be served on a FCFS basis. The queuing model of BCA and USFCA schemes is depicted in Fig. 2.2. Using BCA, the network cannot suspend the transmission of a packet under service, but using USFCA, the network can suspend the transmission of a non-priority

packet, put it back to the priority queue, and start transmitting a new priority packet. This difference is depicted in Fig. 2.2 by dotted line e.

To analyze the performance of the schemes, we made the following assumptions. The arrivals of priority and non-priority packets form Poisson processes with mean λ_P and λ_{nP} respectively. The service time of priority and non-priority packets is assumed to be exponentially distributed with mean $1/\mu_P$ and $1/\mu_{nP}$ respectively. We can use the M/M/*C*/*B* Markov process to model BCA and USFCA.

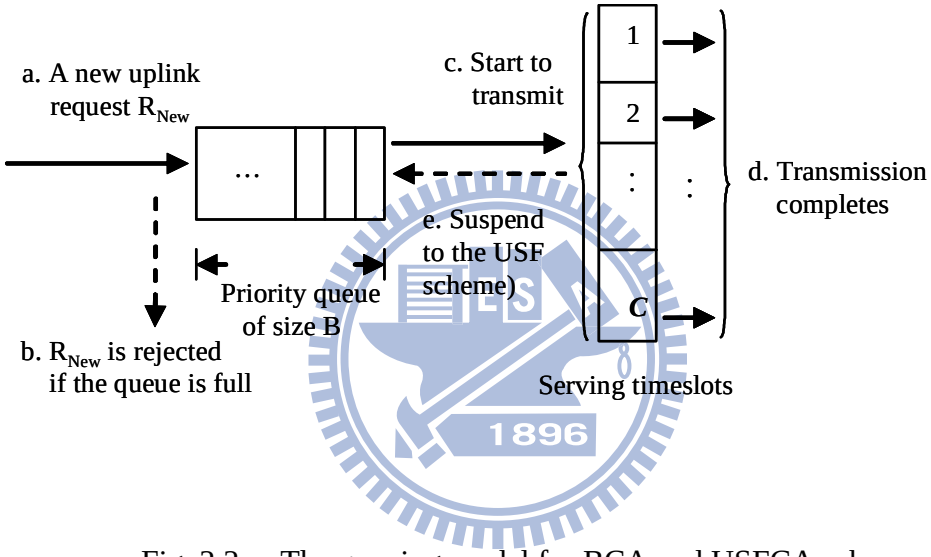


Fig. 2.2: The queuing model for BCA and USFCA schemes

2.3.1 BCA Method

In this process, state (*i*,*j*,*k*) denotes that there are *i* priority packets transmitting in the network, *j* priority packets waiting in the priority queue, *k* non-priority packets transmitting in the network or waiting in the priority queue. Let $P_{i,j,k}$ denote the steady-state probability of the network in state (*i*,*j*,*k*) and S_{bm} be the set of existing states for this process.

$$S_{bm}=\{(i,j,k)\mid 0\le i+j+k\le (C+B),0\le i\le C,0 \le j\le B,0\le k\le (C+B)\ \text{and}((i+k)\ge C\ \text{or}(j=0))\} \quad (2.1)$$

To handle the non-existing states, an indicator $\theta_{i,j,k}$ is used to indicate whether state (i,j,k) exists or not, i.e., $\theta_{i,j,k}=1$ if state (i,j,k) belongs to S_{bm}. In addition, δ_1 - δ_6 indicators are used to indicate whether a specific transition exists or not. The balance equations for this process can be expressed in (2.2) and the parameters are defined in (2.3)-(2.10).

$$\begin{aligned}
&P_{i,j,k}\Big[\delta_1\lambda_p\theta_{i+1,j,k}+\delta_2\left(\lambda_p\theta_{i,j+1,k}+M_{np}(i,j,k)\theta_{i+1,j-1,k-1}\right)\\
&\quad+\delta_3 M_p(i,j,k)\theta_{i-1,j,k}+\delta_4 M_p(i,j,k)\theta_{i,j-1,k}+\delta_5 M_{np}(i,j,k)\theta_{i,j,k-1}+\lambda_{np}\theta_{i,j,k+1}\Big]\\
&=\delta_2\left(P_{i,j-1,k}\lambda_p\theta_{i,j-1,k}+P_{i-1,j+1,k+1}M_{np}(i-1,j+1,k+1)\theta_{i-1,j+1,k+1}\right)+\delta_3 P_{i+1,j,k}M_p(i+1,j,k)\theta_{i+1,j,k}\\
&\quad+\delta_4 P_{i,j+1,k}M_p(i,j+1,k)\theta_{i,j+1,k}+\delta_5 P_{i,j,k+1}M_{np}(i,j,k+1)\theta_{i,j,k+1}+\delta_6 P_{i-1,j,k}\lambda_p\theta_{i-1,j,k}\\
&\quad+P_{i,j,k-1}\lambda_{np}\theta_{i,j,k-1}
\end{aligned} \tag{2.2}$$

$$M_p(l,m,n)=l\,\mu_p \tag{2.3}$$

$$M_{np}(l,m,n)=\begin{cases}(C-l)\,\mu_{np} & \text{if } (n\geq(C-l))\\ n\,\mu_{np}, & \text{otherwise}\end{cases} \tag{2.4}$$

δ_1=1, if $(i+k) < C$; 0, otherwise. (2.5)

δ_2=1, if $(i+k)\geq C$; 0, otherwise. (2.6)

δ_3=1, if $(j==0)$; 0, otherwise. (2.7)

δ_4=1, if $((i+k)\geq C)$ and $(i\neq 0)$; 0, otherwise. (2.8)

δ_5=1, if $(j==0)$ and $(i\neq C)$; 0, otherwise. (2.9)

δ_6=1, if $(i+j+k)\leq C$; 0, otherwise. (2.10)

From the balance equations and the constraints $\sum_{(i,j,k)\in S_{bm}} P_{i,j,k}$ = 1, the steady-state probability $P_{i,j,k}$ can be obtained by an iterative algorithm [5]. The blocking probability of packets ($P_{\_bm}$) ; the mean waiting time and system time of priority packets (W_{p_bm} and T_{p_bm}) ; the mean waiting time and system time of non-priority packets (W_{np_bm} and T_{np_bm}) can be expressed in (2.11)-(2.15).

$$P_{\_bm} = \sum_{(i+j+k=C+B)} P_{i,j,k} \tag{2.11}$$

$$W_{p\_bm} = \frac{1}{\lambda_p (1 - P_{\_bm})} \cdot \sum_{(i,j,k) \in S_{bm}} j \cdot P_{i,j,k} \tag{2.12}$$

$$T_{p\_bm} = \frac{1}{\lambda_p (1 - P_{\_bm})} \cdot \sum_{(i,j,k) \in S_{bm}} (i + j) \cdot P_{i,j,k} \tag{2.13}$$

$$W_{np\_bm} = \frac{1}{\lambda_{np} (1 - P_{\_bm})} \cdot \sum_{(i,j,k) \in S_{bm}, k > (c-i)} [k - (C - i)] \cdot P_{i,j,k} \tag{2.14}$$

$$T_{np\_bm} = \frac{1}{\lambda_{np} (1 - P_{\_bm})} \cdot \sum_{(i,j,k) \in S_{bm}} k \cdot P_{i,j,k} \tag{2.15}$$

2.3.2 USFCA Method

In this process, state (i,j) denotes that there are i priority packets and j non-priority packets transmitting in the network or in the priority queue. Let $P_{i,j}$ denote the steady-state probability of the network in state (i,j) and S_{USF} be the set of existing states for this process. x

$$S_{USF} = \{(i,j) \mid 0 \le i + j \le C + B, 0 \le i \le C + B, \text{and } 0 \le j \le C + B\} \tag{2.16}$$

To handle the un-eisting states, an indicator $\theta_{i,j}$ is used to indicate whether state (i,j) exists or not, i.e., $\theta_{i,j}$=1 if state (i,j) belongs to S_{USF} .The balance equations for this process can be expressed in (2.17) and the parameters are defined in (2.18)-(2.19).

$$\begin{aligned} &P_{i,j}(\lambda_p \theta_{i+1,j} + \lambda_{np} \theta_{i,j+1} + M_p(i,j)\theta_{i-1,j} + M_{np}(i,j)\theta_{i,j-1}) \\ &= P_{i-1,j}\lambda_p \theta_{i-1,j} + P_{i,j-1}\lambda_{np}\theta_{i,j-1} + P_{i+1,j}M_p(i+1,j)\theta_{i+1,j} + P_{i,j+1}M_{np}(i,j+1)\theta_{i,j+1} \end{aligned} \tag{2.17}$$

$$M_p(m,n) = \begin{cases} C\mu_p, & \text{if } (m \ge C) \\ m\mu_p, & \text{otherwise} \end{cases} \tag{2.18}$$

$$M_{np}(m,n) = \begin{cases} 0, & \text{if } (m \ge C) \\ (C - m)\mu_{np}, & \text{if } ((m+n) \ge C) \\ n\mu_{np}, & \text{otherwise} \end{cases} \tag{2.19}$$

From the balance equations and the constraints $\sum_{(i,j) \in S_{USF}} P_{i,j} = 1$, the steady-state probability $P_{i,j}$ can be derived by an iterative algorithm. The blocking probability of

packets ($P_{\_USF}$) ; the mean waiting time and system time of priority packets (W_{p_USF} and T_{p_USF}) ; the mean waiting time and system time of non-priority packets (W_{np_USF} and T_{np_USF}) can be expressed in (2.20)-(2.24).

$$P_{\_USF} = \sum_{(i+j=C+B)} P_{i,j} \tag{2.20}$$

$$W_{p\_USF} = \frac{1}{\lambda_p (1 - P_{\_USF})} \cdot \sum_{(i,j)\in S_{USF},(i>C)} (i - C) \cdot P_{i,j} \tag{2.21}$$

$$T_{p\_USF} = \frac{1}{\lambda_p (1 - P_{\_USF})} \cdot \sum_{(i,j)\in S_{USF}} i \cdot P_{i,j} \tag{2.22}$$

$$W_{np\_USF} = \frac{1}{\lambda_{np} (1 - P_{\_USF})} \cdot \left[\sum_{(i,j)\in S_{USF},(i\geq C)} j \cdot P_{i,j} + \sum_{(i,j)\in S_{USF},(i<cC),(i+j>C)} (i + j - C) \cdot P_{i,j} \right] \tag{2.23}$$

$$T_{np\_USF} = \frac{1}{\lambda_{np} (1 - P_{\_USF})} \cdot \sum_{(i,j)\in S_{USF}} j \cdot P_{i,j} \tag{2.24}$$

2.4 Numeric Results

The total number of data channels (C) is set to be 4 and the queue size (B) is set to be 4. We compare three channel allocation schemes. The first two are BCA and USFCA schemes described in the previous section. The third one is a simple FCFS channel allocation scheme with a FIFO queue of the same size (B). The simple FCFS scheme can also be modeled as a M/M/C/B Markov process, let $W_{\_FCFS}$ and $T_{\_FCFS}$ denote the mean waiting time and system time of packets.

The mean service time of one packet ($1/\mu_P$ and $1/\mu_{nP}$) is assumed to be 0.0625 seconds with one timeslot allocated. This represents approximately an average 105 bytes per packet under the GPRS CS-2 coding scheme and is near the average uplink packet sizes. For packet arrival, λ_{nP} is fixed at 32 packets/second and λ_P varies in the range of 8-32 packets/second.

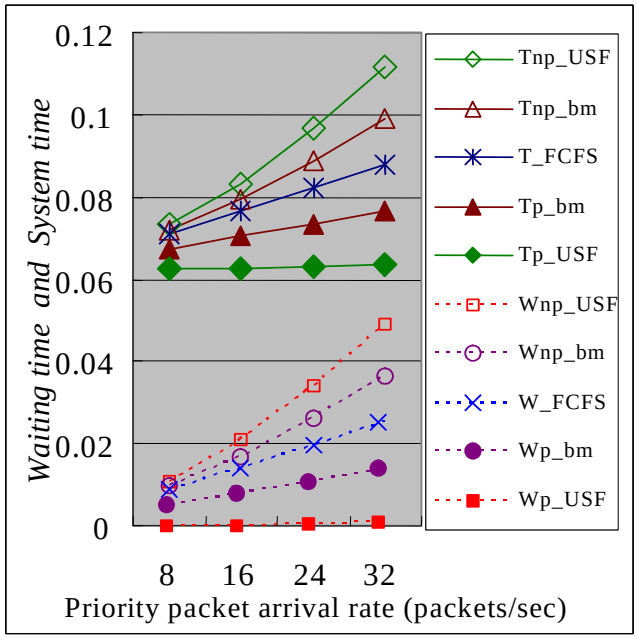


Fig. 2.3: The mean waiting time and system time of uplink packets for λ_{nP} = 32 packets/second and λ_P = 8-32 packets/second

In Fig. 2.3, the results indicate both BCA and USFCA schemes provide shorter mean waiting time and system time for priority packets than the simple FCFS scheme at the cost of longer mean waiting time and system time for non-priority packets. The improvement and the cost become more significant as the priority traffic increases. In addition, the improvement and the cost of USFCA scheme are more significant than those of BCA scheme. This is because when there is no free channel, USFCA scheme can suspend the transmission of a non-priority packet and start transmitting a new priority packet, but BCA scheme cannot. The improvement on mean waiting time and system time for priority packets over the simple FCFS scheme can be as large as 0.025 seconds when the priority packet arrival rate is the 32 packets/sec, the transmission delay can be greatly reduced to an extend of nearly 72%. This 0.025 seconds difference could be critical for real-time voice communications.

2.5 Conclusions

This chapter, we studied BCA and USFCA schemes that implement priority queues in the GPRS network. Both schemes provide shorter mean waiting time and system time for priority packets than the simple FCFS scheme at the cost of longer mean waiting time and system time for non-priority packets. In addition, the transmission delay of priority packets using USFCA can be better guaranteed than that of BCA especially when the GPRS traffic is heavy.

CHAPTER 3

Uplink Connection Scheduling for Flat-Rate Data Services in the UMTS Network

3.1 Introduction

The Universal Mobile Telecommunication System (UMTS) using Wideband Code Division Multiple Access (WCDMA) radio technology represents an evolution in terms of capacity, data rates and service capabilities, from the GSM/General Packet Radio Service (GPRS) network [6]. It is an integrated solution for mobile voice and data with wide area coverage and high data rates. The UMTS network can provide packet data rates up to 384 kbps in high mobility situations, and as high as 2 Mbps for stationary users. The packet data usage of current UMTS users is not popular because of the lack of popular mobile data applications and the high cost of data transmission. To attract more packet data users, UMTS operators have begun to provide flat-rate packet services. Flat-rate users pay fixed monthly charge for un-limited data packet transmission. Since usage incurs no extra charge, flat-rate users tend to keep data connections alive longer, and occupy most of the network resources. Without special treatments for different classes of user connections, normal users who are charged by usage may be blocked from accessing the UMTS network.

Since blocked normal user connections result in revenue loss of the UMTS operator, to increase the revenue, normal users should be given priority on transmission. On the other hand, if flat-rate users experience blocked connections frequently, the discontent flat-rate users may switch to other service providers. The loss of flat-rate users leads to revenue loss

too. Therefore, a balance needs to be found in allocating radio resources to normal users and flat-rate users. In this chapter, we propose a cost function representing the revenue loss due to both blocked normal users and lost flat-rate users. Since the revenue loss on both situations depends on the blocking probabilities, we investigate scheduling techniques, including queueing, guard channels, preemption and rate-adaptation, to keep the blocking probabilities of normal users and flat-rate users at different levels, and to minimize the cost function, i.e., the revenue loss. We consider the aforementioned four scheduling techniques, because they have been repeatedly used in giving transmission priority in mobile networks. However, no one has investigated the effectiveness of these four scheduling techniques in maximizing the revenue of UMTS operators serving flat-rate and normal users.

Much research has been done on the mobile network in giving transmission priority to a certain type of service. In mobile networks, terminating a handoff call is considered a higher cost than blocking a new call. When a handoff call arrives, but there is no free channel in the cell, the handoff call can be placed in a queue and handoff is delayed until free channels become available [7]. To further give a priority to handoff calls, a small number of free channels called guard channels can be reserved for handoff calls. Guard channels significantly reduce the forced termination probability of handoff calls at the cost of blocking more new calls and reducing the system throughput [8]. To increase the total carried traffic and improve the perceived service quality, Guerin put originating calls in a queue when the network has very few free resources [9]. Zeng, et al, also proposed that both the new and handoff calls can be queued, and showed that the forced termination probability of handoff calls decreased drastically with only a small increase in the blocking probability of new calls [10]. For integrated voice and data communications, Zeng, et al, presented a system with two queues for handoff calls, one for voice and the other for data. Their results showed that the forced termination probability of voice handoff calls and the average transmission delay of data connections decreased by increasing the size of handoff queues [11]. Leong, et al, presented a

system with two buffers for data calls, one for new data call and the other for handoff. Their results indicated that the Quality of Service (QoS) can be guaranteed for both voice and data services in a multi-cell environment [12].

Preempting a low priority call to free radio resources for high priority calls is another effective way to ensure transmission priority. However, this approach usually preempts data calls only, because cutting off voice communications can be very annoying to the users. High priority of real-time traffic, such as voice and video, can preempt non-real-time traffic (data). Several researchers have shown that the preemption of non-real-time data can guarantee QoS for real-time classes, and achieve high channel utilization [13-14]. Kim, et al, proposed that high priority voice calls can preempt low priority voice calls. Voice calls that have low SIR and long duration are considered low priority calls, which can be preempted to improve the entire network performance [15].

Sub-rating current calls to free radio resources for new or handoff calls is another way to reduce blocking probabilities. A serving full-rate channel can be temporarily divided into two half-rate channels when the network is fully utilized; one to serve the existing call and the other to serve the handoff call [16]. Chen, et al, studies GPRS networks where a data session can occupy more than one GPRS data channel. When there are no free channels upon the arrival of a voice call, one slot of an existing multi-slot GPRS data session is de-allocated for the new voice arrival [17]. Their results show the voice blocking probability can be greatly reduced, especially at high GPRS traffic load.

Most of the researches focus on reducing the blocking and forced termination probabilities of high-priority connections. However, very few studies have been done on maximizing the operator revenue for mobile networks serving flat-rate users, as well as normal users. In this chapter, we investigate combinations of the scheduling techniques aforementioned to maximize the operator revenue. We propose a cost function that represents the revenue loss of service providers providing both flat-rate and per-packet charging services.

An iterative algorithm has been developed to determine the optimal number of guard channels and the best combination of scheduling techniques in minimizing the revenue loss. Our study may be valuable for UMTS operators in serving flat-rate users. The notations we use in this chapter are listed in Table 3.1.

Table 3.1 The usage of the system notations in uplink connection scheduling for Flat-Rate Users

Notation	Meaning
B	The size of the NUC waiting queue
C	The cost function
C_f	The monthly revenue loss due to lost flat rate users
C_{min}	The minimum value of the cost function
C_n	The monthly revenue loss due to blocked NUCs
G	The number of guard channels
G_{opt}	The optimum number of guard channels
L_{QN}	The average NUC queue lengths
L_{QF}	The average FRUC queue length
N_F	The number of full-rate connections
$N_F^*(y)$	The maximum number of full-rate connections when there are y half-rate connections

$N_{FF}(i,j,k)$	The number of full-rate FRUCs in state (i, j, k)
N_H	The number of half-rate connections
$N_H^*(z)$	The maximum number of half-rate connections when there are z full-rate connections
$N_{HF}(i,j,k)$	The number of half-rate FRUCs in state (i, j, k)
$N_{HF+FF}(i,j,k)$	The total number of full-rate and half-rate FRUCs in state (i, j, k)
P_{BF}	The blocking probability of FRUCs
P_{BN}	The blocking probabilities of NUCs
P_F	The probability that the first events occurs to a serving half-rate FRUC is being full-rated
P_{FC}	The probability that the first events occurs to a serving full-rate FRUC is completion
P_{FP}	The probability that the first events occurs to a full-rate serving FRUC is being preempted
P_{FPrm}	The probability that a serving full-rate FRUC is preempted before its completion
P_{FS}	The probability that a serving full-rate FRUC is sub-rated before its completion or preemption
P_{FST}	$P_{FST} = 1 - P_{FC} - P_S - P_{FP}$
$P_{i,j,k}$	The stationary state probability of the network in state (i,j,k)
P_S	The probability that the first events occurs to a serving full-rate FRUC be being sub-rated
P_{SC}	The probability that the first events occurs to a

	serving half-rate FRUC is completion
P_{SP}	The probability that the first events occurs to a serving half-rate FRUC is being preempted
P_{SPrm}	The probability that a serving half-rate FRUC is preempted before its completion
P_{SST}	$P_{SST} = 1 - P_{SC} - P_F - P_{SP}$
P_{QF}	The queueing probability of FRUCs
P_{QN}	The queueing probability of NUCs
S_{All}	Scheduler with guard channels, waiting queues, rate adaptation, and preemption scheduler
S_G	the set of all existing transition states of SAll
S_{NPrm}	Scheduler without preemption
S_{NRA}	Scheduler without rate adaptation
S_{NWQ}	Scheduler without the NUC waiting queues
T_X	The average transmission rate of serving FRUCs
Q	The size of the FRUC waiting queue
W_{TF}	The waiting time of FRUCs
W_{TN}	The waiting time of NUCs
α	The cost weighting factor of flat-rate users
α_F	The activity factor of full-rate connections
α_H	The activity factor of half-rate connections

β	The departure threshold of FRUC blocking probability
δ_F	The nominal capacity of a full-rate connection
δ_H	The nominal capacity of a half-rate connection
λ_f	The arrival rate of FRUCs
λ_n	The arrival rate of NUCs
$1/\mu_f$	The average service time of full-rate FRUCs
$1/\mu_n$	The average service time of NUCs
$\theta_{i,j,k}$	An indicator to indicate whether state (i,j,k) exists or not
ρ_n	The traffic load of NUCs
ζ	The inter-cell interference factor for a cell
$\Omega(N_F, N_H)$	The total transmission power received by the RNC in a cell

3.2 System Models and Assumptions

A UMTS network consists of three interacting domains: Core Network (CN), UMTS Terrestrial Radio Access Network (UTRAN) and mobile stations (MS). The UTRAN provides the air interface access method for MSs [18]. A Base Station is referred to as Node B; the control node for a group of Node Bs is called a Radio Network Controller (RNC). Wideband CDMA technology was selected to be the air interface of the UTRAN. To be specific, we study the Frequency Division Duplex (FDD) WCDMA operation in this chapter.

A RNC can allocate a physical dedicated radio channel (D_CH) to an MS by through a RAB assignment procedure [18-20]. Fig. 3.1 depicts the message flow of a D_CH assignment procedure. In Step 1, an MS establishes a Radio Resource Control (RRC) connection with the

RNC before creating a Packet Data Protocol (PDP) context between the MS and the GGSN. In Step 2, the MS sends an “Activate PDP Context Request” message to the SGSN with a QoS element indicating service class (conversional, streaming, interactive or background data). In Step 4, the SGSN sends a “RAB assignment request” message with RAB parameters, which will be described in more details later, to the RNC to establish a RAB connection between the MS and SGSN. After the D_CH is established in Step 5, the MS can start to transmit/receive packets to/from the CN in Step 6. When necessary, the RNC can instruct the MS that packet transmission of the connection should be stopped, continued or change the transmission rate on its assigned D_CH by a Radio Bearer (RB) reconfiguration procedure as indicated in Step 7. The MS should comply with the instructions. After the MS completes transmission, the RB and RRC of the MS can be released in Steps 8 and 9.

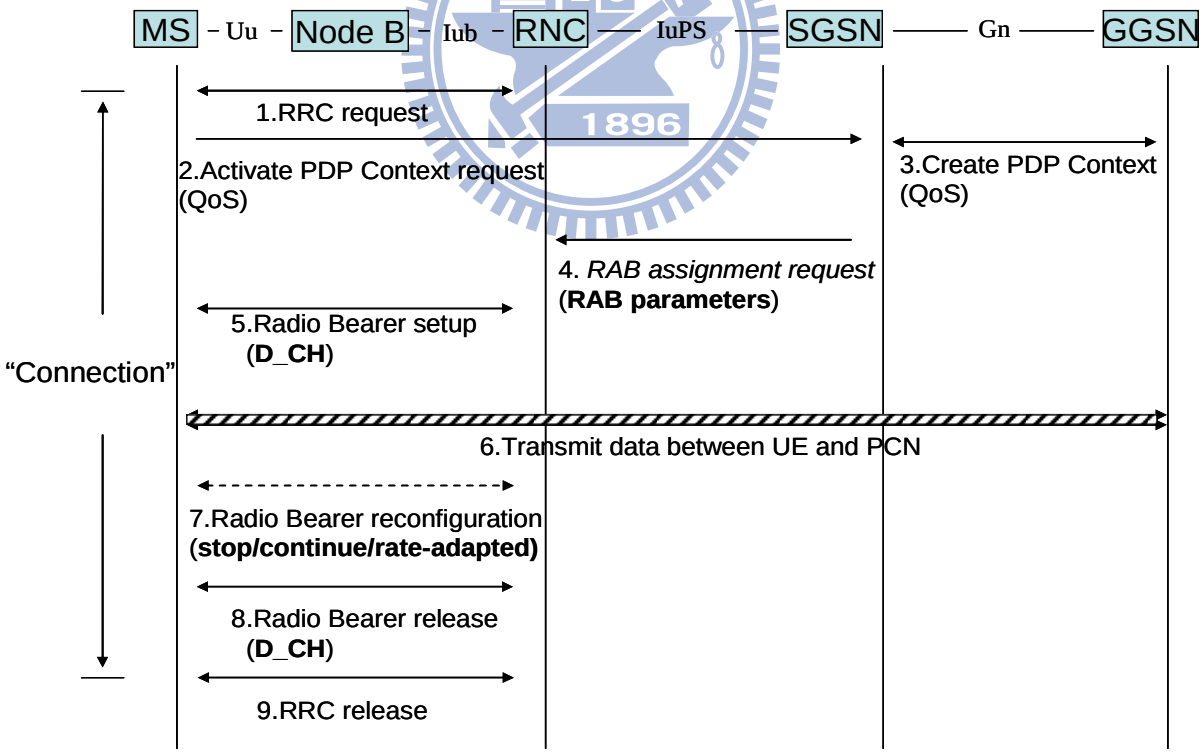


Fig. 3.1: The Radio Access Bearer (RAB) assignment procedure

The RAB parameters sent from the SGSN to the RNC in Step 4 can be used to instruct the RNC the scheduling policy of the data connection. The parameters include a Priority Level element, a Pre-emption Capability element indicating the capability to preempt lower priority RABs, a Pre-emption Vulnerability element indicating whether the DCH is vulnerable to be preempted or not, and a Queuing Allowed element indicating whether the RAB request can be queued. In addition, Maximum Bit Rate and Guaranteed Bit Rate elements indicate the transmission rate of MSs. These RAB parameters can be used to instruct the RNC how to schedule the packet transmission

A user connection starts at the establishment of a RRC between the MS and the RNC, as shown in Step 1, Fig. 3.2, and ends at the disconnection of the RNC. We assume there are two types of user connections in the UMTS network; Normal User Connections (NUCs), which are assigned a higher priority in transmission, and Flat-Rate User Connections (FRUCs), which may be sub-rated or suspended when the network traffic load is high. When a connection is sub-rated, its transmission rate and transmission power can be reduced, and thus transmission power allowance is released for other connections. Since NUCs are charged by the volume of packet transmission, a NUC tends to be shorter, such as sending an e-mail or uploading short files. On the other hand, FRUCs pay fixed monthly fee no matter how many packets they transmit, a FRUC is generally longer, such as playing on-line games and using peer-to-peer applications.

In UMTS R99 network, the uplink data transmission can only be scheduled on connection level, but not on packet level. This is because after DCHs are allocated to MSs, the MSs can start or pause data transmission anytime without notifying the RNC. However, the RNC can suspend or sub-rate the uplink connection as we have described. On the other hand, downlink data transmission can be scheduled on packet level, because all downlink packets are stored and forwarded by the RNC. The RNC can determine priorities in forwarding different classes of packets. As a result, uplink and downlink transmissions may require

different scheduling techniques. In this chapter, we consider the scheduling for uplink data connections only. We use the CDMA uplink soft-capacity model to estimate the uplink total bandwidth of a cell in system.

3.2.1 CDMA Uplink Capacity Model

The capacity of a CDMA network is not fixed; it has so-called "soft capacity". Since a FRUC can be sub-rated, we consider two transmission rates of data services from MSs, full-rate and half-rate data connections. We can obtain the limit on the total transmission power received by the Node B in a cell from Equation (1) [21]. α_F and α_H denote the activity factor of full-rate and half-rate data connections in a cell, respectively. N_F and N_H denote the numbers of MSs using full-rate and half-rate data connections, respectively. δ_F denotes the *nominal capacity* of a full-rate data connection, i.e., the portion of total transmission power received by the Node B in a cell; δ_H denotes that of a half-rate data connection [22]. ζ is the inter-cell interference factor for a cell which can be obtained from measurements [23].

$$\Omega(N_F, N_H) = \alpha_F \times N_F \times \delta_F + \alpha_H \times N_H \times \delta_H < \frac{1}{(1+\zeta)} \qquad (3.1)$$

From Equation (3.1), we can obtain the *Pole capacity* of MSs using full-rate and half-rate data connections in a cell as in Equation (3.2). $N_F^*(y)$ denotes the maximum number of full-rate serving MSs in a cell when there are *y* half-rate serving MSs and $N_H^*(z)$ denotes the maximum number of half-rate serving MSs in a cell when there are *z* full-rate serving MSs. In particular, $N_F^*(0)$ denotes the maximum number of full-rate serving MSs in a cell, and $N_H^*(0)$ half -rate serving MSs.

$$N_F^*(y)=max\left\{N_F \left| \frac{1}{(1+\zeta)} \geq (\alpha_F \times N_F \times \delta_F + \alpha_H \times y \times \delta_H) \right.\right\}$$
$$N_H^*(z)=max\left\{N_H \left| \frac{1}{(1+\zeta)} \geq (\alpha_F \times z \times \delta_F + \alpha_H \times N_H \times \delta_H) \right.\right\} \quad (3.2)$$

In the analysis below, we assume the spread spectrum bandwidth (*W*) of the WCDMA network is 5 MHz, the full-rate data transmission is 128kbps and the half-rate is 64kbps. The two data rates are the default uplink data rates provided in CHT UMTS R99 network. According to the 3GPP specification [24], full-rate and half-rate data transmissions have different Signal-Interference-Ratio (SIR) requirements to achieve Block Error Rate (BLER) $<10^{-2}$ in multipath fading conditions; for full-rate it is 8.4 dB and half-rate 9.2 dB. From the desired SIR, we can obtain the nominal capacity δ_F = 0.177and δ_H =0.106. The activity factor for data services (α_F and α_H) is assumed to be 0.5 in busy hour, and the inter-cell interference factor (ζ) is assumed to be 0.1. These assumptions follow those in [23].

From Equation (2), we can obtain $N_F^*(0)$ = 10 and $N_H^*(0)$ = 17. Note that $N_H^*(0)$ is less than twice of $N_F^*(0)$ because more number of MSs transmitting leads to more signal interference. In other words, half-rate transmission is less efficient in using radio bandwidth.

3.2.2 A Scheduler with all four features

The queueing model of the connection scheduler that implements waiting queues (WQ), guard channels (GC), preemption and rate-adaptation on the RNC is depicted in Fig. 3.2. There are two waiting queues; one for new NUCs, and the other for new and preempted FRUCs. When an on-going up-link connection is put in a waiting queue, the Node B instructs the MS to stop packet transmission. Since there is no packet transmission, no storage space on Node-B is needed for the up-link packets of a queued

connection. Guard channels of dynamic size are reserved for NUCs. The number of guard channels will be determined by an iterative algorithm described later to maximize the revenue. The connection scheduler works as follows. When a new NUC (line a) request arrives, it can be served immediately if the network is not fully utilized (line b). Otherwise, the RNC first tries to sub-rate serving FRUCs (line c) to accommodate the new NUC. If this is not possible, i.e., all serving FRUCs are sub-rated, the RNC preempts FRUCs into the waiting queue (dotted line d). If there is no serving FRUC, the new NUC is put into the NUC waiting queue; if the queue is full, it is rejected (dotted line e).

When a new FRUC (line f) request arrives, it can be served immediately if there are free channels other than the reserved guard channels (line g). Otherwise, serving full-rate FRUCs can be sub-rated (line c) to accommodate the new FRUC, if doing so satisfies the total power limit in Equation (3.2). Otherwise, the new FRUC request can be put into the FRUC WQ; if the queue is full, it is rejected (dotted line h).

When a serving connection finishes, it releases radio channels (line i). The free channel will serve a waiting NUC first. If there is no waiting NUC, waiting FRUCs will be served (line g). If there is no waiting FRUC, a serving half-rate FRUCs can resume full-rate transmission (line j).

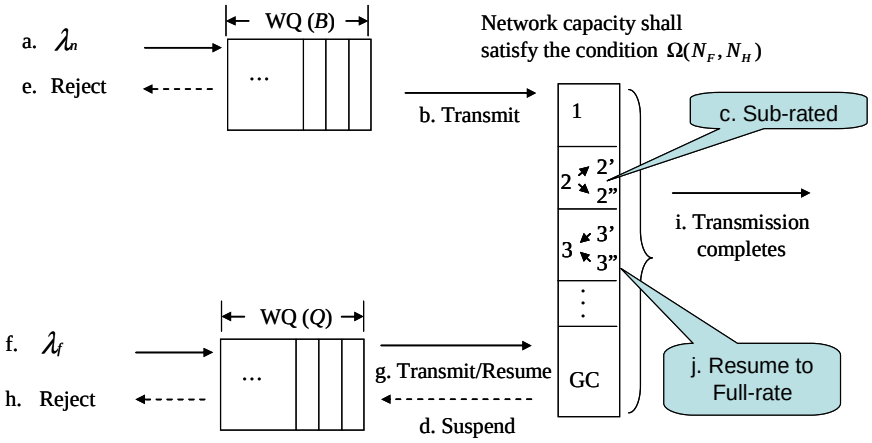


Fig. 3.2: The system queueing model for a reference cell

The scheduler implementing all four scheduling techniques described above will be referred to as S_{All}. To evaluate the effectiveness of waiting queues, rate adaptation, and preemption, we also analyzed three additional schedulers, each of which omits one scheduling technique. Let S_{NWQ} denote the one without a NUC waiting queue, S_{NRA} without rate adaptation, and S_{NPrm} without preemption. All the schedulers implement guard channels. Due to space limitation, the analytic models and the performance measure equations of S_{NRA}, S_{NWQ}, and S_{NPrm} are not presented in this chapter.

3.3 The Analytic Models

In We can use the M/M/*C*/*B* Markov process to analyze the connection schedulers. Let *B* denote the size of the NUC waiting queue, *G* the number of guard channels, *Q* the size of the FRUC waiting queue. The new arrivals of NUCs and FRUCs were assumed to form Poisson processes with rates λ_n and λ_f, respectively. The service times of NUCs and full-rate FRUCs were assumed to be exponentially distributed with mean $1/\mu_n$ and $1/\mu_f$, respectively. The assumption of Poisson arrivals can provide a good approximation when the user population is large; the assumption of exponential service time facilitates the analysis.

3.3.1 The Analytic Model of S_{All}

The analytic model for the scheduler with all features (S_{All}) is described as follows. Let state (*i*,*j*,*k*) denote that there are *i* transmitting NUCs, *j* waiting NUCs, and *k* FRUCs transmitting or waiting. The exact numbers of full-rate and half-rate transmitting FRUCs can be determined by an algorithm described in Fig. 3.3. Let $N_{FF}(i,j,k)$ denotes the number of full-rate FRUCs in state (*i*, *j*, *k*), $N_{HF}(i,j,k)$ that of half-rate, and $N_{FF+HF}(i,j,k)$ denotes the total number of full-rate and half-rate FRUCs in state (*i*, *j*, *k*).

---

if [i >=($N_F^*(0)$-G)],

all k FRUCs are queued, $N_{FF}(i,j,k)$=0 and $N_{HF}(i,j,k)$=0

else if [(i+k) < ($N_F^*(0)$-G)], all FRUCs are served in full rate,

$N_{FF}(i,j,k)$=k and $N_{HF}(i,j,k)$=0

else if there exists a minimal h, such that [(i+k) < ($N_F^*(h)$-G+h)],

$N_{FF}(i,j,k)$=k-h and $N_{HF}(i,j,k)$=h

else if $N_{FF}(i,j,k)$= 0, $N_{HF}(i,j,k)$= $N_H^*(i+G)$, and (k- $N_H^*(i+G)$)

FRUCs are queued

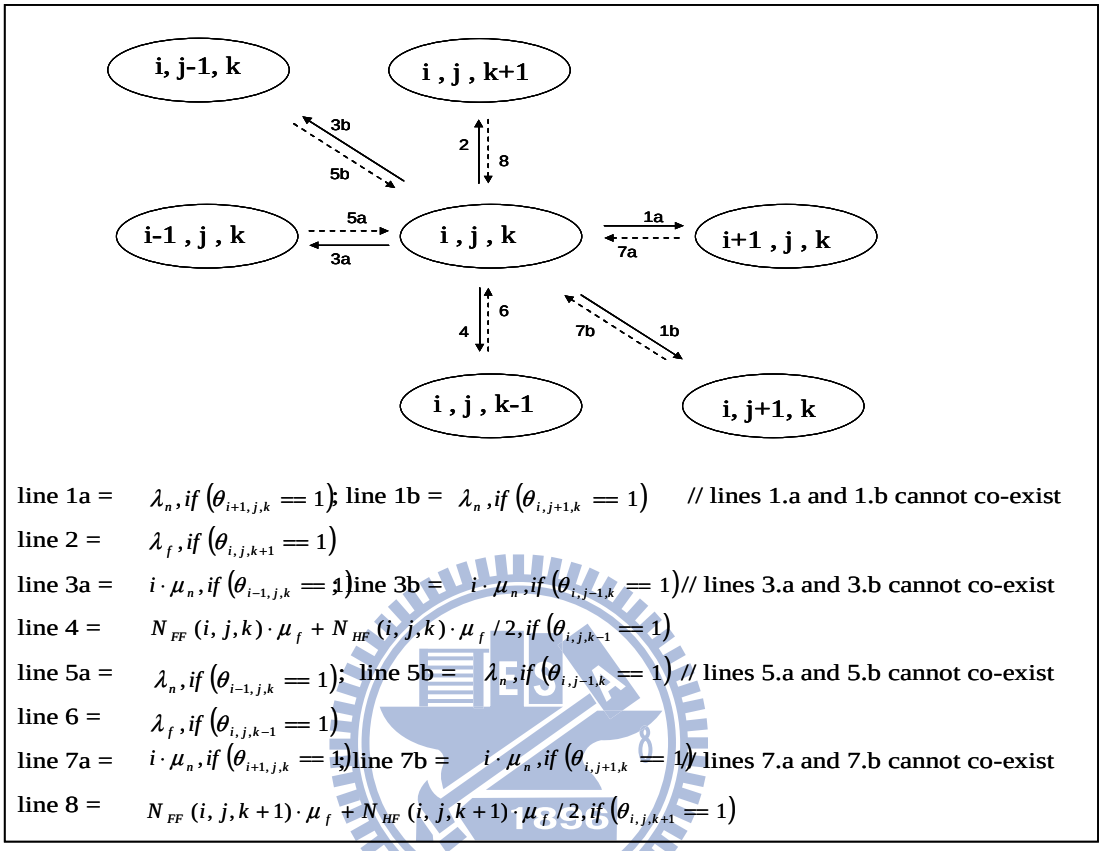
Fig. 3.3: An algorithm determines the numbers of full-rate and half-rate FRUCs in state (i, j, k)

Let S_G be the set of all existing transition states of S_{All}. For each existing state (i ,j, k), the number of serving NUCs cannot be more than $N_F^*(0)$, the number of queued NUCs cannot be more than B, and the number of FRUCs cannot be more than $N_{FF+HF}(i,j,k)$+Q. S_G can be expressed as in Equation (3.3). The maximum size of S_G should be limited to [$N_F^*(0)$+1] * [B +1] * [$N_H^*(0)$+Q+1].

$$S_G=\{(i,j,k)\,|\,[0\le i\le N_F^*(0),\ \mathrm{j}=0,0\le \mathrm{k}\le N_H^*(i)+\mathrm{Q}],\ \mathrm{or}[i=N_F^*(0),0<j\le B,0\le \mathrm{k}\le Q]\} \qquad (3.3)$$

Part of the state transition diagram is depicted in Fig. 3.4. To handle the non-existing states, an indicator, $\theta_{i,j,k}$, is used to indicate whether state (i,j,k) exists or not. $\theta_{i,j,k}$ =1 if state (i,j,k) belongs to ; otherwise, $\theta_{i,j,k}$ =0. Let $P_{i,j,k}$ denote the steady-state probability of the network in state (i,j,k). For existing state (i,j,k), the output flows (lines 1-4), its input flows from other states (dotted lines 5-8), and the transition

rate of each line is depicted in Fig. 3.4.



line 1a = $\lambda_n, if\ (\theta_{i+1,j,k} == 1)$; line 1b = $\lambda_n, if\ (\theta_{i,j+1,k} == 1)$ // lines 1.a and 1.b cannot co-exist

line 2 = $\lambda_f, if\ (\theta_{i,j,k+1} == 1)$

line 3a = $i \cdot \mu_n, if\ (\theta_{i-1,j,k} == 1)$; line 3b = $i \cdot \mu_n, if\ (\theta_{i,j-1,k} == 1)$ // lines 3.a and 3.b cannot co-exist

line 4 = $N_{FF}(i,j,k) \cdot \mu_f + N_{HF}(i,j,k) \cdot \mu_f / 2, if\ (\theta_{i,j,k-1} == 1)$

line 5a = $\lambda_n, if\ (\theta_{i-1,j,k} == 1)$; line 5b = $\lambda_n, if\ (\theta_{i,j-1,k} == 1)$ // lines 5.a and 5.b cannot co-exist

line 6 = $\lambda_f, if\ (\theta_{i,j,k-1} == 1)$

line 7a = $i \cdot \mu_n, if\ (\theta_{i+1,j,k} == 1)$; line 7b = $i \cdot \mu_n, if\ (\theta_{i,j+1,k} == 1)$ // lines 7.a and 7.b cannot co-exist

line 8 = $N_{FF}(i,j,k+1) \cdot \mu_f + N_{HF}(i,j,k+1) \cdot \mu_f / 2, if\ (\theta_{i,j,k+1} == 1)$

Fig. 3.4: The state transition diagram of S_{All}, and the rates of input/output flows

The rate of input flows of state (i,j,k) can be expressed in (3.4), and that of output flows in (3.5). When the system is in equilibrium, the rates are equal; the system equilibrium equation of state (i,j,k) can be expressed in (3.6).

Inflow(i,j,k)=

$$\begin{cases} \lambda_n \cdot P_{i-1,j,k}\theta_{i-1,j,k} + \lambda_n \cdot P_{i,j-1,k}\theta_{i,j-1,k} + \lambda_f \cdot P_{i,j,k-1}\theta_{i,j,k-1} \\ +(i+1)\mu_n \cdot P_{i+1,j,k}\theta_{i+1,j,k} + i\mu_n \cdot P_{i,j+1,k}\theta_{i,j+1,k} \\ +(N_{FF}(i,j,k+1)\cdot \mu_n + (N_{HF}(i,j,k+1)\cdot \mu_n / 2)) \cdot P_{i,j,k+1}\theta_{i,j,k+1} \end{cases} \quad (3.4)$$

Outflow(i,j,k) =

$$\begin{cases} \lambda_n \theta_{i+1,j,k} + \lambda_n \cdot \theta_{i,j+1,k} + \lambda_f \cdot \theta_{i,j,k+1} \\ + i\mu_n \cdot \theta_{i-1,j,k} + i\mu_n \cdot \theta_{i,j-1,k} \\ + \left(N_{FF}(i,j,k) \cdot \mu_n + \left(N_{HF}(i,j,k) \cdot \mu_n / 2\right)\right) \cdot \theta_{i,j,k-1} \end{cases} \quad (3.5)$$

$P_{i,j,k}$ = Inflow(i,j,k)/Outflow(i,j,k); (3.6)

The From Equations (3.3)-(3.6) and the constraint $\sum_{(i,j,k) \in S_G} P_{i,j,k} = 1$, we can use the iterative algorithm proposed in [26] to obtain the stationary state probabilities $P_{i,j,k}$. The iterative algorithm of our system will be described in more details later. It is possible to extend the system model for an arbitrary number of rates, if given the numbers of NUCs and FRUCs, the scheduling scheme can determine the numbers of connections served at each rate.

3.3.2 The Performance Measures

Step The performance measures we considered are described as follows. A NUC or FRUC is blocked when the WQ is full. Based on the Poisson Arrivals See Time Averages (PASTA) property [25], the blocking probabilities of NUCs (P_{BN}) or FRUCs (P_{BF}) can be expressed in (3.7)-(3.8).

$$P_{BN} = \sum_{(i,j,k) \in S_G, (j==B)} P_{i,j,k} \quad (3.7)$$

$$P_{BF} = \sum_{(i,j,k) \in S_G, (k - N_{HF+FF}(i,j,k) == Q)} P_{i,j,k} \quad (3.8)$$

From the stationary state probabilities, we can obtain the average queue lengths of the NUC WQ (L_{QN}) and FRUC WQ (L_{QF}); they can be expressed in (3.9)-(3.10),

respectively.

$$L_{QN} = \sum_{(i,j,k)\in S_G} j \cdot P_{i,j,k} \tag{3.9}$$

$$L_{QF} = \sum_{(i,j,k)\in S_G} \left[k - N_{HF+FF}(i,j,k)\right] \cdot P_{i,j,k} \tag{3.10}$$

The average waiting times of NUCs (W_{TN}) and FRUCs (W_{TF}) in the waiting queue can be obtained using Little's formula; they can be expressed in (3.11)-(3.12), respectively.

$$W_{TN} = \frac{L_{QN}}{\lambda_n \cdot (1 - P_{BN})} \tag{3.11}$$

$$W_{TF} = \frac{L_{QF}}{\lambda_f \cdot (1 - P_{BF})} \tag{3.12}$$

The queueing probability of a connection is defined to be the probability that a connection cannot be served immediately upon its arrival. A new NUC is put into the waiting queue when the network is fully utilized by NUCs. The queueing probability of NUCs (P_{QN}) can be expressed in (3.13). A new FRUC is put into waiting queue when all the serving FRUCs are sub-rated, or sub-rating full-rate serving FRUCs cannot release enough bandwidth for the new FRUC to transmit in half-rate. The queueing probability of FRUCs (P_{QF}) can be expressed in (3.14).

$$P_{QN} = \sum_{(i,j,k)\in S_G,\, i=N_F^*(0),\, (j<B)} P_{i,j,k} \tag{3.13}$$

$$P_{QF} = \sum_{\substack{(i,j,k)\in S_G,\, 0<(k-N_{HF+FF}(i,j,k))<Q, \\ N_{HF+FF}(i,j,k)=N_{HF+FF}(i,j,k+1)}} P_{i,j,k} \tag{3.14}$$

To obtain the probability that a full-rate serving FRUC is sub-rated, consider the first event that occurs to a full-rate serving FRUC. The FRUC may complete (with

probability P_{FC}), be sub-rated (with probability P_S), be preempted (with probability P_{FP}), or none of the aforementioned events occurs but a state transition occurs (with probability P_{FST}). The probabilities can be expressed in (3.15)-(3.18).

$$P_{FC} = \sum_{(i,j,k)\in S_G} P_{i,j,k} \cdot \frac{\mu_a}{i\mu_a + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_a + \lambda_f} \tag{3.15}$$

$$P_S = \sum_{\substack{(i,j,k)\in S_G, N_{FF}(i,j,k)\neq 0, \\ N_{HF}(i+1,j,k) > N_{HF}(i,j,k)}} P_{i,j,k} \cdot \frac{\lambda_a \cdot [N_{HF}(i+1,j,k) - N_{HF}(i,j,k)] / N_{FF}(i,j,k)}{i\mu_a + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_a + \lambda_f}$$
$$+ \sum_{\substack{(i,j,k)\in S_G, N_{FF}(i,j,k)\neq 0, \\ N_{HF}(i,j,k+1) > N_{HF}(i,j,k)}} P_{i,j,k} \cdot \frac{\lambda_f \cdot [N_{HF}(i,j,k+1) - N_{HF}(i,j,k) - 1] / N_{FF}(i,j,k)}{i\mu_a + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_a + \lambda_f} \tag{3.16}$$

$$P_{FP} = \sum_{\substack{(i,j,k)\in S_G, N_{FF}(i,j,k)\neq 0, \\ N_{HF}(i,j,k) = N_{HF}(i+1,j,k) \\ N_{FF}(i,j,k) > N_{FF}(i+1,j,k)}} P_{i,j,k} \cdot \frac{\lambda_m \cdot [N_{FF}(i,j,k) - N_{FF}(i+1,j,k)] / N_{FF}(i,j,k)}{i\mu_a + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_a + \lambda_f} \tag{3.17}$$

$$P_{FST} = 1 - P_{FC} - P_S - P_{FP} \tag{3.18}$$

Let P_{FS} denote the probability that a full-rate serving FRUC is sub-rated before its completion or preemption. From the memory-less property of Markov process, P_{FS} can be expressed as in Equation (3.19).

$$P_{FS} = P_S + P_{FST} P_{FS} = P_S / (1 - P_{FST}) \tag{3.19}$$

In the same way, we consider the first event that occurs to a serving sub-rate FRUC. The FRUC may complete (with probability P_{SC}), be full-rated (with probability P_F), be preempted (with probability P_{SP}), or none of the aforementioned events occurs but a state transition occurs (with probability P_{SST}). The probabilities can be expressed in (3.20)-(3.23).

$$P_{SC} = \sum_{(i,j,k)\in S_G} P_{i,j,k} \cdot \frac{\frac{\mu_f}{2}}{i\mu_n + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \tag{3.20}$$

$$\begin{aligned} P_F = & \sum_{\substack{(i,j,k)\in S_G, N_{HF}(i,j,k)!=0, \\ N_{FF}(i-1,j,k)>N_{FF}(i,j,k)}} P_{i,j,k} \cdot \frac{i\mu_n \cdot [N_{FF}(i-1,j,k) - N_{FF}(i,j,k)]/N_{HF}(i,j,k)}{i\mu_n + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \\ & + \sum_{\substack{(i,j,k)\in S_G, N_{HF}(i,j,k)!=0, \\ N_{FF}(i,j,k-1)\geq N_{FF}(i,j,k)}} P_{i,j,k} \cdot \frac{N_{FF}(i,j,k)\cdot\mu_f \cdot [N_{FF}(i,j,k-1) - N_{FF}(i,j,k)+1]/N_{HF}(i,j,k)}{i\mu_n + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \\ & + \sum_{\substack{(i,j,k)\in S_G, N_{HF}(i,j,k)!=0, \\ N_{FF}(i,j,k-1)>N_{FF}(i,j,k)}} P_{i,j,k} \cdot \frac{N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} \cdot [N_{FF}(i,j,k-1) - N_{FF}(i,j,k)]/N_{HF}(i,j,k)}{i\mu_n + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \end{aligned} \tag{3.21}$$

$$P_{SP} = \sum_{\substack{(i,j,k)\in S_G, N_{HF}(i,j,k)!=0, \\ N_{FF}(i,j,k)==N_{FF}(i+1,j,k) \\ N_{HF}(i,j,k)>N_{HF}(i+1,j,k)}} P_{i,j,k} \cdot \frac{\lambda_n \cdot [N_{HF}(i,j,k) - N_{HF}(i+1,j,k)]/N_{HF}(i,j,k)}{i\mu_n + N_{FF}(i,j,k)\cdot\mu_f + N_{HF}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \tag{3.22}$$

$$P_{SST} = 1 - P_{SC} - P_F - P_{SP} \tag{3.23}$$

Let P_{FPrm} denote the probability that a full-rate serving FRUC is preempted before its completion, and P_{SPrm} denote that of a sub-rated serving FRUC. From the memory-less property of Markov process, they can be expressed in (3.24)-(3.25), respectively.

$$P_{FPrm} = P_{FP} + P_S P_{SPrm} + P_{FST} P_{FPrm} = (P_{FP} + P_S P_{SPrm})/(1 - P_{FST}) \tag{3.24}$$

$$P_{SPrm} = P_{SP} + P_F P_{FPrm} + P_{SST} P_{SPrm} = (P_{SP} + P_F P_{FPrm})/(1 - P_{SST}) \tag{3.25}$$

From equations (3.22-3.25), we can obtain P_{FPrm} and P_{SPrm}; they can be expressed in (3.26)-(3.27), respectively.

$$P_{FPrm} = [(1 - P_{SST})P_{FP} + P_S P_{SP}]/[(1 - P_{FST})(1 - P_{SST}) - P_S P_F] \tag{3.26}$$

$$P_{SPrm} = [(1 - P_{FST})P_{SP} + P_F P_{FP}]/[(1 - P_{FST})(1 - P_{SST}) - P_S P_F] \tag{3.27}$$

From the stationary state probabilities, we can obtain the average transmission rate of serving FRUCs in (3.28)

$$T_F = \sum_{(i,j,k)\in S_G, k>0} P_{i,j,k} \cdot \frac{N_{FF}(i,j,k)\cdot 128k + N_{HF}(i,j,k)*64k}{N_{FF}(i,j,k)+N_{HF}(i,j,k)} \qquad (3.28)$$

3.3.3 Cost Function Scheme

In this chapter, we consider a mobile data operator's revenue that consists of the transmission fee of normal users and the monthly fee of flat-rate users. Instead of calculating the total revenue, we propose a cost function representing the revenue loss due to blocked NUCs and due to the loss of flat-rate users. Since NUCs are charged by the volume of packets transmitted. In a fully utilized network, re-transmitting blocked NUCs only leads to more NUCs blocked. Therefore, we assume that blocked NUCs in a fully utilized network will not be re-transmitted, and thus represent revenue loss. The revenue loss of blocked NUCs is proportional to the blocking probability (P_{BN}) and the traffic load of NUCs ($\rho_n = \lambda_n / \mu_n$). The monthly revenue loss due to blocked NUCs can be expressed in (3.29), where D denotes the transmission charge of a NUC per busy hour, E the number of busy hours per month, and F the number of cells.

$$C_n = D \cdot E \cdot F \cdot \rho_n \cdot P_{BN} \qquad (3.29)$$

The revenue loss due to the loss of flat-rate users also depends on the blocking probability. Since flat-rate users are not charged by the volume of packet transmission, blocked FRUCs do not result in direct revenue loss. However, when the blocking probability is above a departure threshold, β, flat-rate users may become discontent and

start to switch to other operators. We assume that the number of flat-rate users lost per month is proportional to the discrepancy of the blocking probability (P_{BF}) above β. The monthly revenue loss due to lost flat rate users can be expressed in (3.30), where X denotes the total number of flat-rate users, Y the percentage of flat-rate users lost due to each percentage increase of blocking probability above β, and Z the monthly charge of a flat-rate user.

$$C_f = \begin{cases} X \cdot Y \cdot Z \cdot (P_{BF} - \beta) \cdot 100, & if (P_{BF} > \beta) \\ 0, & otherwise \end{cases} \tag{3.30}$$

The total monthly loss (C) is C_n plus C_f. Dividing the monthly revenue loss by D, E, and F, we obtain the cost function, as shown in (3.31-3.33), where α represents the cost weighting factor of flat-rate connections.

$$C = C_n + C_f = \begin{cases} D \cdot E \cdot F \cdot \rho_n \cdot P_{BN} + X \cdot Y \cdot Z \cdot (P_{BF} - \beta) \cdot 100, & if (P_{BF} > \beta) \\ D \cdot E \cdot F \cdot \rho_n \cdot P_{BN}, & otherwise \end{cases} \tag{3.31}$$

$$C = \begin{cases} \rho_n \cdot P_{BN} + \alpha \cdot (P_{BF} - \beta), & if (P_{BF} > \beta) \\ \rho_n \cdot P_{BN}, & otherwise \end{cases} \tag{3.32}$$

where

$$\alpha = \frac{X \cdot Y \cdot Z}{D \cdot E \cdot F} \cdot 100 \tag{3.33}$$

When the cost weighting factor of FRUCs is less than that of NUCs ($\alpha < \rho_n$), the scheduler should give priority to NUCs without considering the FRUC blocking probability. On the other hand, when α is larger than ρ_n, the scheduler should give priority to NUCs when the FRUC blocking probability is below the departure threshold

(β), but it should give priority to FRUCs when FRUC blocking probability is above β. Note that β should be chosen to reflect the beginning of user dissatisfaction as the blocking probability increases; its proper value may be obtained from the past operation data.

3.3.4 An Iterative Algorithm

To minimize the cost function, an iterative algorithm as shown in Fig. 3.5, was developed to obtain the stationary state probabilities, the optimum number of guard channels, and the performance measures. The iterative algorithm first initializes system input parameters, such as the power limit in a cell, the maximum numbers of serving NUCs and FRUCs in a cell, etc., in Steps 1-3. The for loop in step 4 determines the optimum number of guard channels. The while loop in Step 5 is iterations that obtain the stationary probabilities of existing states. In Steps 10-11, based on the stationary state probabilities, we can obtain the performance measures and the cost function. In step 12, we obtain the minimum value of the cost function.

---

1. Obtain $\Omega(N_F, N_H)$ from Equation (3.1), and $N_F^*(0)$ and $N_H^*(0)$ from Equation (3.2);
2. Set $\theta_{i,j,k}$ =1 for each existing state (i,j,k), i.e., (i,j,k) ∈ S_G defined in Equation (3.3);
3. Initialize old $P_{i,j,k} = \frac{1}{|S_G|}$, and C_{min} =0 and G_{opt} =0 ;
4. For (G=0;G<=10; G++) { /* Find the optimum number of guard channels for NUCs*/
5. While (1) { /* Obtain the stationary probabilities of the analytic model*/
6. For all states (i,,j,k) , new $P_{i,j,k}$ = *Inflow(i,j,k)/Outflow(i,j,k)*; /* based on the balance equations (3.4-3.6)*/
7. If |new $P_{i,j,k}$ - old $P_{i,j,k}$ | $\leq 10^{-16}$ for all states, break; /* If system is in equilibrium, go to 10*/
8. For all states (i,,j,k) , old $P_{i,j,k}$ = new $P_{i,j,k}$;
9. } // while
10. Calculate the Performance Measures Equations (7)-(28) based on the G value;
11. Calculate the ρ_n , P_{BN} , P_{BF} and cost function C in (3.32);
12. If {(C_{min} =0) or (C< C_{min})} C_{min} = C; G_{opt} = G;
13. } // next G

Fig. 3.5: An iterative algorithm minimizes the cost function

3.4 Numerical results and Discussions

In the analysis below, we assume the spread spectrum bandwidth (W) of the WCDMA network is 5 MHz, the uplink full-rate transmission is 128kbps and the half-rate 64kbps. The size of the NUC WQ (B) is 4 and the size of FRUC WQ (Q) is 10. The mean service time of NUCs ($1/\mu_n$) is assumed to be 2 minutes and the mean service time of FRUCs ($1/\mu_f$) is assumed to be 10 minutes. The arrival rate of FRUCs (λ_f) is fixed at 0.01 connections/sec. and that of NUC (λ_n) varies in the range of 0.005-0.03 connections/sec, i.e., ρ_n varies in the range of 0.6-3.6 connections. We compare four connection schedulers: S_{All}, S_{NRA}, S_{NPrm} and S_{NWQ}. The iterative algorithm for each scheduler has been developed in C language. The program was run on a laptop PC with 1.6GHz Pentium CPU and 512MB RAM. For each traffic load, the stationary state probabilities can converge in less than one minute.

To choose a suitable α value for the cost function in (3.31), we use the operation data from ChungHwa Telecom (CHT) in Taiwan, and make assumptions if operation information is unavailable. In (3.29), D (the 128kbps transmission charge per hour) is NT$562.5, E=60 (i.e., the number of busy hours per day equals to 2), F (the number of cells) is 1000. The number of flat-rate users, X, is 2 hundred thousands, Y is assumed to be 0.001 (i.e., one out of a thousand users would quit per month due to a percentage increase of blocking probability above β, and Z (the monthly fee of a flat-rate user) is NT$850. β should be chosen to reflect the level of user dissatisfaction; it was chosen to be 0.02, which is the target blocking probability for flat-rate subscribers of CHT. Given that, we can obtain the factor α = 0.504. Note that α is less than ρ_n (0.6-3.6) in our experiments, i.e., the cost weighting factor of FRUCs is less than that of NUCs.

Fig. 3.6.a plots the cost function as NUC traffic increases. The FRUC traffic is fixed at 0.01 connections/sec. The cost function represent the revenue loss of the operator; the less the better. The results indicate S_{NRA} has the least amount of revenue loss among all schedulers.

When the NUC traffic less than 0.015 connections/sec, the revenue losses of S_{All}, S_{NWQ}, and S_{NPrm} are as small as that of S_{NRA}, but the losses rise rapidly as the NUC traffic increases above 0.02 connections/sec, in particular for S_{NPrm}. This indicates when the system traffic load is high, waiting queues and preemption are necessary, but rate-adaptation is not. This is because sub-rated connections are less "bandwidth efficient," and results in system throughput reduction and revenue loss. S_{NPrm} suffers the biggest revenue loss when the NUC traffic load is high. This indicates that preemption is essential in reducing the revenue loss.

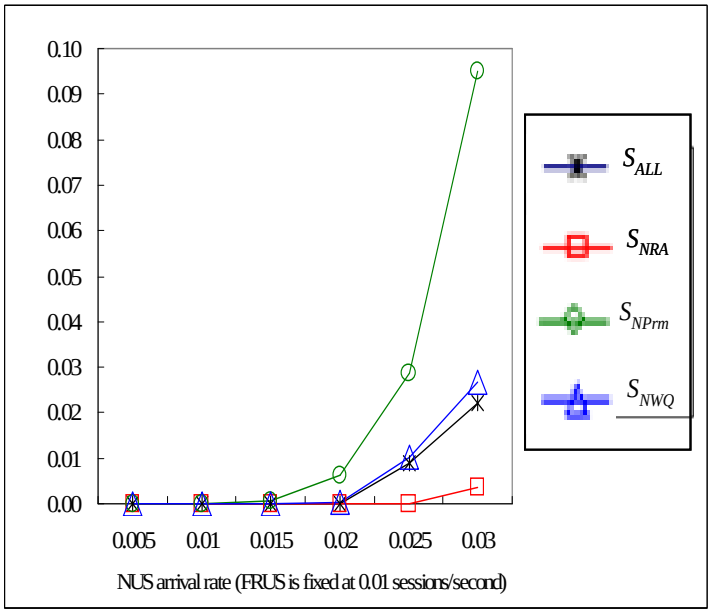


Fig. 3.6.a: The cost function (C) with α=0.504 and β=0.02, (B=4, Q=10)

Fig. 3.6.b presents the optimum number of guard channels for each scheduler under different traffic loads. S_{All}, S_{NRA}, and S_{NPrm} do not need any guard channels. The results indicate the NUC waiting queue plus either preemption or rate-adaptation are effective in giving NUCs priority. Guard channels may reduce the system throughput and thus the revenue. In contrast, S_{NWQ} needs one guard channel when the NUC traffic load is low because it has no NUC waiting queue. However, when the traffic load is high, no guard channel is needed because of the same reason that guard channels leads to system throughput

reduction and revenue loss.

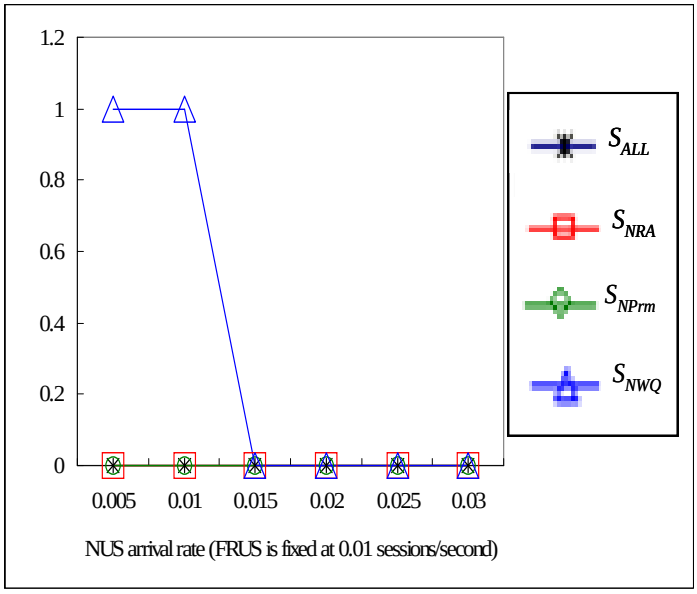


Fig. 3.6.b: The numbers of guard channels (GC) with α=0.504 and β=0.02, (B=4, Q=10)

Fig. 3.7.a plots the blocking probabilities of NUCs as the NUC arrival rate increases. All schedulers provide very low blocking probabilities for NUCs, except S_{NPrm}; the NUC blocking probability of S_{NPrm} increases more rapidly as NUC traffic increases. This is because sub-rated FRUCs are lessefficient in using spectrum. If FRUCs can only be sub-rated, but cannot be preempted, there would be more sub-rated FRUCs when the system traffic load is high. As a result, the overall system throughput decreases, and more NUCs are blocked. Therefore, preempting FRUCs is essential in reducing the blocking probability of NUCs. When the NUC traffic is high and no NUC waiting queue is used (as in S_{NWQ}), the blocking probability slightly rises. This indicates the NUC waiting queue is necessary when the system load is close to its capacity.

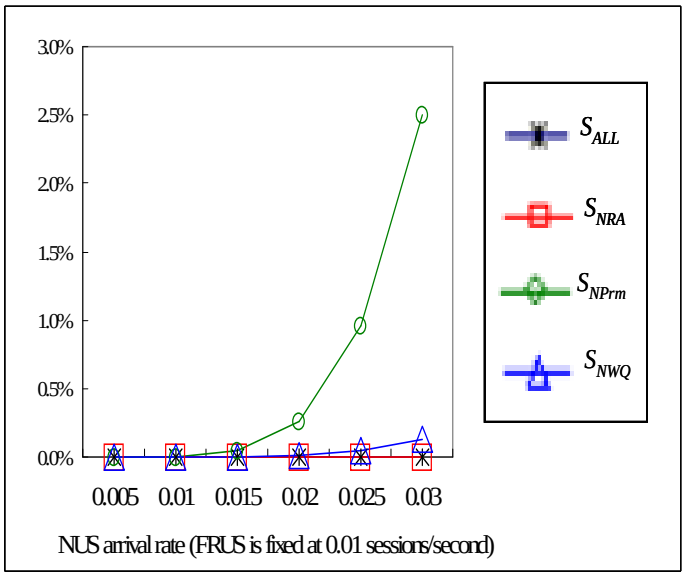


Fig. 3.7.a: Average NUC blocking probabilities (P_{BN}) with B=4, Q=10

Fig. 3.7.b plots the blocking probabilities of FRUCs as the NUC arrival rate increases. The results indicate that the FRUC blocking probabilities of S_{NRA} and S_{NPrm} are about the same; S_{NRA} outperforms S_{NPrm} by a small margin. Even though FRUCs cannot be preempted in S_{NPrm}, the blocking probability of S_{NPrm} is still higher than that of S_{NRA}. This is also because sub-rated FRUCs are less "bandwidth efficient". In addition, the FRUC blocking probabilities of SAll and S_{NWQ} are higher and rise more rapidly as NUC traffic increases, because FRUCs are impaired by both preemption and sub-rating. The fluctuations of FRUC blocking probabilities in S_{NWQ}, when the NUC traffic increases from 0.01 to 0.015 connections/sec, are caused by the change in the number of guard channels

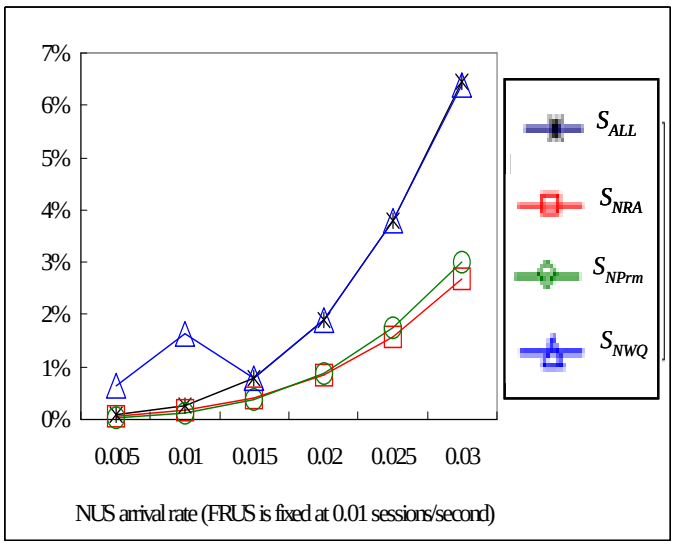


Fig. 3.7.b: Average FRUC blocking probabilities (P_{BF}) with B=4, Q=10

Fig. 3.8.a plots the average waiting times (i.e., queueing times) of NUCs as the NUC traffic increases. The waiting times of NUCs in schedulers S_{All} and S_{NRA} are very insignificant under all traffic loads, i.e., NUCs are rarely queued. This is because serving FURCs can be preempted to free radio resources. If FRUCs cannot be preempted, such as in S_{NPrm}, the average waiting time of NUCs increases steadily as the traffic of NUCs increases.

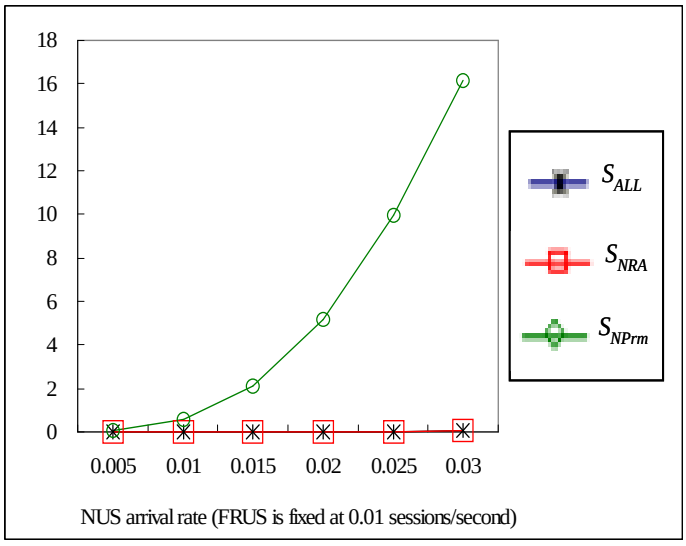


Fig. 3.8.a: Average NUC waiting times (W_{TN}) with B=4, Q=10

Fig. 3.8.b plots the average waiting times of FRUCs as the NUC traffic increases. The waiting times of all schedulers show the same trend of rising as the NUC traffic increases. Even when the system traffic is low, the average waiting time of FRUCs in S_{NRA} is as large as 60 seconds, which is unacceptable for real-time applications. S_{NPrm} provides the shortest waiting time, while S_{NRA} the longest. The difference can be as high as 100 seconds when the NUC traffic is 0.03 connections/sec. Note that the fluctuations of FRUC waiting times in S_{NWQ}, when the NUC traffic increases from 0.01 to 0.015 connections/sec, are also caused by the change in the number of guard channels. This change of guard channels also results in fluctuations of S_{NWQ} results in later figures.

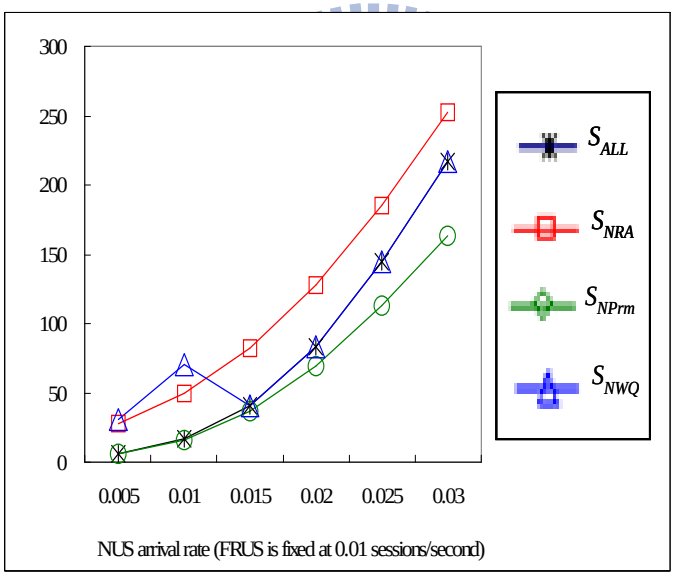


Fig. 3.8.b: Average FRUC waiting times (W_{TF}) with B=4, Q=10

Fig. 3.9.a depicts the probability that a NUC is queued. The results indicate that NUCs in SAll and S_{NRA} are very rarely put into the waiting queue because FRUCs can be preempted to free radio resources. On the other hand, NUCs in S_{NPrm} are more likely to be queued. The probability that a new NUC is queued increases steadily and rapidly as the traffic load increases. The NUC queueing probability in S_{NPrm} can be as high as 50%. This indicates that

preempting FRUCs is critically in reducing the queueing probability of NUCs. Fig. 3.9.b depicts the queueing probabilities of FRUCs as the NUC traffic increases. In general, the probability that a FRUC is queued increases as the traffic load increases. S_{NRA} has the largest FRUC queueing probability, because FRUCs cannot be sub-rated. Other schedulers provide about the same queueing probabilities under all traffic loads.

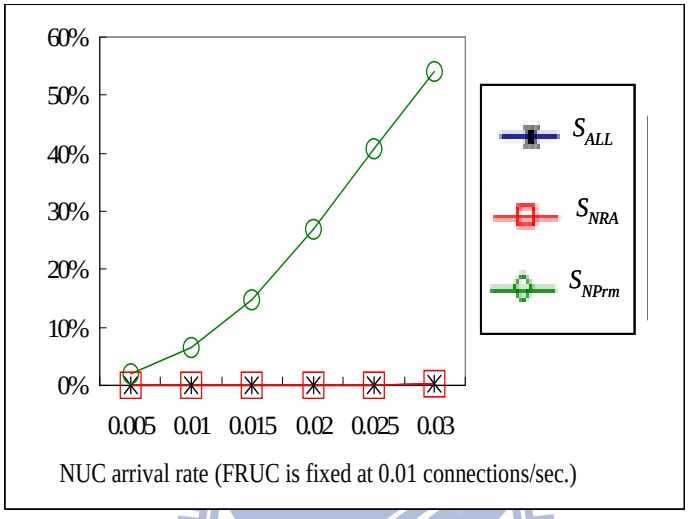


Fig. 3.9.a: Average NUC queueing probabilities (P_{QN}) with B=4, Q=10

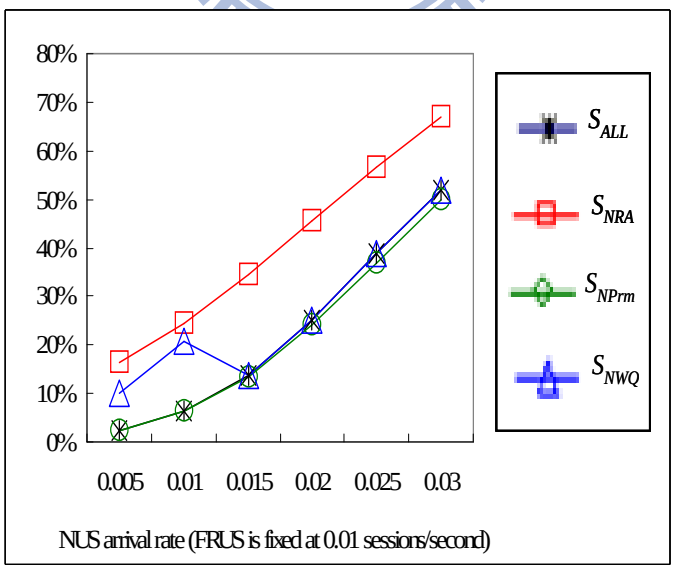


Fig. 3.9.b: Average FRUC queueing probabilities (P_{QF}) with B=4, Q=10

Figs. 3.10.a and 3.10.b present the probabilities that a serving full-rate and half-rate FRUC would be preempted before completion. All schedulers display the same trend of rising preemption probabilities as the traffic load increases. The preemption probability of S_{NRA} is lower than other schemes by a small margin. This is because sub-rated FRUCs are less bandwidth efficient.

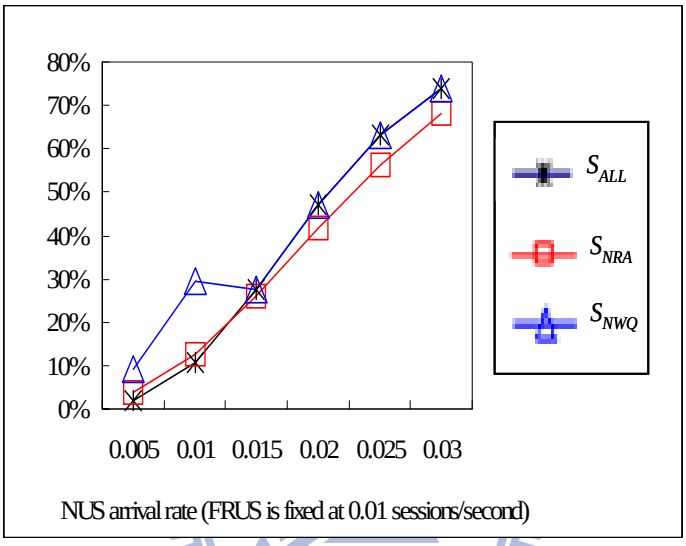


Fig. 3.10.a: Average preempted probabilities of a serving full-rate FRUCs (P_{FPrm}) with B=4, Q=10

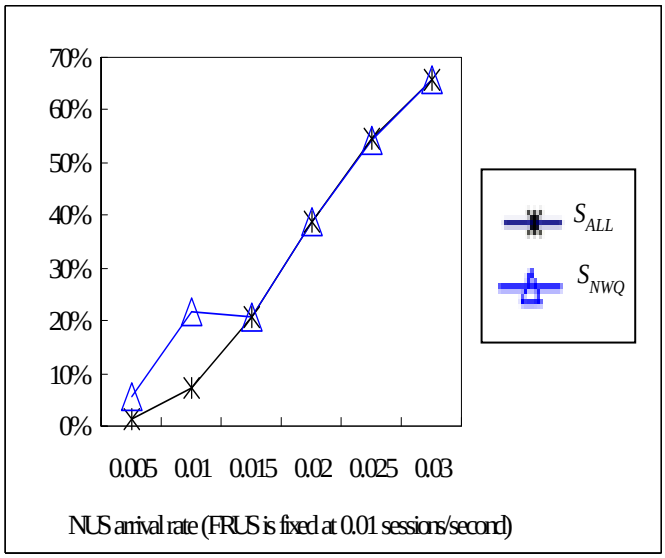


Fig. 3.10.b: Average preempted probabilities of a serving half-rate FRUCs (P_{SPrm}) with B=4, Q=10

Figs. 3.11 presents the probabilities that a serving full-rate FRUC would be sub-rated. As the NUC traffic increases, the sub-rating probabilities of S_{All} and S_{NWQ} first rise and then decline. The decline is because when the system traffic is high, FRUCs are more likely to be preempted. On the other hand, the sub-rating probability of S_{NPrm} increases more rapidly and saturates later as the traffic load increases. This is because as the NUC traffic increases, S_{NPrm} cannot preempt FRUCs; it can only sub-rate more FRUCs.

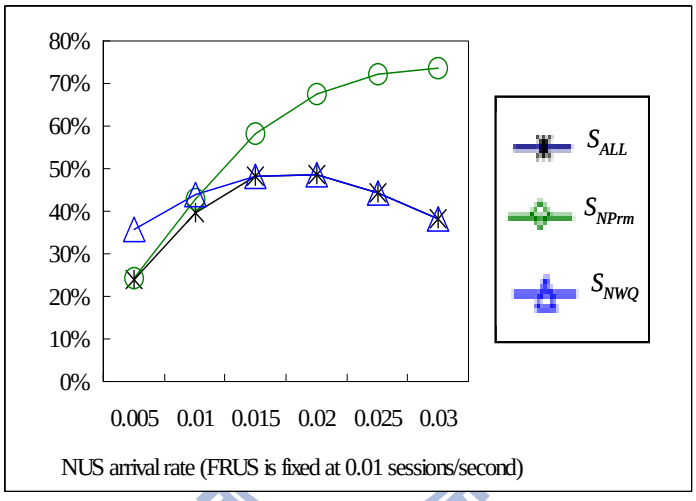


Fig. 3.11 Average sub-rated probabilities of a serving full-rate FRUCs (P_{FS}) (B=4, Q=10)

Fig. 3.12 plots the average transmission rate of FRUCs. Since FRUCs may be sub-rated and/or preempted, the average transmission rate of FRUCs is reduced. In S_{NRA}, no FRUCs are sub-rated. In S_{All} and S_{NWQ}, a FRUC can be sub-rated and preempted; the average transmission rate is reduced to as much as 70% of the full rate transmission when the system traffic is heavy.

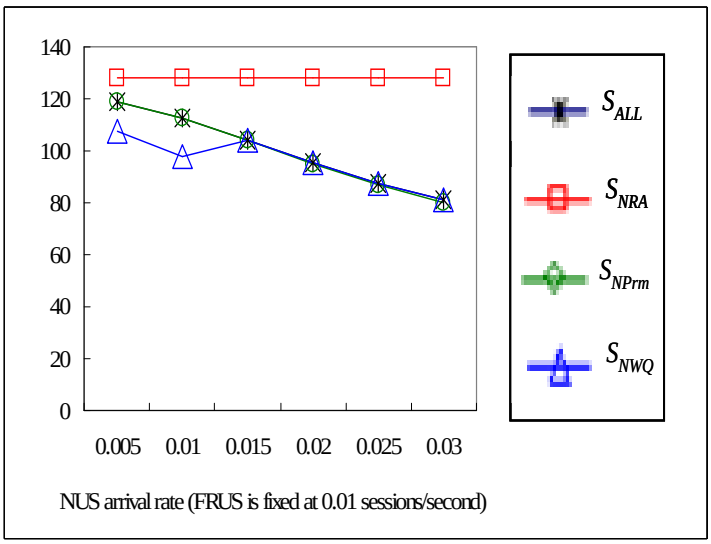


Fig. 3.12 Average transmission rate of serving FRUCs (T_F) (B=4, Q=10)

3.5 Conclusions

This chapter, we investigate four combinations of scheduling techniques, queueing, guard channels, preemption and rate-adaptation, on their effectiveness in scheduling UMTS R99 uplink connections to reduce the revenue loss of the operators serving both normal and flat-rate users. We proposed a cost function representing the revenue loss due to both blocked normal user connections and lost flat-rate users. The optimum numbers of guard channels was determined by an iterative algorithm. The analytic results indicate when α, the cost weighting factor of flat-rate users, is less than ρ_n, queueing and preemption are essential for connection scheduling to maximize the revenue. Rate-adaptation is ineffective, because half-rate connections are less bandwidth-efficient. Sub-rating FRUCs reduced the system throughput and the operator revenue. In addition, no guard channel is needed, if queuing and preemption are used, because guard channels increase the blocking probability of FRUCs and reduces system throughput.

In this chapter, we consider uplink connection scheduling only. We did not study downlink traffic scheduling, which can be done on packet level. In our study, the cost weighting factor of flat-rate users (α), is less than that of normal users (ρ_n). Further study is

needed for UMTS networks with α larger than ρ_n, which is possible when the number of flat-rate users increases or the normal user traffic decreases. In this situation, a more sophisticated scheduler is needed. The scheduler should give priority to NUCs when the FRUC blocking probability is below the departure threshold, β. When FRUC blocking probability is above the threshold, FRUCs should have priority.

CHAPTER 4

Flate-Rate Packet Scheduling for the WCDMA Systems with HSDPA

4.1 Introduction

A Universal Mobile Telecommunications System（UMTS）network consists of three interacting domains: the Core Network (CN), the UMTS Terrestrial Radio Access Network (UTRAN) and Mobile Stations (MSs). Fig. 4.1 depicts the network architecture of the UMTS network. The CN includes Circuited-Switch (CS) domain (i.e., MSC/VLR and GMSC) and Packet-Switched (PS) domain (i.e., SGSN and GGSN). The UTRAN includes multiple Radio Network Controllers (RNCs), each of which connects to multiple Node-Bs. The air interface of UTRAN is based on Wideband CDMA (WCDMA) technology and the details can be found in the 3GPP Release 99 specifications [27]. The peak transmission rate between a Node-B and a stationary mobile station (MS) is 2 Mbps.

To provide a higher data transmission rate for packet data services, WCDMA has evolved into High Speed Downlink Packet Access (HSDPA) described in 3GPP Release 5 specifications [28]. The HSDPA is expected to achieve a peak data rate over 10 Mbps, which is a significant improvement over the peak data rate (2 Mbps) of the 3G WCDMA Release 99. The idea behind HSDPA is that the network transmits the downlink packets to the MS with maximum carrier-to-interference ratio (max. C/I) first at a high data rate. To enable HSDPA, the radio packet scheduler is moved from the radio network controller (RNC) to Node-Bs.

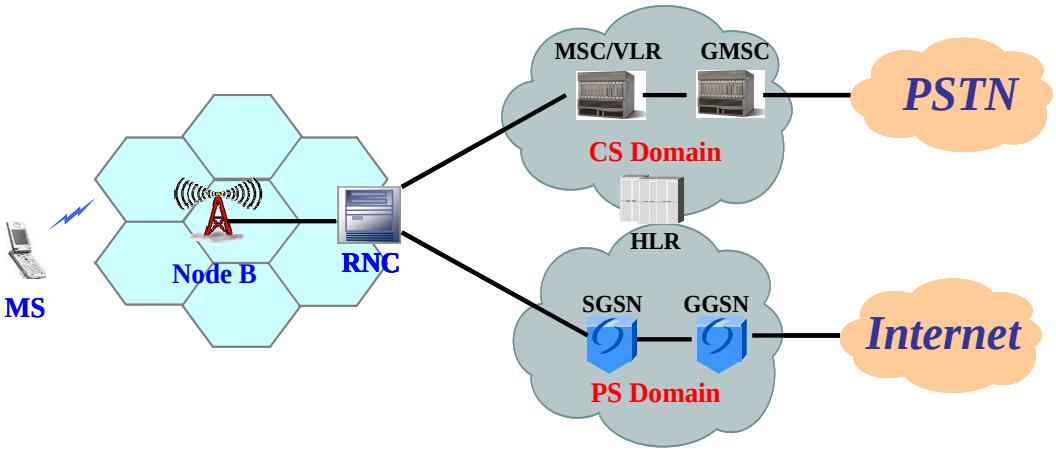


Fig. 4.1: The network architecture of the UMTS network

The packet scheduler in Node-B tracks the channel quality of each MS by measuring the SIR (Signal to Interference Ratio) on the CPICH (Common Pilot Indicator Channel) and allocates the High Speed Downlink Shared Channel (HS-DSCH) to the MS with the best SIR value [29-32]. As a result, the network can achieve the maximum throughput for downlink packets. To prevent the MSs with poor radio channel quality from starvation, traditional Round Robin (RR) packet schedulers can be used to ensure service fairness [33-34], but RR schedulers do not fully utilized the advantages of HSDPA. Proportionally Fair (PF) packet schedulers realize a reasonable trade-off between radio efficiency and fairness [35]. The network transmits downlink packets to the MS whose normalized instantaneous SIR value, the instantaneous SIR value divided by the average SIR value of the on-line transmission period, is the largest among all MSs. The numerical results show that both its system throughput and its worst case user throughput are larger than those of RR schemes.

However, the packet schedulers described above did not consider the revenues of mobile operators. When the mobile operators begins to provide flat-rate packet services to users, revenue, instead of fairness or capacity, is the most important consideration in the HSDPA network for the operators. Flat-rate users pay fixed monthly charge to access the

HSDPA network without limiting the packet transmission. Since usage incurs no extra cost, flat-rate users could occupy most of the network radio resources. Without special treatments for different classes of user's packets, users charged by usage may be blocked out from accessing the HSDPA network and compromise the mobile operator's revenues. Therefore, it is important for a packet scheduler in a Node-B to fairly utilize the network resources for both the users charged by volume and the flat-rate users.

In this chapter, we study how to use packet scheduling techniques to control data packet transmission and guarantee the revenues of mobile operators without impairing flat-rate users too much. We consider two types of downlink packets, Charged Packets (CPs) and Flat-Rate Packets (FRPs). From the viewpoints of mobile operators, revenue and customer satisfaction need to be well balanced. To garner more revenue, the packet scheduler needs to give CPs a higher priority over FRPs. On the other hand, to ensure customer satisfaction, the dropped probability of FRPs needs to be kept below a certain threshold. In this paper, we present two enhanced packet schedulers that constantly monitor the dropped probabilities of both CPs and FRPs, and schedule down-link packet transmission so that the dropped probability of CPs could be below P1 and that of FRPs could be below P2. The scheduling techniques we used include a Priority Queue (PQ) with dynamic guard slots for CPs, and a PQ with Discard Timer (DT) for FRPs. Analytic models have been used to evaluate their performance in terms of packet dropped probability and downlink radio utilization.

4.2 HSDPA Basic Principles

Instead of the Downlink Shared Channel (DSCH) used in the WCDMA, HSDPA provides a new transport channel called High Speed DSCH (HS-DSCH) to transmit the downlink packets to MSs [28, 31]. In HSDPA, a large amount of radio resources can be assigned to a single MS on a Transmission Time Interval (TTI) basis. For each TTI (also referred to as a frame, a 2 ms interval), the Node-B selects an adequate Modulation Coding

Scheme (MCS), such as Quadrature Phase-Shift Keying (QPSK) or 16-Quadrature Amplitude Modulation (QAM), for each served MS according to the quality of downlink radio signal and the current system load. The better quality of downlink radio signal between the Node B and the MS is, the higher data rate of MCS can be selected. Each MCS value chosen for a served MS determines the data transmission rate for the served MS in the next TTI.

The High Speed Shared Control Channel (HS-SCCH) is a downlink control channel at a fixed rate (e.g., 60 kbps) and carries downlink signaling directing the HS-DSCH transmission. It provides packet transmission timing and coding information, so that each served MS listens to the HS-DSCH at the correct time using the correct codes for its downlink packets.

Fig. 4.2 depicts the downlink Spreading Factor (SF) codes allocation tree for the HS-DSCH and HS-SCCH in an HSDPA network. The SF codes for the HS-DSCH and HS-SCCH with orthogonal character must be fixed at 16 and 128, respectively. There are at most 15 downlink SF codes for the HS-DSCH that can be assigned to one MS in a TTI to achieve an ideal peak rate of 14 Mbps when 16- QAM full rate MCS is used in a frequency band of 5MHz. A downlink SF code for the HS-SCCH can instruct only one MS to receive the downlink packets belonged to it, and there are at most four HS-SCCH codes can be used to control downlink packet transmission for all MSs. Other SF codes, except those for the HS-DSCH and HS-SCCH, can be assigned to transport voice calls in parallel with HS-DSCH data transmission or for non-HSDPA data transmission. As a result, in a TTI at most four MSs can be instructed by four HS-SCCH codes, and the selected MSs share 15 HS-DSCH codes to receive their downlink packets.

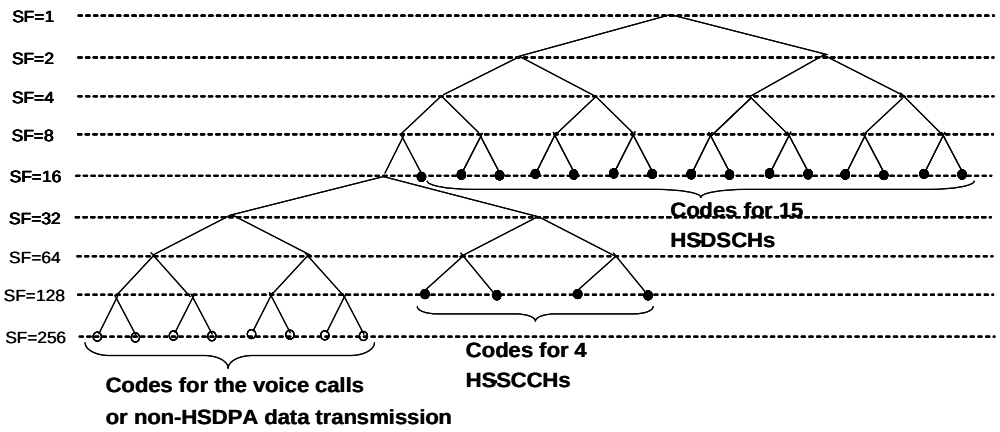


Fig. 4.2: Downlink SF codes allocation tree for HS-DSCH and HS-SCCH

Fig. 4.3 depicts an example downlink packet scheduling for four MSs, MS1-MS4, in a cell. Each MS can at most monitor four HS-SCCH codes and can only be assigned one HS-SCCH code belong to it in a TTI by a Node-B, and then in the assigned HS-SCCH code, the corresponding MS can be instructed to receive its downlink packets using the downlink HS-DSCH codes assigned to it. The time interval between the HS-SCCH instruction for a MS and its correspondent HS-DSCH transmission for this MS is 4/3ms. In the example depicted in Fig. 3, MS1 and MS2 are instructed to receive downlink packets in the first TTI, MS2 and MS3 to receive downlink packets in the second TTI, and all MSs to receive downlink packets in the third TTI.

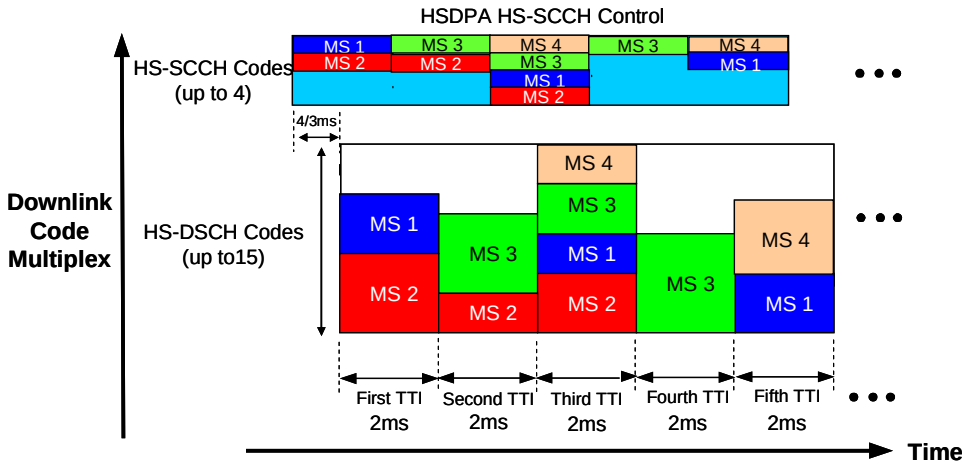


Fig. 4.3: An example downlink packet scheduling for MS1-MS4 in a cell

4.3 System Models and Assumptions

According to the 3GPP specifications, when a MS creates a data session in the PS Domain, a SGSN can send Radio Access Bearer (RAB) parameters in the RAB assignment request message to the RNC to indicate the downlink packet priority [36-37]. In addition, a RNC in a HSDPA network can send the downlink packet scheduling policy to a Node-B, such as packet discard timer and scheduling priority, during the radio link setup procedure [38]. Fig. 4.4 depicts the parameters sent from a SGSN through the RNC to the Node-B during the RAB assignment procedure and the radio link setup procedure. In this paper, these parameters will be used in the Node-B for packet scheduling.

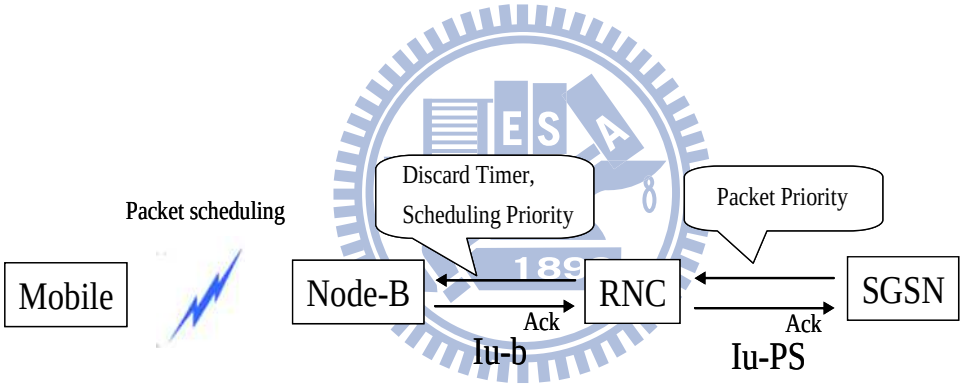


Fig. 4.4: The scheduling parameters sent from an SGSN through a RNC to a Node-B in HSDPA network

To simplify the analytic model, we assume that a cell is partitioned into two zones: a good SIR zone where the MSs have better SIR values, and a poor SIR zone where the MSs have poor SIR value (e.g., due to the distance from the Node-B or multipath signal fading). Using the same number of HS-DSCH codes, an MS in the good SIR zone can receive downlink data packets at twice the data rate of an MS in the poor SIR zone. Since a TTI is only 2ms, we assume that MSs do not move from a zone to another when waiting to receive a packet from the Node-B. Since downlink FRPs are given the lowest priority, in this paper we

only consider three types of downlink packets : CP and FRP packets for the MSs in good SIR zone will be refered to as CP_G and FRP_G; CP packets for the MSs in poor SIR zone will be refered to as CP_P. The queueing models for the four downlink packet scheduling methods we studied are described below.

4.3.1 M-PQ Method

To maximize radio network throughput, downlink packets for the MSs in good SIR zone are transmited first, i.e., CP_G and FRP_G are given priority over CP_P. CP_G and FRP_G are served on a FCFS basis. In implementation, CP_G, FRP_G and CP_P can be put into a PQ when the network has no free HS-DSCH codes. We refer this scheduling method as Max. C/I first in a PQ (M-PQ).

The queueing model for M-PQ scheme is depicted in Fig.4.5. When a new CP_G (line a) or a new FRP_G (line b) request arrives, it is served immediately (line e) if there are free HS-DSCH codes by assigning a HS-SCCH code to this packet and instructing the correspondent MS to receive the packet. Otherwise, this new CP_G or FRP_G request can be put into the PQ before the waiting CP_P. If the PQ is full, the request is rejected (dotted line d). When a new CP_P (line c) request arrives, it is served immediately (line e) if here are free HS-DSCH; otherwise, this new CP_P request can be put into the PQ after the waiting CP_G and FRP_G. If the PQ is full, the request is rejected (dotted line d). When a CP_G or FRP_G or CP_P finishes, it releases radio channels (line f).

4.3.2 P-PQ Method

To maximize the operator's revenue, charging packets (CP_G and CP_P) are given priority over flate rate packets (FRP_G). In addition, since CP_G can be transmitted at a higher data rate, they have priority over CP_P. In implementation, CP_G, FRP_G

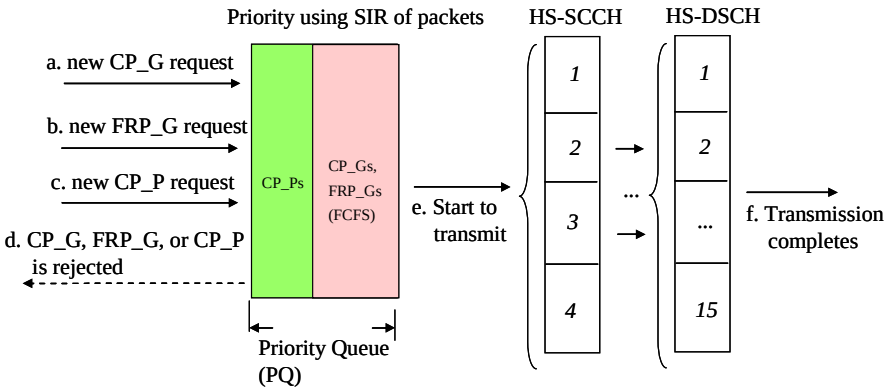


Fig. 4.5: A queueing model for M-PQ scheme

and CP_P can be put into a PQ according to their priority when the network has no free HS-DSCH codes. CP_G can be put before the waiting CP_P and FRP_G in the PQ and are served with the highest priority by the Node_B. In addition, CP_P can be put before the waiting FRP_G in the PQ and are served with the second priority by the Node_B. We refer this scheduling method as CPs first in a PQ (P-PQ).

The queueing model for P-PQ scheme is depicted in Fig. 4.6. The operation scenarios are same as M-PQ method except how the downlink packets are put into the PQ. CPs are always served first by the Node-B when they can be put into the PQ.

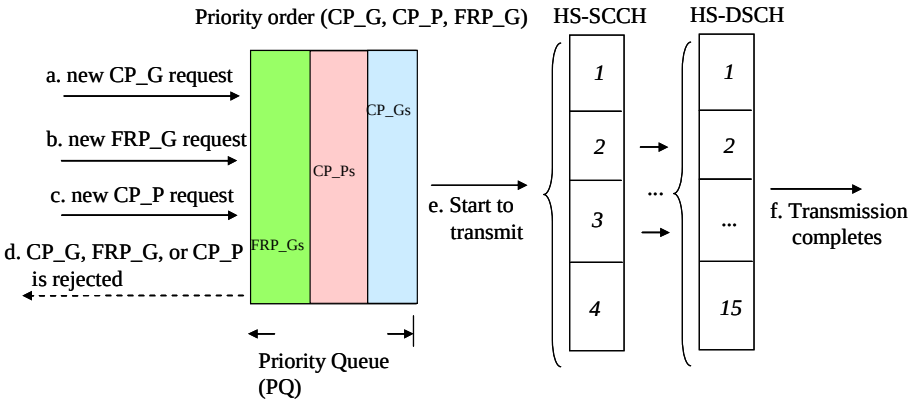


Fig. 4.6: A queueing model for P-PQ scheme

4.3.3 DDT-PQ Method

The FRP_G could occupy most of queue resources in the P-PQ method to block the new downlink charging packets (CP_G and CP_P). To maximize the operator's revenue, FRP_G can be dropped if they stay over a period of time in the PQ without being served. Each downlink packet can have a Discard Timer (DT) value sent from the RNC to Node-B. When the DT expires, the packet is discarded if it is not transmitted yet. In our design, a Node-B can dynamically adjust the DT value, i.e., increase or decrease, for FRPs depending on the traffic load. We refer this scheduling method as Dynamic Discard Timer for FRPs in a PQ (DDT-PQ).

The queueing model for DDT-PQ scheme is depicted in Fig. 4.7. The operation scenarios are same as P-PQ method except that a FRP is discarded if it is still in the PQ when the DT expires (dotted line g). In this scheme, the DT values of the FRPs have a Lower Bound (DTLB) and an Upper Bound (DTUB). A larger DT value increases the probability that a FRP is served; a lower one decreases the probability. By adjusting the DT value based on the system load, we can control the dropped probability of FRP in the PQ. When the DTLB is chosen, a Node-B can not decrease the DT value for the FRPs lower than it. That means the worse case of dropped probability of FRP can be controlled. When the DTUB is chosen, a Node-B can not increase the DT value for the FRPs higher than it. That means the best case of dropped probability of FRP can be controlled.

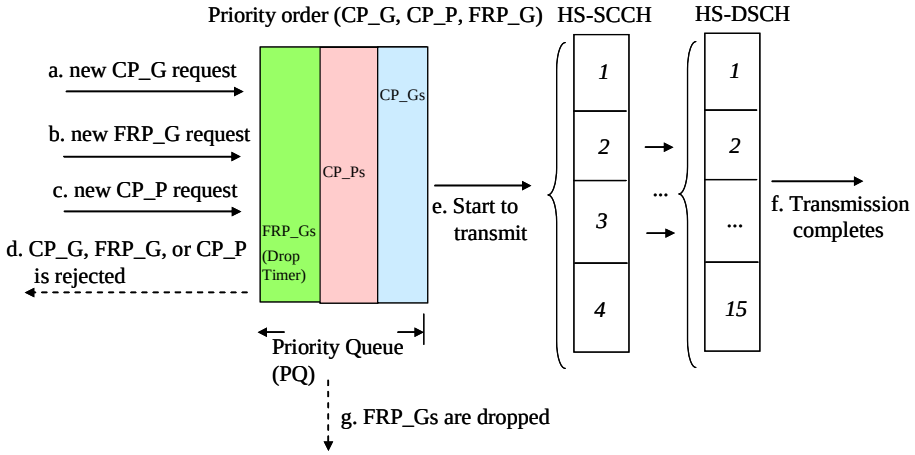


Fig. 4.7: A queueing model for DDT-PQ scheme

The proposed algorithm of dynamically adjusting DT implementation in Node-B can be described as follows. In order to consider the system throughput and the fairness for both CP and FRP, the acceptable Dropped Probability (DP) of CP (P1) and FRP (P2) can be defined in a Node-B. When the DP of CP is larger than P1 and DT is larger than DTLB, then DT can be decreased by a Discard Timer Step Value (DTSV). On the other hand, when the DP of FRP and is larger than P2 and DT is lower than DTUB, then DT can be increased by DTSV. For implementation consideration, a lower P1 or a lower DTLB value can gurantee the throughput for CPs; on the other hand, a lower P2 or a higher DTUB value can gurantee the fairness for FRPs. By dynamically adjusting the DT, we can gurantee the system throughput of CP even when the arrival of FRP is heavy, and can gurantee the fairness of FRP when the arrival of CP is low. The values of the P1,P2,DTSV,DTLB,DTUB and DT can be chosen by mobile operators. The peudocode of the proposed algorithm is described in Fig. 4.8.

Parameters:

DT : Current Discard timer for new FRPs arrive

SW (Slicing Window) : a piece of system processing time

P1: the acceptable Dropped Probability (DP) of CP

P2: the acceptable Dropped Probability (DP) of FRP

DTLB: Discard Timer Lower Bound

DTUB: Discard Timer Upper Bound

DTSV: Step Value of Discard Timer

Pseudocode

Initialize assign *DT* value and renew a timer for current *SW*

Repeat

(1)Any new downlink FRP do not need to be put into PQ goto (3)

(2)Each downlink FRP put into PQ assigns a timer equal to DT

(3) If current SW !=0 goto (4)

(a) All timers for FRPs in PQ - = current SW

(b) Remove all FRPs in PQ with timer = 0

(c) Upadte the number of FRPs which Discard in current *SW*

(d) Recalculate the DP of CP and DP of FRP based on the number of CPs and FRPs which accept or Discard in previous (n-1) *SW*s and current *SW*

(e) If DP of CP > P1 and DT > DTLB , DT - =DTSV , goto g)

(f) If DP of FRP > P2 and DT < DTUB , DT + =DTSV

(g) Renew a SW timer , goto (1)

(4) Upadte the number of CPs and FRPs which accept or Discard in current *SW*

(5) Update Current *SW* (decrease) and goto 1)

End Repeat

Fig. 4.8: A peudocode for DDT-PQ scheme

4.3.4 DGS-PQ Method

To maximize the operator's revenue, a Node-B can reserve some capacities of the PQ for CPs only when the network load is high. That means the CPs have more chance to be served than FRPs when the system load is heavy. Let Guard Slots (GS) denote the number of the reserved capacities of the PQ for the CPs. The concept of GS is similar to that of guard channel used in cellular network [39-42]. In addition, a Node-B can dynamically adjust the value of GS, i.e., increase or decrease, for CPs depending on the traffic load. In our design, the value of GS can have a fraction part to represent the probability of new

FRPs request can be put into PQ. For example, if GS is set to be 1.2, then all new CPs request and 80% of new FRPs request will be allowed to be put into PQ whenever free capacity of PQ is 2. We name this control admission policy as GS admission procedure. We refer this scheduling methd as Dynamic Guard Slots in a PQ for CPs (DGS-PQ).

The queueing model for DGS-PQ scheme is depicted in Fig. 4.9. The working scenario is same with P-PQ method except that an an adjustable portion of the PQ is reserved for the new downlink CPs. In this scheme, the GS values of the CPs have a Lower Bound (GSLB) and an Upper Bound (GSUB). A larger GS value increases the probability that a CP is served; a lower one decreases the probability. By adjusting the GS value based on the system load, we can control the dropped probability of CP in the PQ. When the GSLB is chosen, a Node-B can not decrease the GS value for the CPs lower than it. That means the worse case of dropped probability of CP can be controlled. When the GSUB is chosen, a Node-B can not increase the GS value for the CPs higher than it. That means the best case of dropped probability of CP can be controlled.

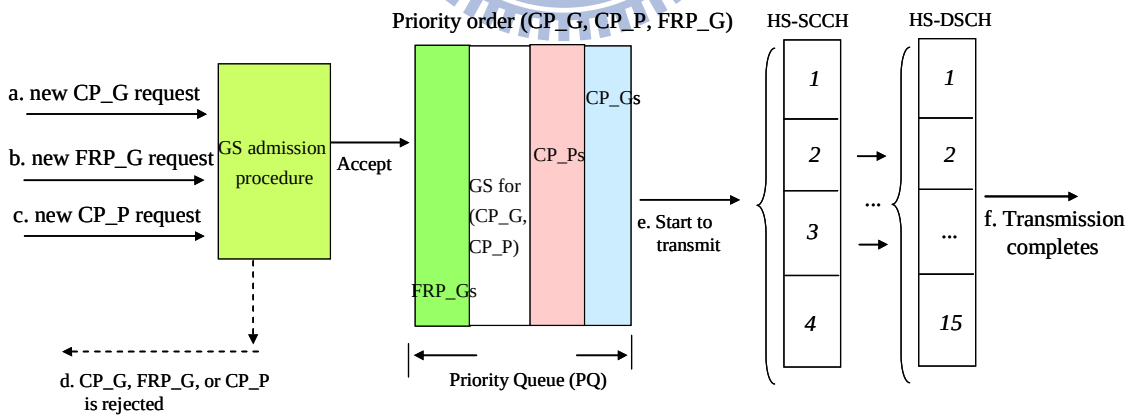


Fig. 4.9: A queueing model for DGS-PQ scheme

The proposed algorithm of dynamically adjusting *GS* implementation in Node-B can be described as follows. The DP of CP (*P1*) and FRP (*P2*) will also use in this method. When the

DP of CP is larger than *P1* and *GS* is lower than *GSUB*, then *GS* can be increased by *GSSV*. Otherwise, when the DP of FRP is larger than *P2* and *GS* is higher than *GSLB*, then *GS* can be decreased by Guard Slots Step Value (*GSSV*). For implementation consideration, a lower *P1* or a higher *GSUB* value can gurantee the throughput for CPs; on the other hand, a lower *P2* or a lower *GSLB* value can gurantee the fairness for FRP. The effects of dynamically adjusting the *GS* can gurantee the system throughput of CP even when the FRP traffic is heavy, and can gurantee the fairness for FRPs. The values of the *P1*,*P2*,*GSSV*,*GSLB*,*GSUB* and *GS* value can be chosen by mobile operators. The peudocode of the proposed algorithm is described in Fig. 4.10.

Parameters:

GS : Guard Slots of PQ for new CP arrives

SW (Slicing Window) : a piece of system processing time

PQR: free capacity of PQ

P1: the acceptable Dropped Probability (DP) of CP

P2: the acceptable Dropped Probability (DP) of FRP

GSLB: Guard Slots Lower Bound

GSUB: Guard Slots Upper Bound

GSSV: Step Value of Guard Slots in a PQ

Pseudocode

Initialize assign *GS* value and renew a timer for current *SW*

Repeat

(1) Any new downlink CP and FRP do not need to be put into PQ goto 3)

(2) GS Admission procedure

(a) if [*PQR*> =($\lfloor GS \rfloor$ +1)], PQ accepts CP and FRP

else if (*PQR* <= $\lfloor GS \rfloor$), PQ rejects CP and FRP

else if (*PQR* = $\lceil GS \rceil$)

PQ accepts CP and ($\lceil GS \rceil$ - *GS*)*100% FRPs in current *SW*

(b) *PQR* - =1 if accept CP or FRP

(3) If current SW !=0 goto (5)

(a) Recalculate the DP of CP and DP of FRP based on the numer of CPs and FRPs which accept or Discard in previous (n-1) *SW*s and current *SW*

(b) If *DP* of *CP* > *P1* and *GS* < *GSUB* , *GS* + =*GSSV* , goto (d)

(c) If DP of FRP > P2 and GS > GSLB , GS - =GSSV

d) Renew a *SW* timer , goto (1)

4) Upadte the number of CPs and FRPs which accept or Discard in current *SW*

5) Update Current *SW* and goto (1)

End Repeat

Fig. 4.10: A peudocode for DGS-PQ scheme

4.4 Analytic Models

In our analysis, The notation of the size of PQ is *B* and the number of HS-SCCH in a cell for all schemes is *C*. The arrivals of CP_G, CP_P and FRP_G form Poisson processes with mean λ_{cg} , λ_{cb} and λ_{fg} , respectively. The service time of CP_G, CP_P and FRP_G is assumed to be exponentially distributed with mean $1/\mu_{cg}$, $1/\mu_{cb}$ and $1/\mu_{fg}$, respectively. We can use the M/M/C/B Markov process to model the M-PQ, P-PQ, DDT-PQ and DGS-PQ schemes, and they are described below.

4.4.1 M-PQ Method

For M-PQ scheme, let state (*i*,*j*,*m*,*n*) denote that there are in total *i* CP_Gs and FRP_Gs transmitting, and *j* CP_Ps transmitting, in total *m* CP_Gs and FRP_Gs waiting, and *n* CP_Ps waiting in the PQ. Part of the state transition diagram of M-PQ scheme is depicted in Fig. 4.11. Let $P_{i,j,m,n}$ denote the steady-state probability of the network in state (*i*,*j*,*m*,*n*) and S_M be the set of existing states for this process. S_M can be expressed in (4.1).

$$S_M = \{(i,j,m,n) \mid [0 \le i \le C, 0 \le j \le C, m = 0, n = 0, 0 \le (i+j) \le C]\} \quad (4.1)$$
$$or [(i+j) = C, 0 \le m \le B, 0 \le n \le B, 0 \le (m+n) \le B]$$

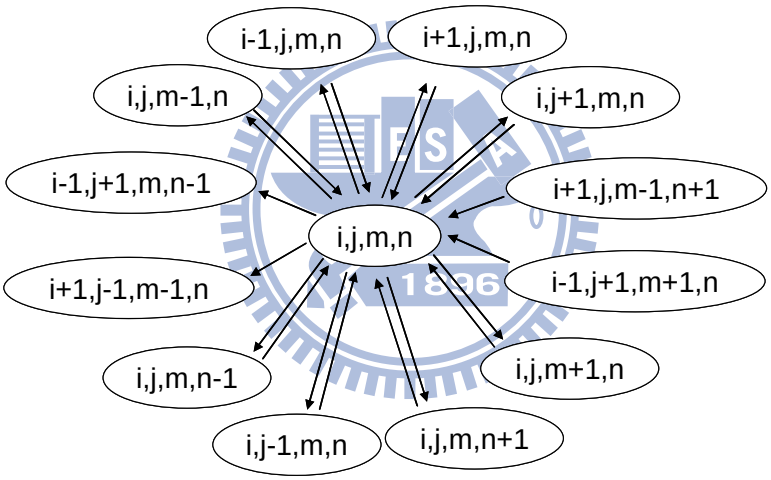


Fig. 4.11: The state transition diagram of M-PQ scheme

From the balance equations which are complicated, not shown here, and the constraint $\sum_{(i,j,m,n) \in S_M} P_{i,j,m,n} = 1$, the steady-state probability $P_{i,j,m,n}$ can be obtained by an iterative algorithm [43]. The dropped probability of CPs (CP_Gs and CP_Ps) and FRPs (P_{cp_M} and P_{frp_M}), the network utilization of CPs (CP_Gs and CP_Ps) (U_{cp_M}) can be expressed in (4.2)-(4.4), respectively.

$$P_{CP\_M} = \sum_{(i,j,m,n)\in S_M,[(i+j)=C,(m+n)=B]} P_{i,j,m,n} * \left[\frac{\lambda_{cg}}{(\lambda_{cg}+\lambda_{fg})}\right] + \sum_{(i,j,m,n)\in S_M,[(i+j)=C,(m+n)=B]} P_{i,j,m,n} \quad (4.2)$$

$$P_{frp\_M} = \sum_{(i,j,m,n)\in S_M,[(i+j)=C,(m+n)=B]} P_{i,j,m,n} * \left[\frac{\lambda_{fg}}{(\lambda_{cg}+\lambda_{fg})}\right] \quad (4.3)$$

$$U_{CP\_M} = \sum_{(i,j,m,n)\in S_M} (i * P_{i,j,m,n}) * \left[\frac{\lambda_{cg}}{(\lambda_{cg}+\lambda_{fg})}\right] + \sum_{(i,j,m,n)\in S_M} (j * P_{i,j,m,n}) \quad (4.4)$$

4.4.2 P-PQ Method

For P-PQ scheme, let state (*i*,*j*,*k*,*x*,*y*,*z*) denote that there are in total *i* CP_Gs transmitting, *j* FRP_Gs transmitting, and *k* CP_Ps transmitting, in total *x* CP_Gs waiting, *y* FRP_Gs waiting, and *z* CP_Ps waiting in the PQ. Part of the state transition diagram of P-PQ scheme is depicted in Fig. 4.12. Let $P_{i,j,k,x,y,z}$ denote the steady-state probability of the network in state (*i*,*j*,*k*,*x*,*y*,*z*) and S_P be the set of existing states for this process. S_P can be expressed in (4.5).

$$S_P = \left\{(i,j,k,x,y,z) \middle| \begin{bmatrix} 0 \le i \le C, 0 \le j \le C, 0 \le k \le C, \\ x = 0, y = 0, z = 0, 0 \le (i+j+k) \le C \end{bmatrix}\right\} \quad (4.5)$$
$$or \begin{bmatrix} (i+j+k) = C, 0 \le x \le B, 0 \le y \le B, 0 \le z \le B, \\ 0 \le (x+y+z) \le B \end{bmatrix}$$

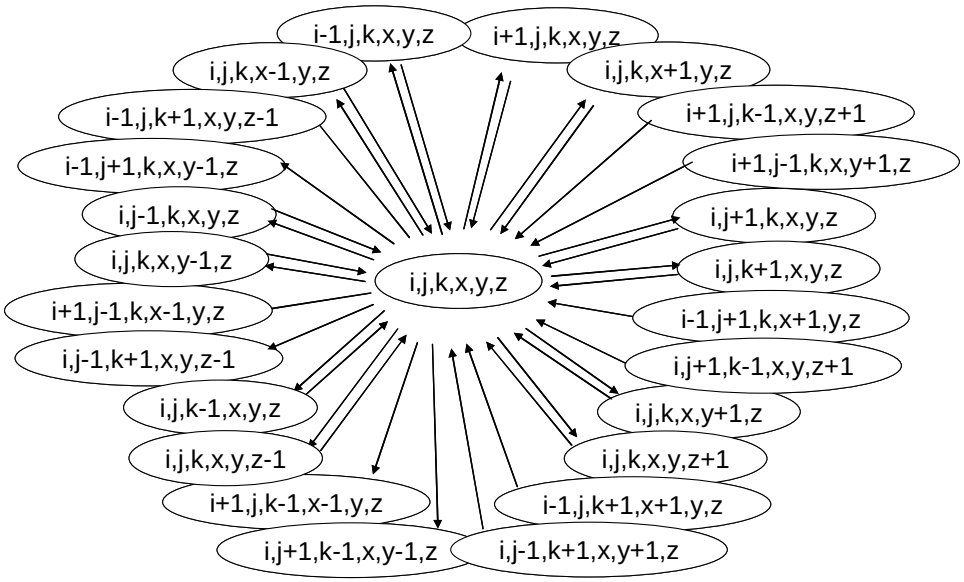


Fig. 4.12: The state transition diagram of P-PQ scheme

From the balance equations which are complicated, not shown here, and the constraint $\sum_{(i,j,m,x,y,z)\in S_P} P_{i,j,m,x,y,z} = 1$, the steady-state probability $P_{i,j,m,x,y,z}$ can be obtained by an iterative algorithm. The dropped probability of CPs (CP_Gs and CP_Ps) and FRP_Gs (P_{cg_P} and P_{frp_P}), the network utilization of CPs (CP_Gs and CP_Ps) (U_{cp_P}) can be expressed in (4.6)-(4.8), respectively.

$$P_{cp\_P} = \sum_{(i,j,k,x,y,z)\in S_p,[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} + \sum_{(i,j,k,x,y,z)\in S_p,[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} \quad (4.6)$$

$$P_{frp\_P} = \sum_{(i,j,k,x,y,z)\in S_p,[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} \quad (4.7)$$

$$U_{cp\_P} = \sum_{(i,j,k,x,y,z)\in S_p} \left[(i+k)* P_{i,j,k,x,y,z}\right] \quad (4.8)$$

4.4.3 DDT-PQ Method

For DDT-PQ scheme, let state (*i*,*j*,*k*,*x*,*y*,*z*) denote that there are in total *i* CP_Gs transmitting, *j* FRP_Gs transmitting, and *k* CP_Ps transmitting, in total *x* CP_Gs waiting, *y* FRP_Gs waiting, and *z* CP_Ps waiting in the PQ. Part of the state transition diagram of DDT-PQ scheme is depicted in Fig. 4.13. The difference of the state transition diagram and that of P-PQ is that there are two extra dotted lines representing the operations of *DT* for FRP_Gs. For example, state (*i*,*j*,*k*,*x*,*y*,*z*) may change to state (*i*,*j*,*k*,*x*,*y-1*,*z*) if there is a FRP_G's *DT* expires and the FRP_G is dropped from the PQ. The *DT* of FRP_G is assumed to be exponentially distributed with mean $1/\mu_{dt}$. Let $P_{i,j,k,x,y,z}$ denote the steady-state probability of the network in state (*i*,*j*,*k*,*x*,*y*,*z*) and S_{DDT} be the set of existing states for this process. S_{DDT} can be expressed in (4.9)

$$S_{DDT} = \left\{ (i,j,k,x,y,z) \middle| \begin{bmatrix} 0 \le i \le C, 0 \le j \le C, 0 \le k \le C, \\ x = 0, y = 0, z = 0, 0 \le (i+j+k) \le C \end{bmatrix} \right\} \quad (4.9)$$
$$or \begin{bmatrix} (i+j+k) = C, 0 \le x \le B, 0 \le y \le B, 0 \le z \le B, \\ 0 \le (x+y+z) \le B \end{bmatrix}$$

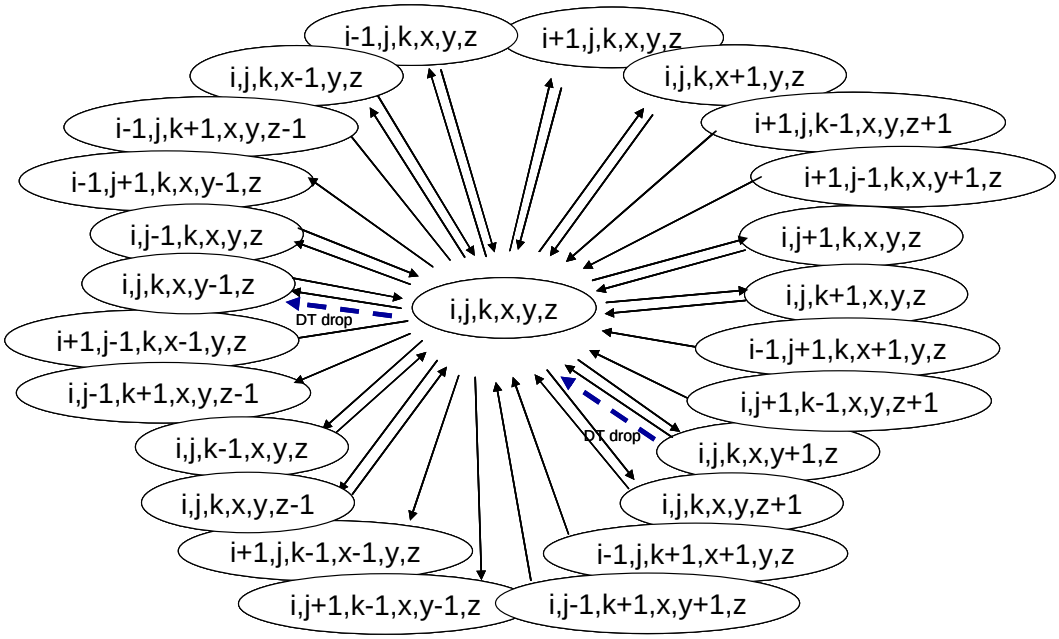


Fig. 4.13: The state transition diagram of DDT-PQ scheme

From the balance equations which are complicated, not shown here, and the constraint $\sum_{(i,j,m,x,y,z)\in S_{DDT}} P_{i,j,m,x,y,z} = 1$, the steady-state probability $P_{i,j,m,x,y,z}$ can be obtained by an iterative algorithm. The dropped probability of CPs (CP_Gs and CP_Ps) and FRP_Gs (P_{cp_DDT} and P_{frp_DDT}), the network utilization of CPs (CP_Gs and CP_Ps) (U_{cp_DDT}) can be expressed in (4.10)-(4.12), respectively.

$$P_{cp\_DDT} = \sum_{(i,j,k,x,y,z)\in S_{DDT},[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} + \sum_{(i,j,k,x,y,z)\in S_{DDT},[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} \quad (4.10)$$

$$P_{frp\_DDT} = \sum_{(i,j,k,x,y,z)\in S_{DDT},[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} + \sum_{(i,j,k,x,y,z)\in S_{DDT}} \left[\frac{y*\mu_{dt}}{(y*\mu_{dt} + i*\mu_{cg} + j*\mu_{fg} + k*\mu_{cb})} \right] * P_{i,j,k,x,y,z} \quad (4.11)$$

$$U_{cp\_DDT} = \sum_{(i,j,k,x,y,z)\in S_{DDT}} [(i+k)* P_{i,j,k,x,y,z}] \quad (4.12)$$

4.4.4 DGS-PQ Method

For DGS-PQ scheme, let state (*i*,*j*,*k*,*x*,*y*,*z*) denote that there are in total *i* CP_Gs transmitting, *j* FRP_Gs transmitting, and *k* CP_Ps transmitting, in total *x* CP_Gs waiting, *y* FRP_Gs waiting, and *z* CP_Ps waiting in the PQ. Part of the state transition diagram of DGS-PQ scheme is depicted in Fig. 4.14. The difference of the state transition diagram

and that of P-PQ is that there are two extra dotted lines representing the operations of *GS* for CP_Gs. For example, state (*i,j,k,x,y,z*) may not change to state (*i,j,k,x,y+1,z*) because the new FRP_G request can not be put into the PQ based on *GS* admission procedure even the PQ still has free capacity and should be dropped. Let $P_{i,j,k,x,y,z}$ denote the steady-state probability of the network in state (*i,j,k,x,y,z*) and S_{DGS} be the set of existing states for this process. S_{DGS} can be expressed in (4.13).

$$S_{DGS}\left\{(i,j,k,x,y,z)\left|\begin{bmatrix}0\le i\le C,0\le j\le C,0\le k\le C,\\ x=0,y=0,z=0,0\le(i+j+k)\le C\end{bmatrix}\right.\right\}$$
$$or\begin{bmatrix}(i+j+k)=C,0\le x\le B,0\le y\le B,0\le z\le B,\\ 0\le(x+y+z)\le B\end{bmatrix} \quad (4.13)$$

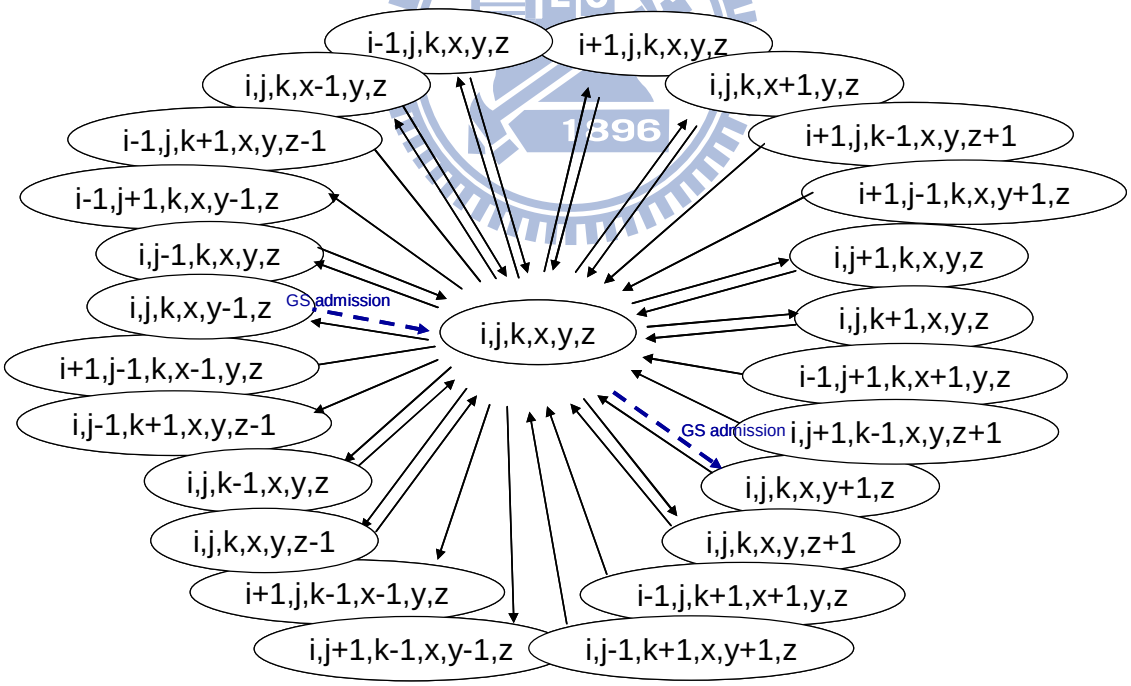


Fig. 4.14: The state transition diagram of DGS-PQ scheme

From the balance equations which are complicated, not shown here, and the constraint $\sum_{(i,j,m,x,y,z)\in S_{DGS}} P_{i,j,m,x,y,z} = 1$, the steady-state probability $P_{i,j,m,x,y,z}$ can be obtained by an iterative algorithm. The dropped probability of CPs (CP_Gs and CP_Ps) and FRP_Gs (P_{cp_DGS} and P_{frp_DGS}), the network utilization of CPs (CP_Gs and CP_Ps) (U_{cp_DGS}) can be expressed in (4.14)-(4.16), respectively.

$$P_{cp\_DGS} = \sum_{(i,j,k,x,y,z)\in S_{DGS},[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} + \sum_{(i,j,k,x,y,z)\in S_{DGS},[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} \quad (4.14)$$

$$P_{frp\_DGS} = \sum_{(i,j,k,x,y,z)\in S_{DGS},[(i+j+k)=C,(x+y+z)=B]} P_{i,j,k,x,y,z} + \sum_{(i,j,k,x,y,z)\in S_{DGS},[\text{GS admission procedure}]} P_{i,j,k,x,y,z} \quad (4.15)$$

$$U_{cp\_DGQ} = \sum_{(i,j,k,x,y,z)\in S_{DGQ}} \left[(i+k)* P_{i,j,k,x,y,z}\right] \quad (4.16)$$

4.5 Cost Function Scheme

In this chapter, we consider a mobile data operator's revenue that consists of the transmission fee of normal users and the monthly fee of flat-rate users. Instead of calculating the total revenue, we propose a cost function representing the revenue loss due to blocked CPs and due to the loss of flat-rate users. Since CPs are charged by the volume of packets transmitted. In a fully utilized network, re-transmitting blocked CPs only leads to more CPs

blocked. Therefore, we assume that blocked CPs in a fully utilized network will not be re-transmitted, and thus represent revenue loss. The revenue loss of blocked CPs is proportional to the blocking probabilities (P_{CP_G} and P_{CP_P}) and the traffic load of CPs (ρ_{CP_G} and ρ_{CP_P}). The monthly revenue loss due to blocked CPs can be expressed in (4.17).

$$C_{CP} = \rho_{CP\_G} \cdot P_{CP\_G} + \rho_{CP\_P} \cdot P_{CP\_P} \qquad (4.17)$$

The revenue loss due to the loss of flat-rate users also depends on the blocking probability. Since flat-rate users are not charged by the volume of packet transmission, blocked FRPs do not result in direct revenue loss. However, when the blocking probability is above a departure threshold, β, flat-rate users may become discontent and start to switch to other operators. We assume that the number of flat-rate users lost per month is proportional to the discrepancy of the blocking probability (P_{FRP_G}) above β. The monthly revenue loss due to lost flat rate users can be expressed in (4.18), where α represents the cost weighting factor of flat-rate users.

$$C_{FRP} = \begin{cases} \alpha \cdot (P_{FRP\_G} - \beta), if\ (P_{FRP\_G} > \beta) \\ 0, otherwise \end{cases} \qquad (4.18)$$

The total monthly loss (C) is C_{CP} plus C_{FRP} . And we assume the the traffic load of CPs (ρ_{CP_G} and ρ_{CP_P}) are ρ . We obtain the cost function, as shown in (4.19).

$$C = C_{CP} + C_{FRP} = \begin{cases} \rho \cdot (P_{CPP\_G} + P_{CPP\_P}) + \alpha \cdot (P_{FRP\_G} - \beta), if\ (P_{FRP\_G} > \beta) \\ \rho \cdot (P_{CPP\_G} + P_{CPP\_P}), otherwise \end{cases} \qquad (4.19)$$

When the cost weighting factor of FRPs is less than that of CPs ($\alpha < \rho$), the scheduler

should give priority to CPs without considering the FRP blocking probability. On the other hand, when α is larger than twice of ρ, the scheduler should give priority to CPs when the FRP blocking probability is below the departure threshold (β), but it should give priority to FRPs when FRP blocking probability is above β. Note that β should be chosen to reflect the beginning of user dissatisfaction as the blocking probability increases; its proper value may be obtained from the past operation data.

To minimize the cost function, a same iterative algorithm shall also be developed as shown in Fig. 3.5. Based on the stationary state probabilities, we can obtain the performance measures and the minimum value of the cost function.

4.6 Numreical Analysis

4.6.1 Case I : $\alpha < \rho$

In the analysis below, we assume the the number of HS-SCCH in a cell (C) is 4 and the size of the PQ (B) is 20. We compare four packet scheduling methods: M-PQ, P-PQ、DDT-PQ and DGS-PQ schemes. The mean service time of a CP_G ($1/\mu_{cg}$) and a FRP_G ($1/\mu_{fg}$) are assumed to be 10 ms, the mean service time of a CP_P ($1/\mu_{cb}$) is assumed to be 20ms. The session arrival rate of CP_G (λ_{cg}) and CP_P (λ_{cb}) are fixed at 50 and 100 packets/second respectively, and that of FRP_G (λ_{fg}) varies in the range of 50-150 packets/second.

In case 1, we let the cost weighting factor of flat-rate users α = 2500 and β should be chosen to be 0.08 to reflect the level of user dissatisfaction. Note that ρ is choosen to be 5000 in our experiments, i.e., the cost weighting factor of FRPs is less than that of CPs.

Extra parameters are needed to be determined in DDT-PQ scheme. *P1* and *P2* are set to be 2% and 3%, *DTLB* and *DTUB* are set between 0.2 and 0.6 seconds, initial *DT* and

DTSV value are set to be 0.4 seconds and 10ms. Extra parameters are also needed to be assumed in DGS-PQ scheme. *P1* and *P2* are set to be 2% and 3% same with the DDT-PQ scheme, the size of *GSLB* and *GSUB* can be reserved for CPs are set between 1 and 3, initial *GS* and *GSSV* value are set to be 2 and 0.1.

Fig. 4.15 plots the mean dropped probabilities of CPs (i.e. CP_G and CP_P) for four schemes as the FRP arrival rates increases. With special treatment for FRPs in DDT-PQ and DGS-PQ, these two schemes are trying to keep the dropped probability of CPs is below *P1* when the arrival rate of FRP is high. The effects of DDT-PQ and DGS-PQ schemes on the mean dropped probability of CPs are changed slowly when the FRP arrival rate is higher than 125 packets/second. This is because in DDT-PQ scheme there is a dynamic *DT* for every FRP to limit the waiting time for it in PQ and in DGS-PQ scheme there is a dynamic *GS* of PQ can be reserved for CPs to let them have more chance to be kept in PQ.

When the FRP arrival rate is 170 packets/second, the *DT* value reachs *DTLB* in DDT-PQ scheme. That means there is no more adjustment of *DT* for FRPs even though the dropped probability of CPs is higher than *P1*. The effect is that the dropped probability of CPs will be higher than *P1* when the FRP arrival rate is higher than 170 packets/second. This indicates that *DTLB* in DDT-PQ scheme can also be dynamically adjust in system implementation consideration. When the FRP arrival rate is 170 packets/second, the *GS* value does not reach *GSUB* in DGS-PQ scheme. That means the adjustment of *GS* for CPs can still control the dropped probability of CPs higher than *P1*. According to the results, DDT-PQ and DGS-PQ schemes can have lower dropped probability of CPs comparing with two other schemes.

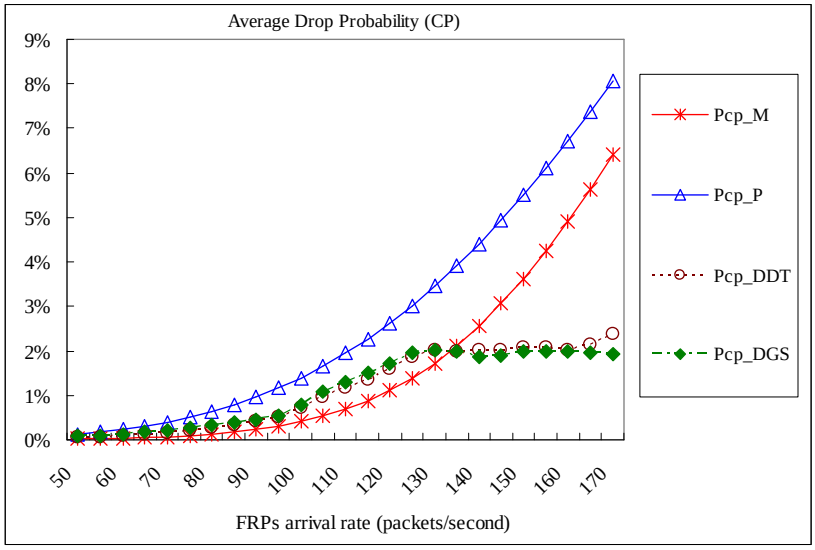


Fig. 4.15: The average dropped probability of CP for four packet scheduling methods

Fig.4.16 plots the mean dropped probabilities of FRPs (i.e. FRP_G) for four schemes as the FRP arrival rates increases. With the special treatment for FRPs in DDT-PQ and DGS-PQ, these two schemes will have higher blocking probabilities for FRUSs. The cost becomes more significant as the traffic of FRPs increases. However, in these two DDT-PQ and DGS-PQ schemes, they are also trying to keep the dropped probability of FRPs is below *P2* when the FRP arrival rate is between 100 and 105 packets/second. But after the FRP arrival rate icreases more, trying to keep the dropped probability of CPs below *P1* is important than trying to keep the dropped probability of FRPs below *P2*.

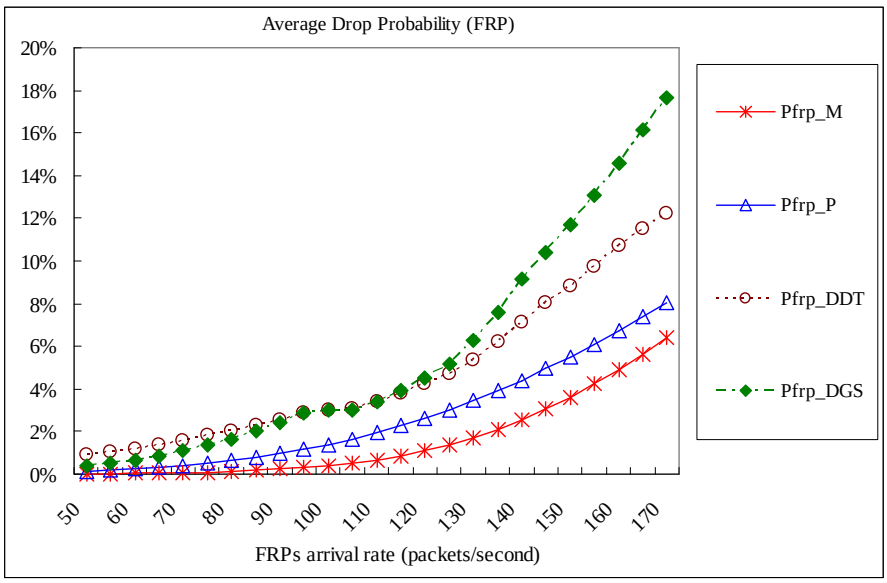


Fig. 4.16: The average dropped probability of FRP for four packet scheduling methods

Fig. 4.17 plots the mean network utilization of CPs (i.e. CP_G and CP_P) for four schemes as the FRP arrival rates increases. With the special treatment for FRPs in DDT-PQ and DGS-PQ, these two schemes are trying to let CPs to have more changes to be served even when the arrival rate of FRP is high. When the FRP arrival rate is 170 packets/second, the *DT* value reachs *DTLB* in DDT-PQ scheme. The effect in DDT-PQ scheme is that the network utilization of CPs are almost the same and can be guranteed when the FRP arrival rate is between 125 and 165 packets/second. When the FRP arrival rate is 170 packets/second, the *GS* value does not reach *GSUB* in DGS-PQ scheme. The effect in DGS-PQ scheme is that the network utilization of CPs are almost the same and can be guranteed when the FRP arrival rate is between 125 and 170 packets/second. Because we choose different simulation parameters assumptions (i.e., *DTLB* and *GSUB*), there is a little different result between these two methods. No matter what parameters we choose, the result can implicitly indicate that these two methods are very effective to gurantee the packet revenues for mobile operators.

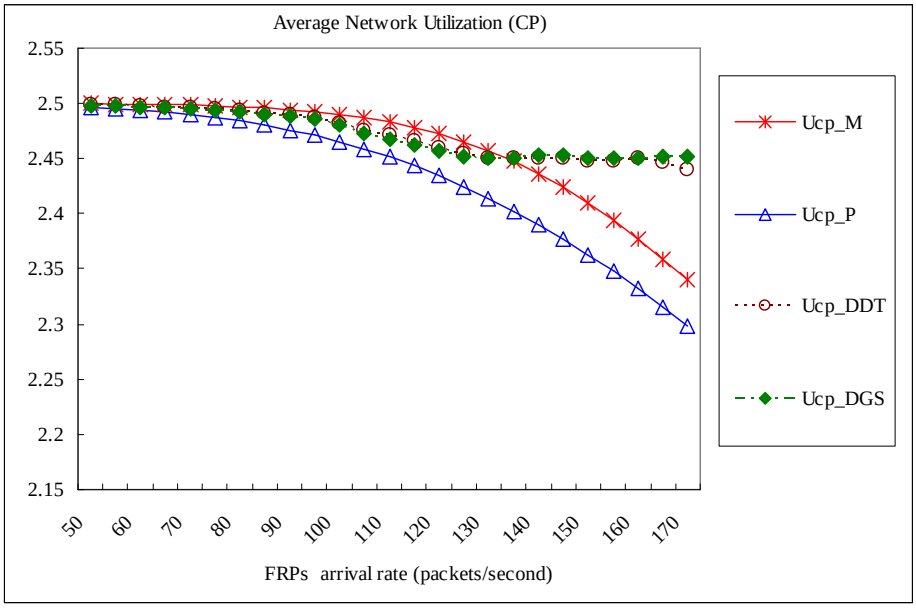


Fig. 4.17: The average network utilization of CP for four packet scheduling methods

Fig. 4.18 plots the result of dynamically adjusting the *DT* value in DDT-PQ scheme as the FRP arrival rates increases. We assume the initial *DT* value is 400ms and *DTLB* , *DTUB* values are 200ms, 600ms. When the FRP arrival rate is 100 packets/second, the dropped probabilities of FRPs reachs *P2*, the *DT* needs to increase to keep the dropped probability of FRP below *P2*. When the FRP arrival rate is 105 packets/second, the *DT* reachs *DTUB* and no more adjustment of *DT* in DDT-PQ scheme. That means the dropped probabilities of FRPs is higher than *P2*. When the FRP arrival rate is 130 packets/second, the dropped probabilities of CPs reachs *P1*, the *DT* needs to decrease to keep the dropped probability of CP below *P1*. When the FRP arrival rate is 165 packets/second, the *DT* reachs *DTLB* and no more adjustment of *DT* in DDT-PQ scheme. That means the dropped probabilities of CPs is higher than *P1*. Different analytic parameters assumptions, i.e., *P1*, *P2*, *DT*, *DTLB* and *DTUB* values in DDT-PQ scheme could be a little different numerical results.

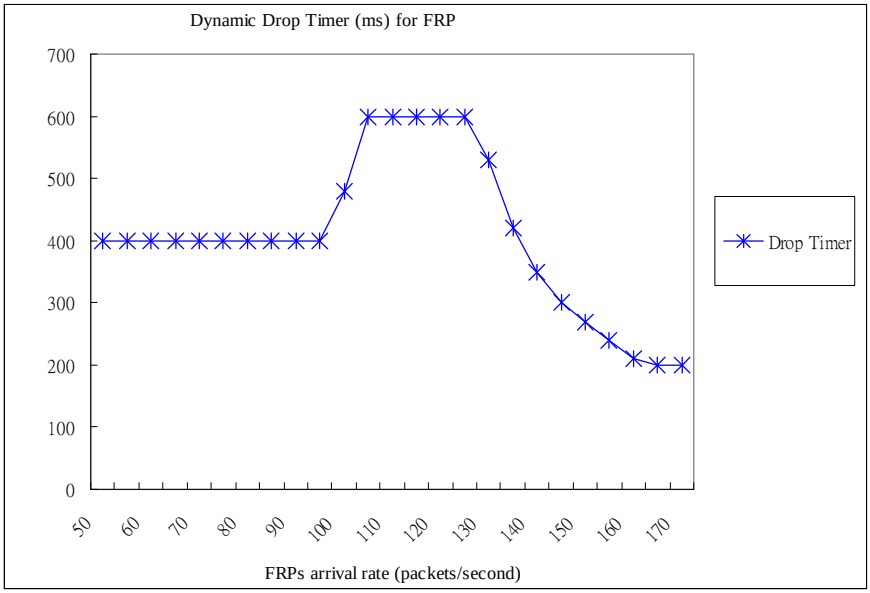


Fig. 4.18: The results of dynamically adjusting the DT value in DDT-PQ scheme

Fig. 4.19 plots the results of dynamically adjusting the *GS* value in DGS-PQ scheme as the FRP arrival rates increases. We assume the initial *GS* value is 2 and *GSLB* , *GSUB* values are 1, 3. When the FRP arrival rate is 100 packets/second, the dropped probabilities of FRPs reachs *P2*, the *GS* needs to decrease to keep the dropped probability of FRP below *P2*. When the FRP arrival rate is 110 packets/second, the *GS* reachs *GSLB* and no more adjustment of *GS* in DGS-PQ scheme. That means the dropped probabilities of FRPs is higher than *P2*. When the FRP arrival rate is 130 packets/second, the dropped probabilities of CPs reachs *P1*, the *GS* needs to increase to keep the dropped probability of CP below *P1*. When the FRP arrival rate is 170 packets/second, the *DT* does not reach G*SUB*. That means the dropped probabilities of CPs is not higher than *P1*. Different analytic parameters assumptions, i.e., *P1*, *P2*, *GS*, *GSLB* and *GSUB* values in DGS-PQ scheme could be a little different numerical results.

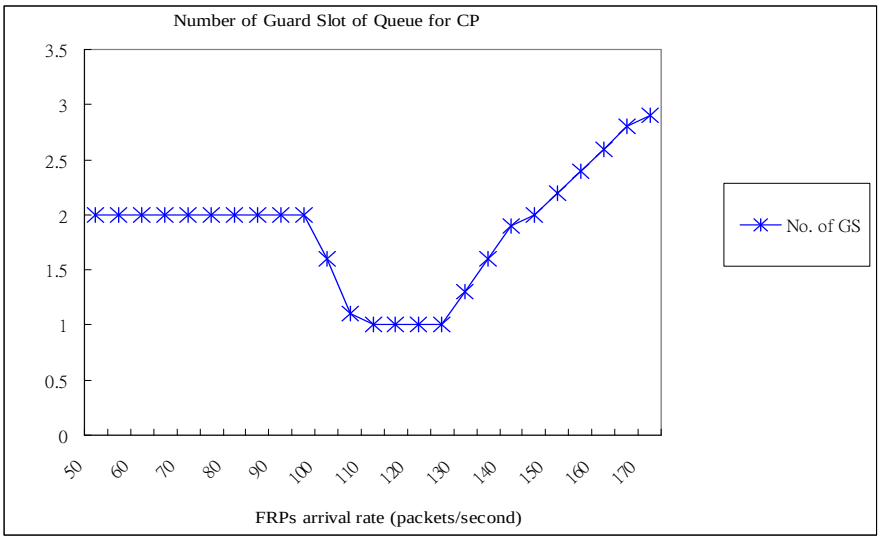


Fig. 4.19: The results of dynamically adjusting the GS value in DGS-PQ scheme

4.6.2 Case II : $\alpha > \rho$

In case 2, we let the cost weighting factor of flat-rate users α = 20000 and β . should be chosen to be 0.08 to reflect the level of user dissatisfaction. Note that ρ is choosen to be 5000 in our experiments, i.e., the cost weighting factor of FRPs is much larger than that of CPs.

Fig. 4.20 plots the cost function as FRP traffic increases. The cost function represent the revenue loss of the operator; the less the better. A more sophisticated scheduler is developed in this case ($\alpha > \rho$). DGS-PQ and DDT-PQ scheduler give priority to CPs when the FRP blocking probability is below the departure threshold, β. When FRP blocking probability is above the threshold, DGS-PQ and DDT-PQ scheduler give priority to FRPs.

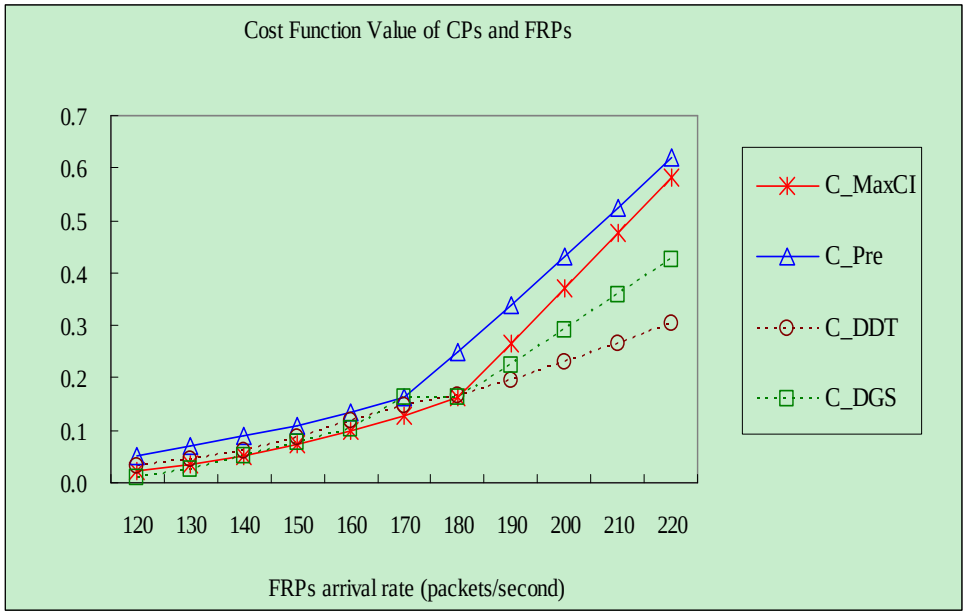


Fig. 4.20: The cost function (C) when $\alpha > \rho$

The results indicate DGS-PQ and DDT-PQ have the least amount of revenue loss among all schedulers. When the FRP traffic less than 180 packets/sec, the revenue losses of two other schedulers are as small as those of DGS-PQ and DDT-PQ, but the losses rise rapidly as the FRP traffic increases above 180 packets/sec.

In DGS-PQ, the reserved GS is for CPs when the FRP blocking probability is below the departure threshold, β when the FRP traffic less than 180 packets/sec. and the reserved GS is for FRPs when the FRP blocking probability is higher the departure threshold, β when the FRP traffic larger than 180 packets/sec.. The number of GS for CPs and FRPs depend on the cost function value and is depicted at Fig. 4.21. We assume the initial *GS* value is 4 and *GSLB* , *GSUB* values are 0, 4.

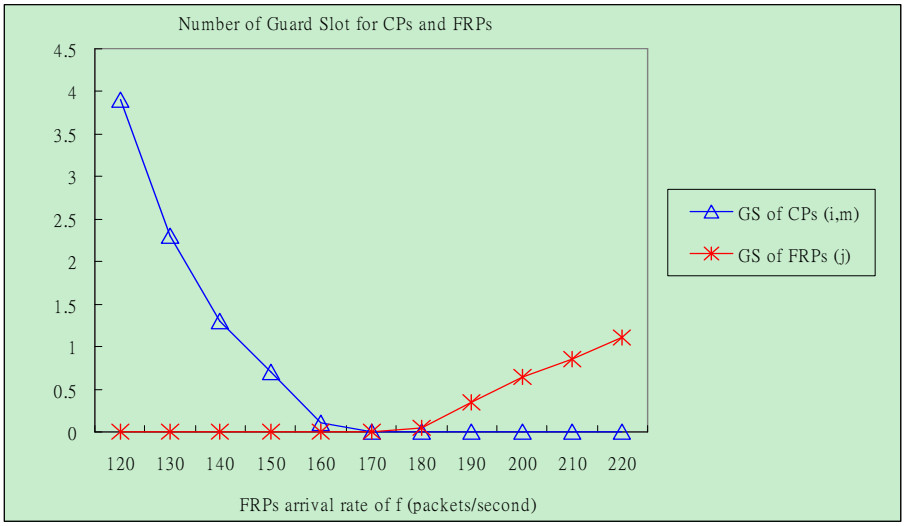


Fig. 4.21: The number of GS for CPs and FRPs when $\alpha > \rho$

In DDT-PQ, the drop timer will set for each packets of FRPs when the FRP blocking probability is below the departure threshold, β when the FRP traffic less than 180 packets/sec and the drop timer will set for each packets of CPs when the FRP blocking probability is higher the departure threshold, β when the FRP traffic less than 180 packets/sec. The number of DT for FRPs and CPs depend on the cost function value and is depicted at Fig. 4.22. We assume the initial *DT* value is 500ms and *DTLB* , *DTUB* values are 500ms,5000ms.

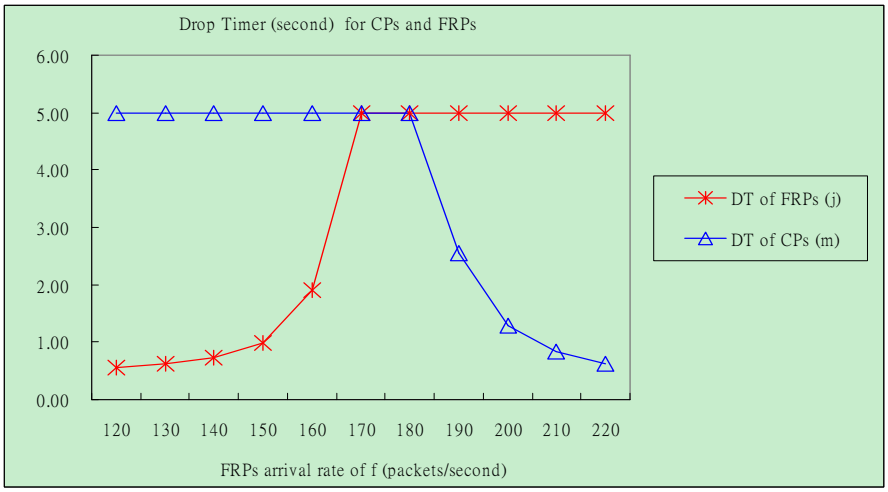


Fig. 4.22: The value of DT for FRPs and CPs when $\alpha > \rho$

Fig. 4.23 plots the blocking probabilities of FRPs as the FRP arrival rate increases. The results indicate DGS-PQ and DDT-PQ can keep the blocking probabilities of FRPs at the same level(i.e., the departure threshold, β) because we adjust the value of GS or DT in DGS-PQ and DDT-PQ for FRPs or CPs. In order to get the minimum value of cost function, the value of GS and DT will be dynamically adjusted.

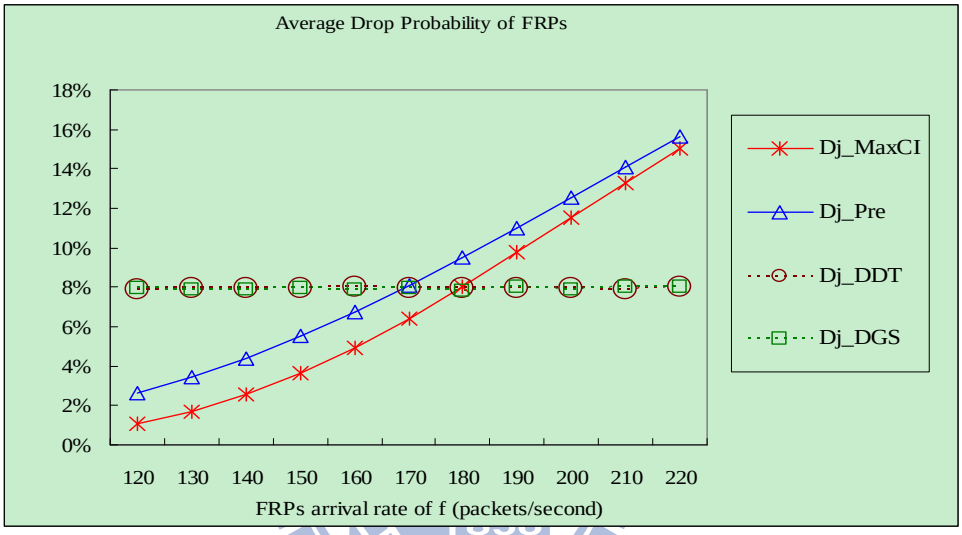


Fig. 4.23: The blocking probabilities of FRPs when $\alpha > \rho$

4.7 Conclusions

Since usage does not incur cost, FRPs data packets may occupy most of the radio channel resources in a HSDPA network. The effects of controlling downlink packet scheduling parameters in HSDPA network, such as SIR, packet priority, Discard Timer (*DT*), Guard Slots (GS) in a PQ for CPs and FRPs have been studied. The analytic models for four packet scheduling schemes, M-PQ, P-PQ, DDT-PQ and DGS-PQ, have been presented and numeric results have been discussed. In our study, M-PQ and P-PQ methods combining downlink packet scheduling techniques studied in this paper, including queueing and priority,

are not suitable for mobile operators to provide flat-rate packet services to users if they care of revenue.

Moreover, DDT-PQ and DGS-PQ methods consider the balance between serving the FRPs and CPs in different system loads no matter $\alpha < \rho$ or $\alpha > \rho$. They combine downlink packet scheduling control techniques studied in this paper, including queueing, priority and dynamic *DT* for FRPs or dynamic *GS* of PQ for CPs, are effective in reducing the blocking probability for CPs at the cost of increasing the blocking probability for FRPs when $\alpha < \rho$, and keep the blocking probability at the same level for FRPs at the cost of increasing the blocking probability for CPs when $\alpha > \rho$. These two methods are much effective to guarantee the revenues for mobile operators especially when system load is high.

CHAPTER 5

Conclusions and Future Work

In this dissertation, we investigated design issues on the scheduling mechanisms for priority transmission in PLMNs. This chapter summaries our study and contributions, and briefly discusses the future directions.

5.1 Summary

In Chapter 2, channel allocation methods for priority packets in the GPRS Network were presented. We studied and described BCA and USFCA schemes that implement priority queues in the GPRS network. Both schemes provide shorter mean waiting time and system time for priority packets than the simple FCFS scheme at the cost of longer mean waiting time and system time for non-priority packets. In addition, the transmission delay of priority packets using USFCA can be better guaranteed than that of BCA especially when the GPRS traffic is heavy.

In Chapter 3, uplink connection scheduling methods for flat-rate data services in the UMTS network were investigated. We described four combinations of scheduling techniques, queueing, guard channels, preemption and rate-adaptation. Moreover, we analyzed their effectiveness in reducing the revenue loss of the operators serving both normal and flat-rate users. We proposed a cost function representing the revenue loss due to both blocked normal user connections and lost flat-rate users. The optimum numbers of guard channels can be determined by an iterative algorithm. The analytic results indicate when α, the cost

weighting factor of flat-rate users, is less than ρ_n, queueing and preemption are essential for connection scheduling to maximize the revenue. Rate-adaptation is ineffective, because half-rate connections are less bandwidth-efficient. Sub-rating FRUCs reduced the system throughput and the operator revenue. In addition, no guard channel is needed, if queuing and preemption are used, because guard channels increase the blocking probability of FRUCs and reduces system throughput. We considered uplink connection scheduling only. We did not study downlink traffic scheduling, which can be done on packet level. In our study, the cost weighting factor of flat-rate users (α), is less than that of normal users (ρ_n). Further study is needed for UMTS networks with α larger than ρ_n, which is possible when the number of flat-rate users increases or the normal user traffic decreases. In this situation, a more sophisticated scheduler is needed. The scheduler should give priority to NUCs when the FRUC blocking probability is below the departure threshold, β. When FRUC blocking probability is above the threshold, FRUCs should have priority.

In Chapter 4, flate-rate packet scheduling techniques for the WCDMA systems with HSDPA were presented. Since usage does not incur cost, FRPs data packets may occupy most of the radio channel resources in a HSDPA network. The effects of controlling downlink packet scheduling parameters in HSDPA network, such as SIR, packet priority, Discard Timer (DT), Guard Slots (GS) in a PQ for CPs and FRPs have been studied. The analytic models for four packet scheduling schemes, M-PQ, P-PQ, DDT-PQ and DGS-PQ, have been presented and numeric results have been discussed. In our study, M-PQ and P-PQ methods combining downlink packet scheduling techniques studied in this paper, including queueing and priority, are not suitable for mobile operators to provide flat-rate packet services to users if they care of revenue.

Moreover, DDT-PQ and DGS-PQ methods consider the balance between serving the FRPs and CPs in different system loads no matter $\alpha < \rho$ or $\alpha > \rho$. They combine downlink packet scheduling control techniques studied in this paper, including queueing,

priority and dynamic DT for FRPs or dynamic GS of PQ for CPs, are effective in reducing the blocking probability for CPs at the cost of increasing the blocking probability for FRPs when $\alpha < \rho$, and keep the blocking probability at the same level for FRPs at the cost of increasing the blocking probability for CPs when $\alpha > \rho$. These two methods are much effective to guarantee the revenues for mobile operators especially when system load is high.

5.2 Future Works

Based on the research results in this dissertation, the following design issues on the scheduling mechanisms for priority transmissions in wireless network can be investigated further.

(1) Integrated uplink and downlink scheduling mechanisms for priority transmission in wireless networks

In this dissertation, we only consider one way scheduling mechanisms (i.e., uplink or downlink) to address the QoS issues of different customers and the revenues of wireless operators. To provide a better service to the users, one should consider both uplink and downlink scheduling mechanisms in providing priority services in wireless networks.

(2) To propose a new cost function in evaluating the performance of scheduling mechanisms

In Chapter 3, we consider a mobile data operator's revenue that consists of the transmission fee of normal users and the monthly fee of flat-rate users. Instead of calculating the total revenue, we propose a cost function representing the revenue loss due to blocked NUCs and due to the loss of flat-rate users. In reality, multiple QoS requirements of users may be provided by wireless operators. A new cost function considers multiple parameters is needed in evaluation the performance of scheduling methods.

Furthermore, there are strong demands from the customers for the operators to support QoS in wireless networks. Different scheduling mechanisms should be provided by the wireless operators in their business models.

Bibliography

[1] Phone Lin and Yi-Bing Lin, “Channel Allocation for GPRS”, IEEE Trans. Vehicular Technology, vol. 50, no. 2, pp. 375-387, Mar. 2001.

[2] Wei-Yeh Chen, Jean-Lien C.Wu, and Li Liann Lu, “Performances Comparsions of Dynamic Resource Allocation With/Without Channel De-Allocation in GSM/GPRS Networks”, IEEE Comm. Lett., vol. 7, no. 1, pp. 10-12, Jan. 2003.

[3] Karann Chew, Rahim Tafazolli, “Performance Analysis for GPRS Prioritized and Non-prioritized Mobility Management Procedures”, 3G Mobile Comm. Technologies Conf. Publication, no. 489, pp. 544-549, May IEE 2000.

[4] “An approach to graphs of linear forms (Unpublished work style),” unpublished. “General Packet Radio Service (GPRS); Mobile Station (MS) – Base Station System (BSS) interface; Radio Link Control / Medium Access Control (RLC/MAC) protocol”, 3GPP TS 04.60, version 7.10.0, Release 1998.

[5] Y.-B.Lin, “Performance Modeling for Mobile Telephone Networks”, IEEE Network Mag., vol.11, pp.63-68, Nov./Dec. 1997.

[6] 3GPP TS 23.002, “Network Architecture”, version 3.6, Release 1999.

[7] Y.-B. Lin, S. Mohan, and A. Noerpel, “Queueing priority channel assignment strategies for handoff and initial access for a PCS network”, IEEE Trans. Veh. Technol., vol. 43, no. 3, pp. 704712, 1994.

[8] D. Hong and S.S. Rappaport, “Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures”, IEEE Trans. Vehicular Technology, vol. 35, no. 3, pp. 77-92, Aug. 1986.

[9] R. Guerin, “Queuing-Blocking System with Two Arrival Streams and Guard Channels”, IEEE Trans. Comm., vol. 36, no. 2, pp.153- 163, Feb. 1998.

[10] Q.-A. Zeng, K. Mukumoto, and A. Fukuda, “Performance Analysis of Mobile Cellular Radio System with Priority Reservation Handoff Procedures”, Proc. IEEE VTC-94, vol. 3, pp. 1829- 1833, June 1994.

[11] Q.-A. Zeng and D.P. Agrawal, “Performance Analysis of a Handoff Scheduler in Integrated Voice/Data Wireless Networks”, Proc. IEEE VTC-2000, pp. 1986-1992, Sept. 2000.

[12] C. W. Leong, W. Zhuang, Y. Cheng, and L. Wang, “Call Admission Control for Integrated On/Off Voice and Best-Effort Data Services in Mobile Cellular Communications”, IEEE Trans. Comm., vol. 52, no. 5, pp.778- 790, May. 2004.

[13] J. Wang, Q.A. Zeng, and D. P. Agrawal, “Performance Analysis of a Preemptive and Priority Reservation Handoff Scheduler for Integrated Service-Based Wireless Mobile Networks”, IEEE Trans. on Mobile Computing., vol. 2, no. 1, pp. 65-75 Jan-Mar. 2003.

[14] M.S. Do, Y. Park, and J.Y. Lee, “Channel Assignment With QoS Guarantees for a Multiclass Multicode CDMA System”, IEEE Trans. Vehicular Technology, vol. 51, no. 5, pp. 935-948, Sep. 2002.

[15] S. Kim and P. K. Varshney,” An Integrated Adaptive Bandwidth-Management Framework for QoS-Sensitive Multimedia Cellular Networks”, IEEE Trans. Vehicular Technology, vol. 53, no. 3, pp. 847-864, May. 2004.

[16] Y.-B. Lin, A. Noerpel, and D. Harasty, “The Sub-Rating Channel Assignment Strategy for PCS Hand-Offs”, IEEE Trans. Vehicular Technology, vol. 45, no. 2, pp. 122-130, Feb. 1996.

[17] Wei-Yeh Chen, Jean-Lien C.Wu, and Li Liann Lu, “Performances Comparisons of Dynamic Resource Allocation With/Without Channel De-Allocation in GSM/GPRS Networks”, IEEE Comm. Lett., vol. 7, no. 1, pp. 10-12, Jan. 2003.

[18] 3GPP TS 25.413,“UTRAN Iu interface RANAP signaling”, version 3.14, Release 1999.

[19] 3GPP TS 25.331,“Radio Resource Control (RRC) protocol specification”, version 3.21, Release 1999.

[20] 3GPP TS 23.060,“General Packet Radio Service (GPRS) Service description; Stage 2”, version 3.a.0, Release 1999.

[21] D. Niyato, and E. Hossain, “Call-Level and Packet-Level Quality of Service and User Utility in Rate-Adaptive Cellular CDMA Networks: A Queuing Analysis”, IEEE Trans. on Mobile Computing., vol. 5, no. 12, pp. 1749-1763 Dec. 2006.

[22] L. Xu, X. Shen, and J.W. Mark, “Dynamic Fair Scheduling with QoS Constraints in Multimedia Wideband CDMA Cellular Networks”, IEEE Trans. Wireless Comm., vol. 3, no. 1, pp. 60-73, Jan. 2004.

[23] J. Laiho and A. Wacker, “Radio Network Planning Process and Methods for WCDMA”, Annals de Telecomm., p. 56, 2000.

[24] 3GPP TS 25.104,“BS Radio transmission and Reception (FDD)”, version 3.13, Release 1999.

[25] Wolff, R. W. 1982. Poisson Arrivals See Time Averages. Ops. Res. 30:2, 223-231.

[26] Y.-B. Lin, “Performance Modeling for Mobile Telephone Networks”, IEEE Network Mag., vol.11, pp.63-68, Nov./Dec. 1997.

[27] 3GPP TS 23.002, “Network Architecture”, version 3.6, Release 1999, September 2002.

[28] 3GPP TS 25.038, “High Speed Downlink Packet Access (HSDPA); Overall description; Stage 2”, version 5.7, Release 5, December 2004.

[29] 3GPP TR 25.848, “Physical layer aspects of UTRA High Speed Downlink Packet Access”, version 4.0 , Release 5 , 2001.

[30] S . Borst, “User-level Performance of Channel-aware Scheduling Algorithms in Wireless Data Networks,“ *Proc. of the IEEE IFOCOM,* vol. 1, March 2003, pp.321-331.

[31] M. Kazmi and N. Wiberg, “Scheduling Algorithms for HS- DSCH in a WCDMA Mixed Traffic Scenario”, *Proc. of the IEEE PIMRC,* Beijing, China, September 2003, pp. 1485-1489.

[32] A. Farrokh, F. Blomer, V. Krishnamurthy, “A Comparison of Opportunistic Scheduling Algorithms for Streaming Media in High-Speed Downlink Packet Access (HSDPA)”, in *Proc. of MIPS*, November 16-19, 2004, Grenoble, France.

[33] Kelly F. “Charging and Rate Control for Elastic Traffic”. *Europeans Transactions on Telecommunications*, Volume 8, 1997. pp. 33-37. http://www.statslab.cam.ac.uk/~frank/elastic.html.

[34] Ameigeiras P. “Packet Scheduling and QoS in HSDPA. Ph.D. Thesis Dissertation”. October 2003. Center for Person Kommunikation. Aalborg University.

[35] Young I.S and Dan K.S, “Analytical Comparison of Three Packet Scheduling Schemes under a Per-User Minimum Throughput Assurance Requirement in HSDPA”. 2005 *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1708-1712.

[36] 3GPP TS 25.331,“Radio Resource Control (RRC) protocol specification”, version 3.21, Release 1999.

[37] 3GPP TS 25.413,“UTRAN Iu interface RANAP signaling”, version 3.14, Release 1999.

[38] 3GPP TS 25.433, “UTRAN Iub interface Node-B Application Part (NBAP) signaling”, version 5.16, Release 5, September 2006.

[39] R. Ramjee, R. Nagarajan, and D. Towsley, “On optimal call admission control in cellular networks,” in *Proc. IEEE Infocom*, San Francisco, CA, Mar. 1996, pp. 43-50.

[40] F. A. Cruz-P´erez, D. Lara-Rodr´ıguez, and M. Lara, “Fractional channel reservation in mobile communication systems,” *IEE Elect. Lett.*., vol. 35, no. 23, pp. 2000-2002, Nov. 1999.

[41] Y. Fang and Y. Zhang, “Call admission control schemes and performance analysis in wireless mobile networks,” *IEEE Trans. Veh. Technol*., vol. 51, pp. 371-382, Mar. 2002.

[42] J. L. V´azquez-´Avila, F. A. Cruz-P´erez,, and L. Ortigoza-Guerrero, “Performance Analysis of Fractional Guard Channel Policies in Mobile Cellular Networks”, *IEEE Trans. Wireless Comm.*, vol. 5, no. 2, pp. 301-305, Feb. 2006.

[43] Y.-B. Lin, “Performance Modeling for Mobile Telephone Networks”, *IEEE Network Mag.*, vol.11, pp.63-68, Nov./Dec. 1997.

Curriculum Vitae

Chung-Yung Chia born in Taipei, Taiwan, R.O.C., in 1963. He received the B.S. and M.S. degrees in Computer Engineering from National Chiao Tong University in 1987 and 1989, respectively. He is currently a PhD candidate in Computer Science at National Chiao Tong University , Hsinchu, R.O.C. Since 1989, he has been working in Telecommunication Laboratories, Chunghwa Telcom Co., Ltd, where he is currently a researcher and a project manager. His research interests include design and analysis of wireless communications network, development of telecommunication scheduling and monitoring systems, and performance modeling.

Publication List

- **Journal Publications**

1. Chung-Yung Chia*, Ming-Feng Chang. “Channel Allocation for Priority Packets in the GPRS Network.” *IEEE Communication Letters*, vol. 10, no. 8, pp. 602-604, Aug. 2006.

2. Chung-Yung Chia*, Ming-Feng Chang., “Uplink Connection Scheduling for Flat-Rate Data Services in the UMTS Network.” *IEEE Transaction on Vehicular Technology*, vol. 58, no. 5, pp. 2354-2365, Jun. 2009.

- **Conference Papers**

1. Chung-Yung Chia* , Jung-Tai Lee and Ming-Feng Chang , " Fast Handoff for Voice over WLANs," *National Computer Symposium* (NCS'07) , 2007.

- **Revision**

1. Meng-Fang Chang, Fang-Sun Lu and Chung-Yung Chia* , " A Callback Mechanism for Private Telecommunications Network," submitted to *Computer Communications*, 2008.

- **In Preparation**

1. Chung-Yung Chia* and Meng-Fang Chang , "Flate-Rate Packet Scheduling for the WCDMA Systems with HSDPA," prepared to submit to *IEEE Wireless Communication*.