Performance Analysis and Call Control Procedures in High-Speed Multimedia Personal Wireless Communications

SAM (Shaokai) YU

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

19 December 1999

Centre for Telecommunications
Information Networking
(CTIN)
Department of Electrical and Electronic Engineering
Faculty of Engineering

The University of Adelaide
South Australia
Australia

Abstract

Traffic integration and **global reach** are two major issues for emerging wireless multimedia communications. The fundamental criteria for quality of service (QoS) are to guarantee the requirements for real time traffic and provide best effort for non-real time traffic. However, the resources in wireless networks are limited and they must be shared according to the principle of dynamic and efficient allocation. The main objective of this dissertation is to develop efficient analytical methodologies for multimedia traffic integration and improved design tools to enhance data performance.

To evaluate integrated traffic performance, both spatial traffic variability and handoff must be taken into account. Matrix-analytic methods (MAMs) are developed as the main theoretical tool due to their efficiency and accuracy. To provide robust performance, the important requirements are usually considered to be: low call blocking and reduced handoff failure rate, maximum channel utilisation and minimum interference, optimised throughput and minimal delay. In order to maximise system capacity for multi-services and hierarchical cell structure (HCS) services, several new call admission control (CAC) policies are proposed to tune system performance and balance these parameter trade-offs.

For the provision of variable bit rate services, high-speed data (HSD) transmission is the key to the delivery of video or Internet multimedia applications in the near future. A dynamic call control scheme is proposed to improve call performance and a multipriority scheme is exploited to ease packet congestion in heavy traffic loads. Meanwhile, simulation results are shown to be consistent with analytical solutions. To further the understanding of blocking-delay performance in reservation protocols, different allocation strategies are compared and a non-boundary scheme is found to exhibit the best performance. Subsequently, a design benchmark for packet delay is numerically derived.

For design improvement, an optimal threshold design method is developed to satisfy the required loss-blocking probability in the interference-limited systems. A design guideline for guaranteeing loss-delay limit is then derived to smooth interference fluctuation. Finally, for HCS systems, the burstiness of overflow traffic is investigated by three different moment matching techniques. The final analysis favours using the higher moment method. Moreover, a disposition policy is employed to improve call dropping probability and the network is subsequently optimised by taking overflow traffic into account.

The main conclusions are: firstly, we observe that MAMs are robust enough to analyse the performance of HSD integrated services; secondly, the optimal resource allocation in integrated services largely depends on the traffic characteristics and input traffic statistics. Furthermore, terminal distribution, data message length and cell structure have a vital impact on integrated traffic performance. In particular, the effect of terminal mobility on call performance under high traffic loads is highly significant and therefore important to examine; finally, the CAC policies proposed in this dissertation are not only able to enhance traffic performance but also maintain communication quality at an acceptable level. In particular, the multiple priority scheme is a promising technique for the control of packet congestion in a high-speed wireless multimedia environment.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the university library, being available for loan and photocopying.

SIGNED:_____ DATE: 19 Dec 1999

Acknowledgments

I would like to take this unique opportunity to thank all the staff at the Centre for Telecommunications Information Networking (CTIN) of the University of Adelaide, Australia, especially the founding father and director of this centre Prof. Reg Coutts, who has provided tremendous support to me through the course of my study from day one.

Next, thanks must go to Dr. Theodore V. Buot, now in Nokia Research Centre, Helsinki, and Dr. Jianmin Li, now in Telstra Research Laboratory, Melbourne, who guided me and also engaged in countless discussions about the problems which occupied me most in this dissertation. During the first year, I received assistance from Dr. Ken Sarkies, while the lectures of Associate Prof. Bill Henderson, Dr. Peter Taylor, and Dr. Nigel Bean from Teletraffic Research Centre (TRC) in the University of Adelaide gave me invaluable insight and knowledge of stochastic modelling. Meanwhile, I would like to thank the Association of International Education, Japan (AIEJ) for the award of a scholarship, which allowed me to stay in the Takahashi's laboratory of the University of Electro-Communications, Tokyo, Japan during 1998.

Subsequently, my gratitude must be expressed to Mr. Peter Crook and Ms. Collette Snowden, for their time in the careful editing of this dissertation. In particular, I must appreciate Prof. Lang White, Dr. Theodore V. Buot, Dr. Linda Davis and Dr. Christine O'Keefe for meticulous reviewing and providing some advice to improve this thesis significantly. Meanwhile, the assistance of Dr. Derek Rogers, Dr. Sergey Nesterov, John Leske, Dr. Matthew Sorell, Phong Nguyen, and others in CTIN has also been valuable.

At last, I would like to thank my parents and my brother who have given me enormous support in my home country, and in my early, often extremely difficult years in Australia. They have truly inspired and encouraged me through those arduous and aimless days.

List of Publications

Journal Publication:

Yu Shaokai and T. Buot, "Data traffic distributed control scheme for wideband and narrowband integrated services in PWC," Institute of Electronics, Information and Communication Engineers (IEICE) Transaction on Communications, Japan, vol. E82-B, no. 6, pp. 834-840, June 1999.

Conferences Publications:

Yu Shaokai, "Data traffic control scheme for wideband and narrowband integrated services in PCS," Proceedings of IEEE Globecom'98, Vol. 3, pp. 1437-1442, Sydney, November 8-12, 1998.

Yu Shaokai, "Analysis of wideband and narrowband integrated services in cellular systems," Proceedings of the 3rd workshop on Personal Wireless Communications (PWC), pp. 183-190, Tokyo, April 8-9, 1998.

Yu Shaokai and Ted Buot, "Analysis of integrated services in GPRS cellular systems," Proceedings of the Third Asia-Pacific Conference on Communications (APCC), pp. 474-478, Sydney, December 7-10, 1997.

Ted Buot and Yu Shaokai, "Video transmission over reservation-TDMA protocols," Proceedings of the Third Asia-Pacific Conference on Communications (APCC), Sydney, pp. 1453-1456, December 7-10, 1997.

Yu Shaokai, "Channel allocations in NC-PRMA," The third Melbourne-Adelaide Teletraffic workshop, December 17-18, 1998.

Yu Shaokai and Ted Buot, "Teletraffic issues in overlay mobile networks with multimedia traffic," The first Melbourne-Adelaide Teletraffic workshop, December 7-8, 1996.

List of Abbreviations

ACTS	Advanced Communications Technologies and Services
AIUR	Air Interface User Rate
AR	AutoRegressive process
ARIB	Association of Radio Industries and Business in Japan
ARQ	Automatic Retransmission Request
B (W)-ATM	Broadband (Wireless) Asynchronous Transfer Mode
BA	Basic Algorithm
BAR	Basic Algorithm with Reassignment
BCO	Borrowing with Channel Ordering
BDCL	Borrowing with Directional Channel Locking
BER	Bit Error Rate
BFA	Borrow First Available
B-ISDN	Broadband-Integrated Services Digital Network
BS	Base Station
CABR	Channel Assignment with Borrowing and Reassignment
CAC	Call Admission Control
CBR	Constant Bit Rate
CCH(s)	Control Channel(s)
CDF	Cumulative Distribution Function
CDPD	Cellular Digital Packet Data
CHT	Channel Holding Time
CoD	Capacity on Demand
CODIT	Code Division Testbed
COST	Cooperation in the field of Scientific and Technical Research
CPDC	Common Packet Data Channel
CPSM	Contention-type Packet Switching Method
CRC	Cyclic Redundancy Check
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
CST	Cell Sojourn Time
CTD	Control Time Difference
C-TDMA	Consistent Time Division Multiple Access
CUT	Call Unencumbered Time
DCA	Dynamic Channel Allocation
DCS	Dynamic Channel Selection
DECT	Digital European Cordless Telephone
DS-CDMA	Direct Sequence Code Division Multiple Access
D-TDMA	Dynamic Time Division Multiple Access
EDGE	Enhanced Data Rates for GSM Evolution
EPA	Equilibrium Point Analysis
ERM	Equivalent Random Method
E-TDMA	Enhanced-Time Division Multiple Access
ETSI	Europe Telecommunications Standards Institute
FA	First Available
FBS	Fixed Boundary Scheme

FCA	Fixed Channel Allocation
FDMA	Frequency Division Multiple Access
FIFO	First-in First-out
FRAMES	Future Radio Wideband Multiple Access System
GFM(s)	Generation Function Method(s)
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Services
GSM	Global Systems for Mobile
HA	Hayward Approximation
HCA	Hybrid Channel Allocation
HCS(s)	Hierarchical Cell Structure(s)
HRS	Hybrid Reservation Scheme
HSCSD	High Speed Circuit Switched Data
HSD	High-Speed Data
I.I.D	Independent, Identically Distributed
IDC	Index of Dispersion for Counts
IMT-2000	International Mobile Telecommunications in the year 2000
IPP	Interrupted Poisson Process
IPRMA	Integrated Packet Reservation Multiple Access
ISMA	Idle Signal Multiple Access
IWSs	Integrated Wireless Systems
LANs	Local Area Networks
LEO	Low Earth Orbit
LIFO	Last-in First-out
LODA	Locally Optimized Dynamic Assignment
LOS	Line-of-sight
LRA	Logarithmic Reduction Algorithm
LRD	Low Rate Data
LST	Laplace-Stieljets Transform
MAC	Medium Access Control
MAHO	Mobile-Assisted Handoff
MAI	Multiple Access Interference
MAM(s)	Matrix-analytic Method(s)
MAU	Maximum Allowable Users
MBPS	Measurement-Based Prioritization Scheme
MBS	Movable Boundary Scheme
MCHO	Mobile-controlled Handoff
MD	Moving Direction
MMPP	Markov-Modulated Poisson Process
MMT(s)	Moment Matching Technique(s)
MPBS	Multiple Priority-Based Scheme
MS	Mobile Station
MSC	Mobile Switching Centre
MSQ	Mean Square
NBS	Non-Boundary Scheme
NC-CDMA	Non-Collision Code Division Multiple Access
NCHO	Network-controlled Handoff
NC-PRMA	Non-Collision Packet Reservation Multiple Access
ODCA	Ordered Channel Assignment Scheme with Rearrangement

PASTA	Poisson Arrivals See Time Averages
PCS	Personal Communication Systems
PDF	Probability Density Function
PHS	Personal Handyphone System
PP	Preemptive Priority
PRACH	Packet Random Access Channel
PWC	Personal Wireless Communications
QBD	Quasi-Birth-Death
QoS	Quality of Service
R.V.	Random Variables
RACE	Research into Advanced Communications in Europe
R-ALOHA	Reservation-ALOHA
RF	Radio Frequency
RING	Selection with Maximum Usage on the Reuse Ring
RoD	Reservation on Demand
RPSM	Reservation-type Packet Switching Method
R-TDMA	Reservation-Time Division Multiple Access
SAMA	Spread ALOHA Multiple Access
SBR	Simple Borrow from the Richest
SCS	Sequential Channel Search
SGSN	Serving GPRS Support Node
SHB	Sharing with Bias
SHCB	Simple Hybrid Channel Borrowing Strategy
SIR	Signal-to-interference Ratio
SNR	Signal-to-noise Ratio
SOR	Successive-over-relaxation
SSMA	Spread Spectrum Multiple Access
S-UMTS	Satellite Universal Mobile Telecommunication System
TCH(s)	Traffic Channel(s)
TCP/IP	Transmission Control Protocol/Internet Protocol
TDMA	Time Division Multiple Access
TIA	Telecommunications Industry Association
TRXs	Transmitters/Receivers
T-UMTS	Terrestrial Universal Mobile Telecommunication System
UMTS	Universal Mobile Telecommunication System
USFs	Uplink State Flags
UTRA	UMTS terrestrial radio access
VAD	Voice Activity Detector
VAF	Voice Activity Factor
VBR	Variable Bit Rate
VRRA	Variable Rate Reservation Access
WACS	Wireless Access Communications Systems
W-CDMA	Wideband Code Division Multiple Access
WWW	World Wide Web

Important Symbols and Notations

d_i	base station to terminal distance
P_B	blocking probability
t_{CUT}	call unencumbered time
CIR	carrier-interference ratio
L	cell perimeter
R_c	cell radius
t_{CST}	cell sojourn time
t_{CHT}	channel holding time
N_0	channel noise
r_c	channel rate
S	channel throughput
G	channel traffic
Q	conservative rate matrix
C_{con}	conventional additive white Gaussian noise channel
$f_{cost}(C_h)$	cost function
λ_{hd}	handoff data call arrival rate
λ_d	data call arrival rate
P_{ermd}	data permission probability
T	frame duration in PRMA
$g_r(t)$	Gamma distribution
$f_m^*(s)$	Laplace-Stieljets Transform
P_l	The probability for idle users
P_{loss}	loss probability
C_{ma}	multiple access channel capacity
M_e	overflow traffic mean
Var	overflow traffic variance
P_{drop}	packet dropping probability
h	path loss exponent
$\eta_{0.01}$	PRMA efficiency
PG	processing gain
R	rate matrix
C_h	reserved channels
r_s	source rate
$\pi^{(\infty)}$	stationary distribution
$W^{(k)}$	the *kth* moment of long term waiting time
λ_1	aggregated arrival rate in macrocells
λ_0	aggregated arrival rate in microcells

$\Lambda_{(p)}$	arrival matrix
$E(N_d)$	the average number of data packets
T_d	the average time of data calls in a system
$T_{dQ\|x}$	average waiting time in buffers
Bi_d	binomial distribution for data calls
Bi_v	binomial distribution for voice calls
β_M	cell cross-over rate in macrocells
β_n	cell cross-over rate in microcells
μ_1	channel service rate in macrocells
μ_2	channel service rate in microcells
$E(N_{d\|x})$	the conditional average number of data packets
α_d	data activity factor
$f(w)$	the density function of virtual waiting time
P_F	forced termination probability
λ_{h1}	handoff arrival rate in macrocells
λ_{h0}	handoff arrival rate in microcells
P_{hf}	handoff failure probability
Q_Q	infinitesimal generator for waiting time distribution
$\Pr(W > D_{d\max})$	maximum allowable queuing time for data packets
$D_{\max}$	maximum delay for speech packet
$D_{d\max}$	maximum allowable queuing time for data packets
$M_{0.01}$	the maximum number of conversations with 1% packet dropping probability
t_2	the mean duration of a silent gap
t_1	the mean duration of a talkspurt
λ_{v1}	new voice arrival rate in macrocells
λ_{v0}	new voice arrival rate in microcells
C_2	the number of channels in macrocells
C_1	the number of channels in microcells
m_0	the number of macrocells
n_0	the number of microcells
N_x	the number of slots occupied by x voice users
N	the number of slots per frame in PRMA
N_u	the number of users in NCAC
M_u	the number of voice users in CDMA
v	the order of a Gamma distribution
H	packet header in PRMA
α_g	the parameter of a Gamma distribution

$\Pr(K)$	the probability of a K-handoff call
W_{Delay}	the queuing delay time for data packets
γ	the rate of data call cell sojourn time
μ_{dCH}	the rate of data call channel holding time
μ_d	the rate of data call unencumbered time
η	the rate of voice call cell sojourn time
μ_{vCH}	the rate of voice call channel holding time
μ_v	the rate of voice call unencumbered time
$\eta_{\inf}$	The ratio of the required interference to noise power density
α_w	the ratio of voice duration to data duration
M	service matrix
Ω	the state space of integrated services
$\xi_j(t)$	virtual waiting time at *jth* server
α_v	voice activity factor
$W(t)$	waiting time distribution
C	total available channels
λ_2	total data arrival rate
λ_1	total voice arrival rate
P_{owi}	transmitting power
I	unit matrix
V	vehicle speed
λ_v	new voice call arrival rate
λ_{hv}	voice call handoff rate
P_{ermv}	voice permission probability

List of Figures and Tables

To:

my parents, Yu Gongming, Lin Sufen,
brother Yu Shaoyan for their support,
and lovely Zheng Ni for her patience.

Chapter 1

Introduction

1.1 Background, Identified Problems and Challenges

Future wireless networks are expected to provide personal communications with anywhere, anytime, and anyone connectivity. Correspondingly, future terminals are required to carry a mix of voice, data and video source traffic as well as provide personalised services. The overwhelming demand for more capacity and new services is propelling the speeding development of wireless technology during this decade. However, existing networks find it difficult to meet such requirements and can only provide users with limited accessible services of *some*-where, *some*-time, and *some*-persons. Therefore, there are many research activities being carried out in the wireless world at present.

The aim of ubiquitous personal wireless communications (PWC) is to provide person-to-person wireless communications without tethered constraint [*Cox95*]. By and large, the main problem in future wireless networks can be simply expressed by the contradiction between the bandwidth-required multimedia services and the bandwidth-limited or interference-limited wireless systems. It is known that high quality delivery of multimedia applications requires high network bandwidth. However, in a unified transport network, the resources in wireless networks are physically limited. The scramble for limited resources ends with conflict and leads to denial of servicing. As a consequence, network QoS can be severely degraded.

The response to this problem is either to increase channel capacity for accommodating more users or to improve design methods for optimising network resources. While many investigations concentrate on the former issue [*Viterbi93*], [*Lavery93*], [*Grieco94*], [*Fuka96*], and [*Evans99b*], etc, which still remains open to some extent, the latter issue is more crucial but has lacked particular attention and systematic investigation in the area of integrated services [*Calleg95*] and [*Jabba96*]. As a result, this becomes a focus of this dissertation. The objectives of this dissertation are mainly concentrated on the development of the generic analytical methodologies and the improvement of call control strategies to enhance data performance. Specifically, it is intended to improve QoS, e.g., reducing blocking or loss probability and also eliminating undue delay while maintaining

required throughput. As a result, we do not discuss the techniques of capacity enhancement in detail.

In contrast to wireline networks, which aim to carry a maximum amount of calls on minimum links, the ultimate objective of wireless systems is to maximise the number of subscribers under a given limited resource while maintaining acceptable QoS [*Everitt94*]. In order to provide high performance networking and meet increasing demands, it is essential to develop new theoretical methodologies and design tools for heterogenous environments. Because the design of wireless systems needs to take spatial traffic variability, different source traffic patterns, interference, multiple access techniques, and multi-layer structures into account [*Jabba96*], the evolving and next-generation wireless networks pose significant challenges in design advancement and multimedia traffic modelling. The problem of traffic integration lies in how to find efficient and accurate theoretical modelling methods for traffic integration. On the other hand, although the source generating traffic is beyond control, call control policies can be enforced on the network side to enhance overall system performance. The problem of design improvement is concerned with the issue of the required QoS through the development of suitable and flexible call admission control (CAC) policies, which pertain to the wireless multimedia environment. Generally, the study of multimedia traffic needs to consider four major issues: mixed traffic performance, switching architecture, radio frequency (RF) modulation and encoding techniques. From the viewpoint of traffic performance analysis, our study mainly focuses on the QoS aspects rather than radio transmission quality.

Following the introduction of nomadic mobile computing concepts in wireless systems, the importance of non-voice services becomes more significant in traffic performance studies. Apart from the consideration of both movable terminals and source traffic, the difficulty in performance evaluation lies in voice, data and video services requiring different service mechanisms and QoS. While voice and video signals must be transported in a relatively low error channel, data users need to be transmitted in a variable bandwidth. If voice activity is considered, time slots are allocated to different traffic sources to gain the benefit of statistical multiplexing. On the other hand, delay and bit error rates must be sustained according to the users' needs at data application levels. In addition, the delivery of high rate data services requires more channels than single voice service. With regard to video traffic, there are two kinds of video services. One is called retrieval-type service and the other conversational-type service. In particular, the different types of services have

different transmission requirements. In comparison with the retrieval-type services, the conversation-type services require lower delays and thus need more capacity. Simply speaking, the key to QoS in multimedia traffic in wireless networks is to guarantee the required QoS for real-time traffic and also provide best effort QoS for non real-time traffic. Here best effort means the provision of minimum guarantees for performance while an unspecified variance is allowed. Specifically, maximal throughput is usually required, while excessive delay is seen as undesirable from a user point of view. As a consequence, channel allocation in network planning for bursty packet data is of significance and also is quite different from that for conventional real time traffic in integrated services.

It is known that data applications in Time Division Multiple Access (TDMA) systems can operate in two modes, i.e., circuit switching and packet switching. In order to deliver high data rate services in TDMA systems, more than one slot can be dedicated to the same terminal, which is usually termed multislot service. As a result, the overall data rate becomes a multiple of the basic rate. In this study, once the number of multiple slots becomes high enough, this is regarded as high-speed data (HSD) services. High-rate data services will be very important in the future. For instance, the high-speed wireless data communications are to deliver video services in a circuit access mode, whereas HSD packets are essential for wireless Transmission Control Protocol/Internet Protocol (TCP/IP) based data applications.

For circuit mode services, High Speed Circuit Switched Data (HSCSD) is an enhancement service that can provide data rates higher than the original 9.6 kb/s in GSM systems. The concept of HSCSD is to allow parallelled multiple full rate traffic channels (TCH/F) for the use of a single connection. As a result, the capacity of HSCSD becomes a multiple of a single TCH/F capacity and hence leads to a significant increase in data rate. During implementation, this can be realised by combining several physical channels. At the beginning of call set-up, the number of TCH/F needs to be indicated by the air interface user rate (AIUR). In addition, because multiple traffic channels are used for the same HSCSD, handoff operation should be controlled as one radio link only [*ETSI334*]. Although HSCSD can provide high data rate transmission, the transport mechanism actually limits the efficiency of transmission. Therefore, the use of circuit switching is unsuitable for the efficient delivery of bursty data packets [*Har80*] and [*Ross82*].

On the other hand, packet switching is found to be more suitable for packet transmission because data packets are bursty in nature [*Kaya98*]. It can provide data packets with asynchronous, asymmetric and multicast transmissions. In particular, packet switching is able to support both constant bit-rate (CBR) and bursty variable bit-rate (VBR) services. Moreover, because packet switching exploits idle channels, it has efficient, flexible and economic advantages.

It is evident that, in the near future, these two different modes of traffic handling mechanisms will eventually merge and provide efficient services for wireless users. For instance, motivated by overlaying the existing infrastructure, General Packet Radio Service (GPRS) is proposed for wireless data access in the same band as the present GSM system [*ETSI364*]. However, based on the author's search of current literature, the traffic performance analysis of data traffic in such integrated systems is not widely documented, especially for the study of high data rates. The problem in HSD services is how to efficiently reuse the scarce radio spectrum resource to provide best effort QoS. Therefore, the study of HSD services forms the core of this investigation. In addition, because the most important aspect of the proposals for the next generation wireless system is the backward compatibility with the present systems, the investigation of integrated services in the evolution of GSM technology and emerging Code Division Multiple Access (CDMA) technology becomes of interest in this dissertation. Focusing on the performance of HSD services, this dissertation also concentrates on call control strategies and draws on analysis of a wide range of integrated wireless systems (IWSs), which are defined as including both the integrated terrestrial systems and the integrated satellite systems in this study.

Since 1917, when the Danish mathematician A. K. Erlang published his well-known loss formula for the telephone system, traffic theory has experienced significant development and has expanded to other areas such as ecological modelling [*Kelly91*]. For over a decade, the traffic theory in telecommunications has also experienced extraordinarily rapid development, especially in the wireless area.

Generally speaking, there are three main methods that can be used to analyse the traffic performance: traffic measurement, simulation and analytical methods. Traffic measurement collects statistics from field data, while analytical techniques are efficient in computation but have to be based upon some strict assumptions. On the other hand, simulation can be used to evaluate the network behaviour without such constraints. In this

dissertation, in order to analyse the mixed traffic performance, both the analytical method and the simulation method are adopted.

It is well known that simulation models and approximate analytical models have been successfully used in the development of first generation and second generation systems. However, more precise analytical methods and more effective call control policies are needed to guide the design of future networks. Design and management decision-making requires us to accurately forecast the performance of networks. Only if we can provide more accurate traffic performance analysis and effective control methods, can we better predict traffic behaviour and efficiently allocate network resources. Otherwise, once a 'hot spot' cell occurs, which means that the offered traffic is greater than the design traffic, the required QoS can not be attained and the performance of networks will be significantly degraded.

Using traditional analytical methodology, it is difficult to derive statistical performance measures for bursty source traffic, especially for emerging multimedia services, because of large state space and complexity. Fortunately, the development of MAMs is becoming mature and robust [*Bright96*]. With the aid of such promising methods, mathematical tools are developed to solve specific wireless integration problems in this dissertation.

For two classes of traffic, a two-dimensional canonical setting can be used to represent the state space. Usually, the state space is described by two integer random variables j and i, where j has a finite range and i takes any non-negative value. Therefore, generation function methods (GFMs) can be used to solve the equations of state space [*Pavli94b*]. Specifically, the GFM is to list global balance equations according to the global balance rule, which states that all the rates flowing into a state must be equal to all the rates flowing out from that state. Subsequently, the use of transform techniques can convert a difference equation into an algebraic equation, which can be solved by algebraic techniques [*Schwartz87*]. The moment generating function is usually denoted as: $G(z) = \sum_{n=0}^{\infty} p_{\infty}^{(n)} z^n$, where $p_{\infty}^{(n)}$ is the discrete probability. By differentiating $G(z)$ with respect to z, we are able to find the first and second moments of the distribution. In the end, Cramer's rule and l'Hospital's rule can be used for analytical manipulation [*Serres88*]. As a matter of fact, the difficulty in the use of GFMs is in determining the root of the determinant equation. If the roots belong to the case of multiple roots or become close to

unity, this increases the difficulty in solving these equations [*Willi84*]. In addition, in some complex queuing systems, e.g., involving infinite queues and taking handoff calls into account, the balance equations are not readily solved by simple recursive techniques.

Alternatively, as another efficient and promising technique, **matrix-analytical methods (MAMs)** are proposed to solve such a problem. "*MAMs are defined as a tractable approach to algorithmic probability, which involve the modelling development with underlying matrix structure*" [*Bright96*]. As stated in [*Bright96*], although MAMs are at an advanced stage of their development, they still develop at a rapid pace and are found in many applications such as manufacturing processes, inventory systems, biological processes and telecommunication areas.

The key to using the MAMs is to determine the minimal positive solution of the rate matrix R from a non-linear matrix equation. Subsequently, the invariant vector is expressed in terms of the powers of the rate matrix R. Although conventional iterative methods are usually used to determine steady state probability distributions, convergence is very slow and hence inefficient for complex problems. This motivates us to find more efficient methods. In addition, although Williams [*Willi84*] proposes an iterative method for the rate matrix, complexity and stability have not been well addressed in that study. In particular, the previous algorithms are found to be unsatisfactory for some difficult problems. The use of logarithmic algorithms is able to circumvent this difficulty in numerical solutions [*Latou93*]. In order to solve the traffic integration problem in wireless multimedia services, we propose and develop MAMs as an important analytical tool for these novel applications in this dissertation.

By using such theoretical tools, the analysis of traffic integration in wireless multimedia systems becomes the focus. In general, three important aspects are seen as essential for designing a future wireless network, namely, speed, variety and reliability [*Bout97*]. Specifically, speed is analogous to transmission rate, e.g., high-speed services are required in future wireless networks. Variety is important for the provision of a wide range of data services to target different customer niches, while reliability is necessary to guarantee error recovery and also meet the required QoS.

Keeping such criteria in mind, this dissertation principally tackles three important aspects, i.e., multi-services, CAC strategies in interference-limited systems and multi-layer systems, which are expected to form the key areas for traffic studies of future wireless multimedia networks [*Jabba96*]. Examples include integrated cellular systems, integrated

cordless and cellular systems, integrated terrestrial and satellite systems. These systems are expected to coexist in future PWC systems and therefore warrant further study. Throughout this thesis, CAC can be regarded as a unifying thread, which is used to enhance integrated traffic performance.

Firstly, future wireless networks require us to integrate data services into existing voice services. Ultimately, video services will be integrated into voice and data services to create multi-service in an efficient manner. The convergence of the telephone and computer industries requires that the provision of wireless Internet multimedia services becomes a reality in the near future. The TCP/IP based data applications will play an important role in future wireless data networks. At the heart of multimedia services is the problem of optimal resource allocation. In particular, packet channel allocation and packet network planning are the most challenging problems for data packet over circuit access. This requires us to analyse traffic performance of data packet in integrated services. The key challenge is how to integrate data services, especially for HSD services, with voice services in an efficient way for a large number of users over distributed geographical areas.

While voice services are already well in place, variable bit transmission has now become the focus of attention. In particular, only a high-speed system can deliver a wide range of wireless data services. By using optimised modulation techniques, the Enhanced Data Rates for GSM Evolution (EDGE) technology is expected to provide data rates as high as 384 kb/s. To satisfy the different levels of QoS, call control policies can be adopted to assure a high level of network performance. For the QoS in traffic integration, minimum blocking and least delay are regarded as two criteria. For instance, call blocking is usually required at 2% and the allowable delay in a Packet Reservation Multiple Access (PRMA) system needs no more than 32 ms [*Goodm91*]. In addition, because handoff calls require a relatively lower failure rate than that of a new call, handoff controls should be handled differently from new call controls.

In order to provide for the strict requirements of QoS, CAC is seen as a key design improvement tool. Traffic integration will increase traffic intensity in existing systems. In highly congested wireless traffic environments, especially in dense urban areas, we are faced with the problem of efficiently managing traffic congestion and rationally allocating limited resources. In order to alleviate the burden of the network controllers, the use of distributed control techniques has been proposed for future packetized multiple access

[*Goodm89*]. By using packet transmission, the conventional dedicated channel is replaced by the concept of statistical multiplexing. However, because optimal resource allocation relies on the study of traffic integration, an efficient theoretical methodology needs to be developed for integrated services and multi-level traffic control schemes need to be addressed further. In addition, taking the importance of mobility into account, we concentrate on the effect of terminal mobility on multimedia traffic variability. We investigate how volatile voice traffic affects overall system performance, especially for HSD performance.

The second problem addressed in this dissertation concerns CAC schemes in interference-limited services. It is known that the studies of traffic performance rely largely on the specific network's architecture and different access control schemes. Meanwhile, although multiple access design only contributes to a small part of network design, multiple access mechanisms must be taken into account in the design of networks. The selection of radio interfaces not only has a significant effect on the spectral efficiency, which can be evaluated in terms of *Erlangs/MHz/km*^{2}, but also influences Integrated Services Digital Network (ISDN), Broad-ISDN or even Asynchronous Transfer Mode (ATM) back-bone fixed networks. Moreover, this will eventually determine the cost of resources and the complexity of equipment.

As a result of the high spectral efficiency of CDMA systems, there have been many research and development activities carried out in recent years. For instance, Code Division Testbed (CODIT) is part of Research into Advanced Communications in Europe (RACE) projects, originally intended for the interface of the Universal Mobile Telecommunication Systems (UMTS), which aims to provide for data bit rates up to 2 Mb/s [*Ander95*]. More specifically, CODIT is to provide multirate services, mixed cell structure, macrodiversity, simultaneous voice, data and even video transmission under the principle of efficient radio resource use.

As in the case of integrated services in TDMA, packet services are allowed to gain access but not to impair voice services. However, the circuit mode voice calls and packet mode data in CDMA can simultaneously gain access to the channel without constraint and therefore do not need a special protocol for slot assignment. To some extent, this can drive the system design in CDMA to become simpler than that in TDMA. The challenging problem in the interference-limited services is to design an appropriate access control mechanism so as to smooth interference fluctuation for multimedia services.

For multiuser communication in CDMA systems, the loss probability is defined as the probability of the total interference exceeding the pre-defined level [*Viterbi93*]. Accordingly, the Erlang capacity can be determined. The availability of channels in CDMA is defined in terms of a pool of available code channels. Compared with the conventional FDMA or TDMA systems, the arrivals in CDMA are not blocked in terms of the occupancy of the frequency slots or time slots. In other words, blocking would not occur if receiver capacity is not limited. Therefore, as long as the aggregate interference does not exceed the predefined threshold, the QoS in CDMA systems is regarded as being satisfactory [*Mermel93*].

Due to the interference averaging property, any reduction of interference results in the proportionate increase of system capacity. Therefore, there exists a tradeoff relationship between system capacity and communication QoS. Compared with the conventional TDMA system, the maximum capacity in a Direct Sequence-CDMA (DS-CDMA) system is more susceptible to interference. Poor QoS will occur once the total interference level exceeds a certain limit. Therefore, from a system capacity viewpoint, the key problem for DS-CDMA integrated services is how to smooth interference fluctuation so as to provide a maximum number of users for each class service with required QoS. As a result, a new design philosophy that focuses on the tradeoff relationship between system capacity and QoS must be further pursued in integrated services.

Since CAC can effectively control the maximum number of users, the study of CAC is essential in order to guarantee the required SIR level. The main focus for the design of CAC is to define the effective threshold of CAC while maintaining maximum Erlang capacity. Generally, the maximum number of users can be restricted so as to achieve a required level of interference. There are two different forms for admission control. One is based upon the number of users [*Lavery93*] and [*Liu94*]. The other is based on the interference level [*Viterbi93*]. For integrated services, because the tradeoff relationship between system capacity and communication quality still remains unclear [*Ishika97*] and [*Naga98*], the problem of how to choose the effective threshold and how to quantify the effect of CAC on the QoS needs to be further studied. This motivates us to develop a new design guideline for satisfying the loss-delay relationship for integrated traffic.

The last challenge is how to handle call control strategies in HCSs. Apart from improved coding schemes, optimised modulation techniques and using frequency hopping, microcells can be used as a means to increase system capacity. In order to exploit

spectrum utilisation efficiently, a layered structure can be adopted for access to large and small areas. In this case, not only can microcells utilise the same spectrum as macrocells, but upper cells can also be used as an overflow group to prevent the loss of lower layer traffic. However, increased handoff times among layers will lead to an increase in the use of system resources. As handoff crossing rates increase, the increased signalling traffic will impose an extra cost on switching centres of limited resources. Therefore, an improvement for handoff design is to reduce the number of handoff events by enhancing handoff algorithms so as to ease the effect of delay as well as to minimise the degradation of QoS. As a result, the optimal design of multi-tier networks is crucial in performance enhancement.

Since resource allocation among layers is dependent upon overflow traffic, the methodologies of overflow traffic modelling are important and are key to the provision of optimal resource allocation among layers. A more accurate moment matching technique (MMT) is needed to describe the behaviour of overflow traffic. The overflow traffic control policy and network optimisation are subsequently investigated in this dissertation. Furthermore, the effect of terminal mobility on the performance of overlaid systems is investigated.

In summary, taking wireless distinguishing characteristics into account, this dissertation focuses on how to allocate resources for multimedia services in an efficient and effective manner through the study of traffic integration and the facilitation of CAC strategies. The ultimate objective is to provide multi-services and multi-layer services for future personalised wireless multimedia networks with required QoS.

1.2 Dissertation Overview

Briefly, this dissertation concentrates on the development of new analytical methodologies and new CAC strategies for multimedia services in future wireless networks. TDMA, Reservation-TDMA (R-TDMA), DS-CDMA and HCS integrated systems are taken into consideration. In particular, optimal resource allocation, spatial traffic distribution and transmission rates are of particular interest. Generic theoretical methods are developed to analyse data performance, especially for HSD in traffic integration. Several new call control strategies are proposed to enhance system

performance. Additionally, global access systems are considered through the investigation of overflow traffic.

The background, problem definition and challenges of this study appear in the first chapter. Subsequently, the main contents of the dissertation are comprised of five components.

Chapter 2 reviews the existing studies of traffic performance for integrated services. Network control strategies, including radio resource control, new and handoff call controls, and mobility management are then discussed. Performance measurement methods, such as the conventional approximation methods and generation function methods are then discussed. Access control methods for integrated services are systematically classified. Subsequently, the existing analyses for integrated services in TDMA systems, R-TDMA systems, DS-CDMA systems and HCSs are presented.

In Chapter 3, generalised call sojourn time and call performance in multiple platforms with multi-mobility are investigated in Subsection §3.1.1. Handoff call control policies are analysed in Subsection §3.1.2. Then the performance of multiservice in loss systems is investigated. Furthermore, an approximation method for integrating low rate data packets into voice services is proposed in Section §3.2. Meanwhile, multimedia source models for voice, data and video traffic are described.

The studies of HSD performance are highlighted in Chapter 4. Using the developed robust MAM in Section §4.1, a hybrid reservation scheme (HRS) for HSCSD systems is proposed so as to improve voice call blocking probability. For HSD packets, a multiple priority-based scheme (MPBS) is used to ease packet congestion in Section §4.2. Subsequently, an application of reservation-TDMA, namely PRMA, is used to investigate channel allocation schemes. During the study, we elaborate on the blocking-delay relationship of voice and data packets in a single-carrier multiple-client environment. Three slot allocation strategies, i.e., fixed boundary scheme (FBS), movable boundary scheme (MBS) and non-boundary scheme (NBS) are then contrasted. Moreover, an efficient MAM is developed to evaluate queuing performance of data packets. In order to ease data congestion in the heavy load region, we subsequently develop a multipriority scheme for data packets.

In Section §5.1 of Chapter 5, quality-based CAC schemes in DS-CDMA systems are considered. An optimal threshold design method for the interference-limited voice system is then developed. Voice control and data control schemes are subsequently analysed. A

design guideline for integrated services is proposed. Subsequently, for HCS systems, the overflow traffic performance is analysed in Section §5.2. We contrast three moment matching techniques and then compare their results. Accordingly, we develop a disposition policy for overflow call handling. Additionally, network optimisation is then discussed. Finally, we follow up with conclusions and future work in the last chapter.

1.3 Original Contributions in this Dissertation

The author recognises the importance of both HSD performance and call control schemes in future multimedia services. Over the years, there have been some difficulties in developing efficient analytical methods to obtain the performance of HSD integrated services and employing effective call control procedures to the multimedia wireless services. This dissertation mainly tackles these two important issues. To my knowledge, there are no prior results that have been reported by others in these areas.

Some of the original contributions have been published in journals and conference proceedings. For this dissertation, the main contributions are in the area of analysis of systematic performance, especially for HSD performance in time channel systems, and improvement of call control strategies. We show in this dissertation that the new proposed call control schemes and design methods can enhance call performance under arbitrary traffic load conditions, especially under heavy traffic load conditions, which have a potential use for future personalised multimedia systems. Such an approach is in contrast to the capacity enhancement methods, which are widely adopted in most previous studies. The specific contributions are:

Firstly, the impact of traffic variability and priority handoff schemes on call performance in multiple platforms is investigated. As a result, the concept of integrated radio resource management has been advanced. Subsequently, diverse multimedia source traffic models are described. Furthermore, a marginal distribution method is proposed to analyse single slot integrated services with consideration of prioritised handoff traffic and terminal mobility. Although this technique is approximate, it is found to be able to estimate the performance of integrated services efficiently.

Secondly, to integrate HSD services into voice services, an analytical methodology called the matrix-analytic method (MAM) is proposed and developed to analyse the systematic performance of a novel prioritized wireless application by using a matrix-

geometric solution for a quasi-birth-and-death (QBD) process. For HSCSD, a new hybrid reservation scheme is proposed. It can be used to reduce new voice call blocking probability with the increased traffic loads. For integrating HSD packets into voice services, a new multipriority scheme and a source coding scheme are proposed. It shows that the congestion of data packets can be effectively eased by using optimal coefficients. Meanwhile, variable voice coding schemes can be used to enhance data performance without degrading voice services significantly. In addition, the impact of spatial traffic distribution due to terminal mobility on integrated traffic performance is also investigated. It shows that data packets can gain benefit from high variability of voice traffic. For the R-TDMA systems, different kinds of channel access strategies are contrasted and a non-boundary scheme is suggested for the integrated services with reservation protocols. Moreover, a new benchmark parameter for data packets is numerically analysed. In the meantime, data packet performance is enhanced by using the new proposed multipriority scheme.

Next, for CAC in interference-limited systems, an optimal threshold design method is proposed to guarantee rigid voice call QoS. A new design guideline for integrated voice and data services is proposed to justify the loss-delay performance. For HCSs, the Hayward approximation method is proposed to analyse the burstiness of overflow traffic. Subsequently, different kinds of MMTs for non-random traffic are compared and it concludes that only the high MMT can be used to evaluate the accurate performance of the overlaid systems. Meanwhile, a new disposition policy is proposed to improve overflow call dropping probability and a new optimisation algorithm is developed while taking the overflow traffic into account.

Finally, other minor contributions are embedded in the studies of this dissertation, which include the effect of CST on call performance, the impact of terminal mobility on voice waiting time distribution, the time-out model for handoff traffic, and the performance of multiservice in loss wireless systems.

Chapter 2

An Overview of Recent Research

Resource allocation and **traffic integration** are two major issues in future integrated networks. The aim of a traffic integration study is to achieve optimal resource allocation, while limited resources have to be assigned to different traffic classes. Resource allocation must be handled by using network control strategies in an efficient manner. On the other hand, integrated traffic performance can be enhanced by optimally enforcing call control strategies. Our concern in this study is to analyse integrated traffic performance and to achieve optimal resource allocation through the use of appropriate call control strategies.

The contents of this chapter are organised as follows. In Section §2.1, network control strategies are discussed. Firstly, radio resource management is introduced, which includes bandwidth assignment schemes, resource sharing policies and multi-service resource allocation divisions. Secondly, new call and handoff call controls are discussed. In Section §2.2, different performance measurement methods are examined. Subsequently, the previous studies for integrated services in TDMA systems are investigated in Section §2.3. Next, traffic control procedures in both single layer and multi-layer systems are presented. Finally, a summary follows.

2.1 Network Control Strategies

It is known that network resources must be rationally allocated by using optimal control strategies. In this study, network control strategies are classified into three categories: resource management, call admission control (CAC) and handoff control. Specifically, CAC determines whether a system accepts, delays or even rejects a request for a new communication set-up. After a call request is accepted, the system has to allocate resources for the continuing communication platform. Resource management includes bandwidth assignment, channel allocation and power control. In order to maintain a high transmission quality for a progressing call on the move, handoff control can be employed to reduce handoff call dropping probability. These three aspects of network control strategies are described as follows.

2.1.1 Dynamic Radio Resource Management

Strictly speaking, personal wireless communications, either using TDMA or CDMA, is a radio resource-limited system. The precious network resource must be rationally allocated between different users and also be well managed. Optimal and dynamic radio resource management is required whilst maintaining the required QoS, especially for packet mode channel allocation. The study of resource management is of paramount importance because it can determine spectrum efficiency, transmission quality, power consumption, and even network infrastructure cost.

From the viewpoint of traffic analysis, the performance of circuit switching in wireless networks is a variant of wireline networks. For example, Tekinay [*Tekin93*] has studied heterogenous services in a circuit mode with the consideration of channel assignment schemes. Generally, circuit mode is suitable for the transmission of real time services, e.g., voice and video, which require guaranteed QoS. In contrast, non-real time services like bursty data packets, which require best effort QoS, tend to be more suitable for the transmission of packet mode [*Kaya98*]. Some early versions of packet radio systems like ARDIS and Mobitex have been introduced in several countries [*Pahla94*]. Furthermore, a unified transport network, which can exploit these two different transfer modes and share the scarce network resources in a most efficient and prudent manner, has been in operation in some countries. For example, Cellular Digital Packet Data (CDPD) is designed to overlay the existing analog Advanced Mobile Phone System (AMPS) system in North America [*Pahla94*].

To manage limited resources, channel access techniques use certain default formats, commonly known as protocols. As we are going to analyse traffic performance under a specific configuration, some basic resource assignment methods, according to classification of protocols, need to be clarified.

2.1.1.1) Bandwidth Assignment Schemes

It is known that robustness, efficiency and security are seen as three main objectives of protocol design. The criteria for the choice of multiple access control protocols for data networks are dependent upon traffic statistic characteristics and the framework of the technology development at that time [*Abram94*]. Among efficient resource allocation methods, the most basic is bandwidth assignment, which is realised by medium access

control (MAC) protocols. Multiaccess protocols can be classified according to whether they are *static* or *dynamic*. There are three most important categories: fixed assignment, random access and demand assignment [*Prasad96*]. Specifically,

- **Fixed Assignment.** This scheme assigns a certain bandwidth to specific users. Namely, the total channels are divided into orthogonal channels for different users whether signals are transmitted or not. Obviously, the use of fixed assignment in FDMA or TDMA systems may become inappropriate for bursty traffic in integrated services because of inefficient radio spectrum utilisation.

- **Random Assignment.** If the number of potential users is much larger than the available channels, the fixed assignment scheme is found unsuitable. Instead, another contention scheme called random access scheme, in which different users compete for the same channel, can be adopted. Although random access is suitable for bursty data traffic, it is not desirable for delay-sensitive traffic and has the apparent problem of hidden terminals. For example, although Carrier Sense Multiple Access with Collision Detection (CSMA/CD) can provide the wireline networks with very high throughput, this technique is inhibited by the difficulties in sensing the remote carriers in the presence of local transmission in a radio environment. The signal from a local transmitter can overload the receiver and then disable any attempt to sense the remote transmissions. To conquer this problem, the use of a contention-based protocol called Idle Signal Multiple Access (ISMA) has been proposed by Mukumoto [*Muku81*]. Another important protocol called Packet Reservation Multiple Access (PRMA) is later proposed by Goodman [*Goodm89*]. Moreover, it is notable that the use of CDMA is a combination of fixed and random assignments.

- **Demand Assignment**. Bandwidth can be assigned according to the requirements of users. Under such an arrangement, bandwidth is allocated only when users have messages to transmit. Users transmit a request access packet to a base station (BS) and then the BS assigns channels for users once available. While the VBR users are in idle mode, the bandwidth is allocated to the other users. Therefore, demand

assignment is suitable for VBR traffic and multimedia traffic. Examples of this assignment include centralised control roll-call polling and distributed control token-bus. Most of these applications are used in high-speed Local Area Networks (LANs). Unlike random assignment, demand assignment can avoid the waste of bandwidth due to collision by providing the connections with contention-free bandwidth during active periods. However, it does introduce additional overheads during the request and assignment process.

In summary, according to the characteristics of multimedia traffic, different bandwidth assignment schemes or their hybrids can be used flexibly. In particular, the problem of resource allocation in multimedia must incorporate a consideration of both physical layer properties and source traffic characteristics.

2.1.1.2) Resource Sharing Policies

Apart from bandwidth assignment as mentioned early, another apparent resource allocation problem related to traffic studies is channel allocation, which needs to make a decision for which sets of channels can be reused and which calls they might be assigned to in each cell. The main idea behind channel allocation is to minimise the signal-to-interference ratio (SIR) by making use of radio propagation path loss and thus increase spectrum reuse efficiency.

It is well known that spectrum reuse can cause interference from a reuse distance. For TDMA and PRMA systems, the interference signals mainly come from the co-channel cells because active users are not allowed to share the same time slot [*Frull94*]. The effect of the co-channel interference in TDMA or PRMA systems is shown as in [Figure 2.1]. Here we use d_i to represent the distance between a transmitting BS and a receiving mobile station (MS), in which the transmitted power is P_{owi}. Obviously, the reuse distance is a trade-off between the co-channel interference and system capacity.

If the channel noise signal is represented by N_0, the average SIR at the reference station R (with distance d_R) can be given by [*Katze96*]:

$$SIR = \frac{P_{owR} d_R^{-h}}{\sum_{i=1}^{n} P_{owi} d_i^{-h} + N_0} \tag{2.1}$$

where h denotes the path loss exponent for narrowband propagation.

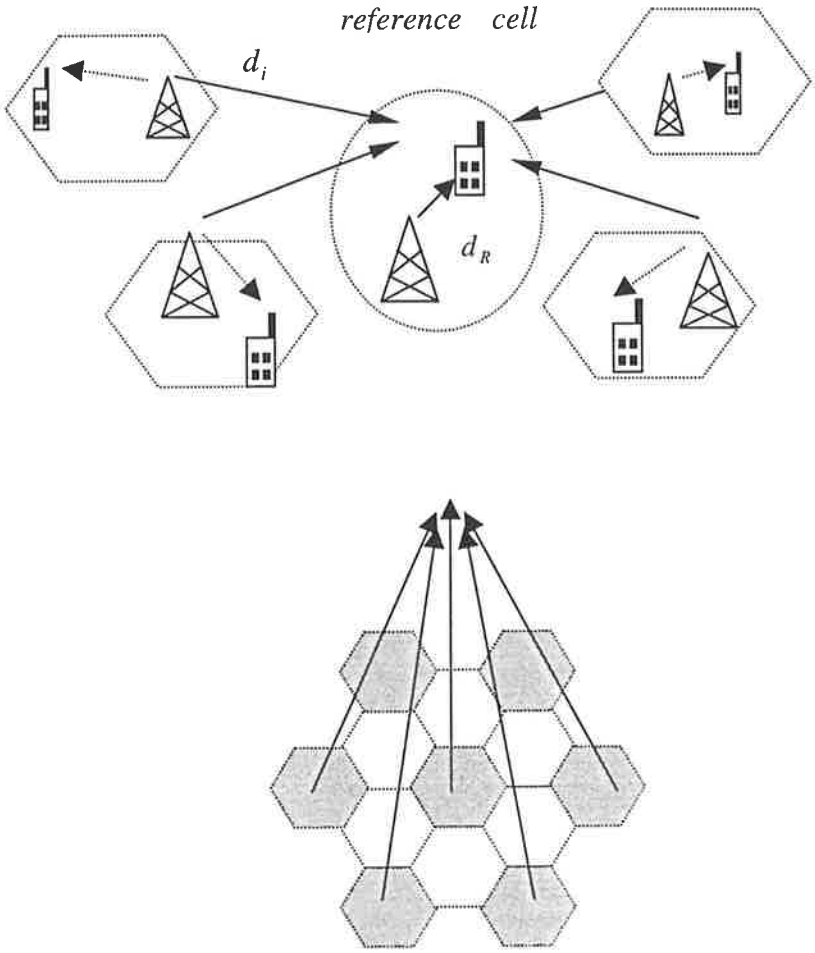


Figure 2. 1: The impact of co-channel on the reference cell

In TDMA systems, besides the co-channel interference, adjacent channel interference must be taken into account during network planning as well. In contrast, for CDMA systems, the concept of interference mechanisms is totally different, because all users share the same bandwidth and performance can be improved by using interference cancellation techniques [*Frull94*].

The ultimate objective of the use of channel allocation strategies is to achieve a required QoS for efficient spectrum utilisation. In general, channel allocation is classified as fixed channel allocation (FCA), dynamic channel allocation (DCA), and hybrid channel allocation (HCA) as summarised in Figure 2.2.

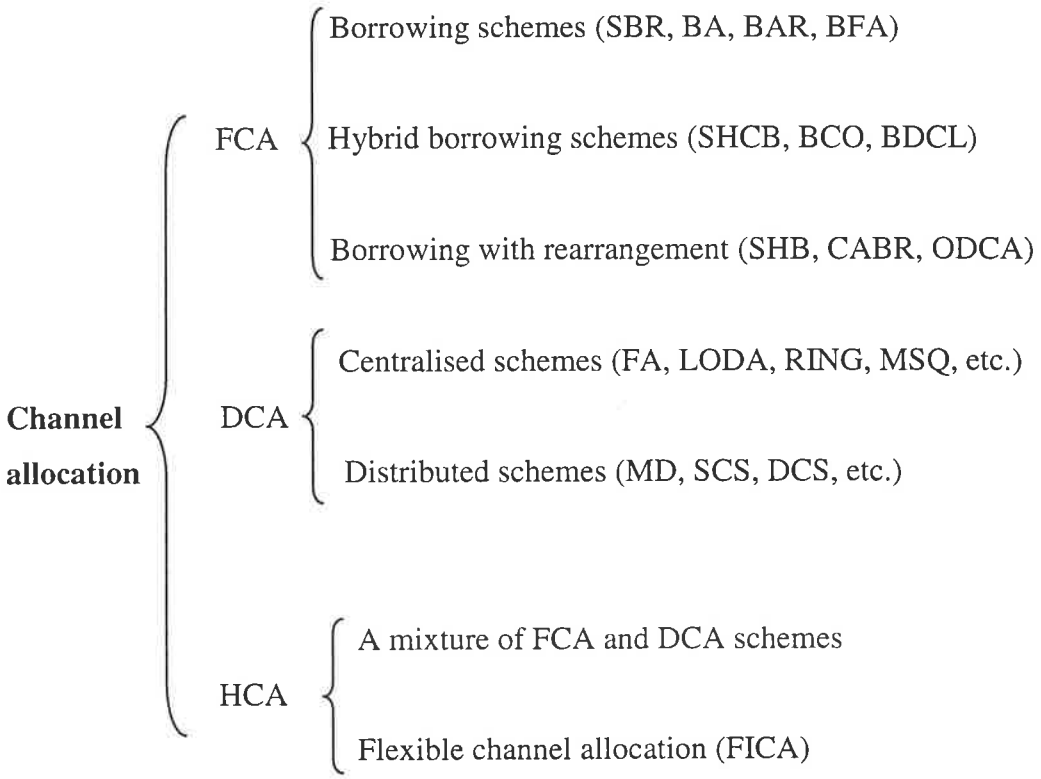


Figure 2. 2: An overview of channel allocation schemes

- **Fixed Channel Allocation (FCA).** For FCA, different cells are permanently assigned with a set of channels. Because the same frequency can be reused at the co-channel reuse distance away, the performance of a cluster of cells can be treated independently. Generally speaking, FCA includes borrowing schemes, hybrid borrowing schemes, and borrowing with rearrangement schemes.

  Firstly, for the borrowing scheme, a channel can be borrowed from neighbour cells, known as donors. After the channel is borrowed, this channel will be locked up and can't be used by the other cells. Specifically, the borrowing schemes include borrow from the richest (SBR), basic algorithm (BA), basic algorithm with reassignment (BAR) and borrow first available (BFA) schemes. Secondly, for the hybrid borrowing scheme, channels are classified into permanent channels and borrowing channels by a priori criteria under different traffic conditions. The hybrid borrowing schemes have three categories, including simple hybrid channel borrowing strategy (SHCB), borrowing with channel ordering (BCO) and borrowing with directional channel locking (BDCL). As an example, for the BCO

scheme, channels are labelled by ranking, where the borrowing order is dependent on a hierarchical rank. The advantage of this channel borrowing scheme is to provide extra channels to the temporally fluctuating offered traffic in the light or medium traffic conditions. However, in heavy traffic, the use of the borrowing scheme may cause increased blocking probability and lower channel efficiency [*Katze96*]. Finally, the borrowing with rearrangement scheme includes sharing with bias (SHB), channel assignment with borrowing and reassignment (CABR) and ordered channel assignment with rearrangement (ODCA) schemes.

- **Dynamic Channel Allocation (DCA).** FCA is usually analysed by assuming uniform traffic conditions and it is also easy to implement. However, FCA is found to be less efficient, especially under non-uniform traffic conditions. Instead, a DCA scheme can be adopted to overcome this problem. DCA has centralised schemes and distributed schemes divisions. The centralised schemes include first available (FA), locally optimized dynamic assignment (LODA), selection with maximum usage on the reuse ring (RING), and mean square (MSQ) schemes, etc. By contrast, the distributed schemes include moving direction (MD), sequential channel search (SCS) and dynamic channel selection (DCS) schemes, etc. The main advantage of DCA is to exploit its resilience to volatile traffic patterns. Due to propagation in small cell systems, cell boundaries frequently overlap. Traffic volatility is hard to predict and the cell layout becomes extremely difficult. As a consequence, the implementation of FCA becomes difficult and thus DCA is used instead. In particular, the forced termination probability can be reduced by using DCA [*Katze96*]. Using the DCA, channels are dynamically allocated to an individual call according to a minimum cost function, e.g., the CIR function. The cost function is dependent on blocking probability, channel usage frequency and reuse distance. As an example, a simple but tractable algorithm called maximum packing is proposed in [*Everitt83*] and [*Everitt89b*]. The basic idea of maximum packing policies is to determine the minimum number of channels for the progressing calls within all cells at a given time. This policy assumes that a new call will be blocked only if it is impossible to reallocate channels to calls. After the call completes, the channel must return to a central channel pool. Therefore, DCA

is more effective in terms of channel utilisation. However, DCA is more difficult to implement in practice. Although the use of DCA has attracted much attention in recent years, it still remains unclear how to trade off the performance parameters [*Katze96*]. Further study is needed to determine the substantial relationships between the gains and the cost, especially by the use of analytical methods. Adversely, an inappropriate DCA algorithm can lead to poor performance under heavy load conditions [*Everitt94*].

- **Hybrid Channel Allocation (HCA).** In comparison with FCA and DCA, the HCA scheme includes a mixture scheme and flexible channel allocation scheme (FICA). The mixture scheme is a combination of FCA and DCA schemes, while the FICA scheme means that a set of permanent channels is still assigned to cells but some flexible channels are retained for the use of increasing traffic loads. The flexible channels can be measured on a scheduling or predictive basis. In order to achieve the best performance, an optimum ratio of the fixed to dynamic channels must be found. HCA is more suitable to non-uniform traffic conditions. Although HCA can not achieve the high performance of DCA [*Everitt94*], it does have some practical advantages in being easier to implement.

In summary, the choice of these three channel allocation methods in wireless systems will be dependent upon the required QoS and traffic distribution. In particular, complexity and flexibility are two important factors in considering the implementation of channel allocation algorithms. There exist some tradeoffs among these factors. More importantly, channel allocation for packet-access systems is different from that for circuit-access systems. Using the packet access, channels are assigned only when there are packets to be delivered. Most of the studies found in the literature concentrate on channel allocation in the circuit-access systems [*Katze96*]. Therefore, it remains a challenge to apply channel allocation to packet data traffic. In this dissertation, in order to obtain a tractable solution, the FCA is assumed throughout the analysis of traffic integration.

2.1.1.3) Resource Allocation in Multiservice

Although the problem of resource allocation for multiclass services has been addressed extensively in wireline networks, the study of wireless networks still remains open. As

mentioned earlier, performance evaluation in wireless networks has to take into account mobility, prioritised traffic and interference.

For a loss system with single source traffic, the well-known Erlang formula can be applied [*Kelly91*]. For a circuit-mode system with multi-service heterogenous traffic, the equilibrium probabilities can be obtained based upon multi-dimensional birth-death processes [*Purzy95*]. In a recent study, the product form solution is found to be applicable to multiservice CDMA loss networks [*Evans99a*]. However, if queuing mechanisms and handoffs are taken into account for multimedia traffic, the solution is not easily obtained for multiclass and multirate services in a unified queuing system [*Pavlid94b*]. In order to obtain a tractable solution, two classes of services are considered in this dissertation. It is worth noting that this approach is generic and can be extended to the case of multi-services.

2.1.2 Call Admission Control Strategies

In order to achieve the best use of network resources, call admission control (CAC) strategies can be employed in system design. CAC means that certain measures or mechanisms are adopted during call access processing so that better QoS can be achieved. In other words, the use of CAC ensures that blocking and delay do not exceed a pre-specified threshold while new arrivals are still permitted.

As described in [*McMil93*], CAC policies for voice service can be classified as fixed routing, alternation routing, and repacking policies. In fact, such classifications have been widely applied to fixed networks. However, the use of CAC in a multimedia wireless environment is rather new and is worthy of further investigation, because it needs to take both VBR transmission and fluctuating traffic loads into account.

In general, the CAC policies must consider the QoS both at call level and at packet level [*Calleg95*]. The conventional QoS at call level is evaluated by new call blocking and handoff forced termination probability properties. However, at the packet level, packet dropping rate must be taken into account. As a matter of fact, there exist some trade-off relationships between the packet level control and the call level control. For the integration of voice calls and data calls, we can use three admission regions to represent the CAC function as shown in [Figure 2.3].

From a user point of view, it is usually recommended that the handoff users should be handled with priority over the new calls. Because of the use of statistical multiplexing, the reservation concept in multiservice packet data wireless networks can be converted to *bandwidth reservation* rather than the conventional *channel reservation* [*Calleg95*].

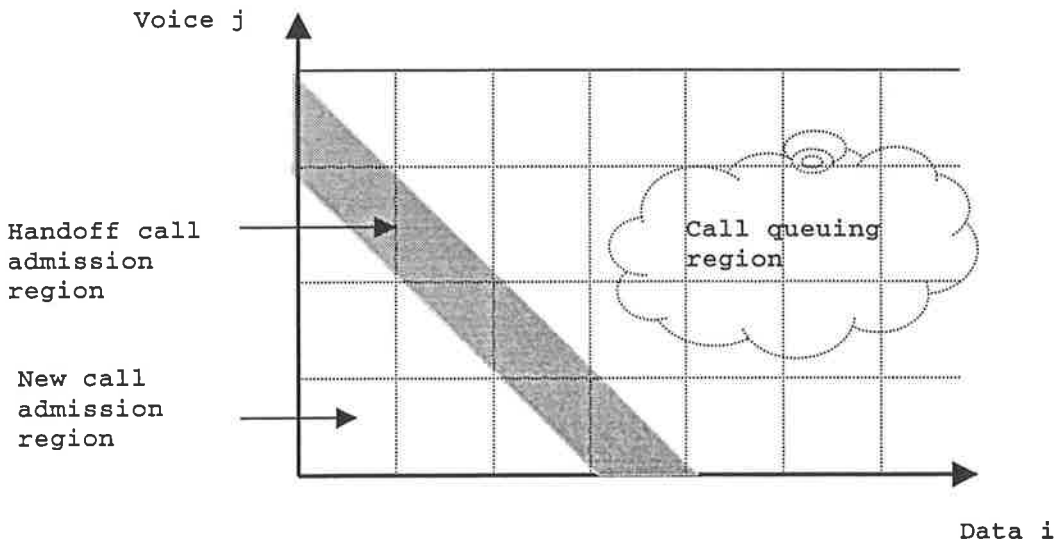


Figure 2. 3: Call admission regions for multiservice

2.1.3 Handoff Call Control and Performance

Due to the movement of terminals, calls are usually differentiated as new calls and handoff calls. Handoff refers to the call process that transfers an ongoing call from one serving station or channel to another while terminals are traversing from one location to another. The main objective for handoff study is to reduce handoff forced probability and improve spectrum utilisation with the use of fast handoff algorithms. In addition, we have to minimise handoff times and the impact of handoff calls on the originating calls. First, we introduce the concept of spatial traffic variability, which is mainly caused by the movement of terminals.

2.1.3.1) Mobility Management

Although the integrated services of voice and data in a wireless environment inherit similar characteristics from a wireline network, the untethered characteristics in a wireless

network add another dimension to the difficulty of network assessment. The reasons for mobility study in wireless systems are summarised in the following.

Firstly, mobility determines cell sojourn distribution and hence affects channel holding distribution. The frequent movement leads to increased handoff times in the target cells. This can dramatically increase the burden of signalling processing. Secondly, the study of terminal mobility can also provide information for location management, e.g., location update and location registration. More importantly, because terminal movement introduces the fluctuation of signal strengths, the investigation of spatial traffic variability is seen to be at the heart of multimedia wireless studies. Finally, the investigation of mobility is the key to the support of seamless multimedia handoff, such as mobile IP.

In some previous studies, an analytical model in the presence of terminal mobility is proposed in [*Hong86*], while a fluid flow model for mobility modelling is used in [*Thomas88*]. In addition, Lin [*Lin94*] uses an excess life theorem to derive the relationship of new call arrival rate and handoff rate. However, as mentioned in [*Zonoo97*], although mobility-related traffic modelling is so important, few in-depth papers have studied the impact of cell sojourn time and channel holding time on resource allocation, especially for multimedia services. Therefore, traffic integration in multimedia services needs to include the investigation of spatial traffic variability.

2.1.3.2) Handoff Call Control Schemes

Due to the degradation of a vulnerable radio link, the power level of channels must be constantly measured so that a terminal which moves away from a base station can be detected in time. This area, where a call changes during its whole holding time, is usually regarded as a handoff area. Call dwell time in the handoff area is dependent on vehicle speed, moving direction, cell structure as well as signal availability. Note that a handoff call is required to complete the handoff process during the cell overlap regions. The lack of channels in the target cells or the breakdown of the signalling link can cause the failure of the handoff call.

The loss properties of handoff calls can be measured by handoff failure probability. In order to derive the distribution of handoff requests in a queue, the random nature of handoff calls requires some probabilistic measures to evaluate their performance, such as handoff attempt failure and forced termination probability. Handoff attempt failure rate refers to the proportion of handoff requests that fail to gain access to a channel in the new

BS, while forced termination probability is defined as a fraction of a call, which is not blocked but eventually incomplete. In a small cell system, even if the subscribers are insensitive to the number of handoff times, the forced termination probability is seen as a critical issue. In addition, another important performance measure for handoff is called incomplete probability, which characterises the proportion of a call either blocking or forced termination [*Hong86*]. Therefore, the failure of handoff calls and the blocking of new calls are treated differently in this dissertation.

From a traffic study point of view, the way to improve handoff performance is by means of handoff control schemes. Generally, prioritised handoff is seen as a promising technique. There are four popular prioritised handoff schemes, which are described as follows.

The first one is guard channel assignment. It is known that there are two kinds of handoff control policies. One is the non-reservation scheme and the other one the reservation scheme, which are differentiated by whether a certain number of channels are exclusively preserved for handoff calls or not. In the non-reservation scheme, the BS handles handoff requests in the same manner as new calls.

In fact, the concept of guard channels in wireless systems originated from the trunk reservation scheme or cut-off priority scheme [*Fisch76*]. It was then introduced into cellular networks in the middle of 1980s [*Posner85*] and [*Hong86*]. Considering only a limited number of channels, if new calls are accepted more often than the proportion of handoff calls, the handoff calls can be seriously interrupted. This will cause an increase in forced termination of voice calls. Under such a circumstance, a reservation scheme can be used to favour handoff calls and also to provide biasing against new arrival calls. However, if too many channels are reserved, this can result in the rejection of new calls and the reduction of total carried traffic. It will lead to a reduction of spectrum efficiency while handoff arrival loads are low, because new calls have less chance to gain access to the available channels. This requires us to determine the optimal number of guard channels, which needs prior knowledge of traffic patterns as well as channel occupancy distribution. In addition, guard channels can be applied in the aforementioned FCA. However, for HCA and DCA, there is no guard channel concept.

The reservation scheme improves handoff failure probability but at the expense of total carried traffic and new call blocking probability. The solution for this is to allow new calls to wait [*Guer88*] and [*Daigle92*]. A series of studies for guard channels can be found in

[*Guer88*], [*Daigle92*], [*Tekinay92*], [*Yoon93*] and [*Lin94*]. However, most of these studies concentrate on the voice services only. The impact of reservation scheme on the performance of integrated services needs to be further investigated.

The second priority scheme is called handoff request queued. Handoff calls are allowed to stay in the queue if there are no available channels in the new BS [*Hong86*] and [*Tekinay92*], which is known as request waiting. The maximum waiting time in the queue is the interval, when a communicating vehicle sojourns in the handoff zone. In the meantime, the old BS channels are not released until channels in the new BS become available [*Lin94*]. Once the power level of a terminal is lower than the threshold, a handoff call has to switch the handling channel to the new BS. As a tradeoff, the use of queuing handoff request will increase the new call blocking probability and reduce the ratio of carried to admitted traffic because new calls can not gain access to channels until the completion of the waiting handoff calls.

Another handoff control scheme called measurement-based prioritization scheme (MBPS) is developed in [*Tekinay92*]. MBPS is a non-preemptive dynamic priority queuing policy, which sorts the order of queuing handoff requests according to power levels. The advantage of using the MPBS is to reduce forced terminated probability and provide better tradeoffs between QoS and spectrum utilisation. The price paid for this is the increase of call blocking probability and the reduction of carried traffic under a certain capacity.

The last handoff control scheme is the sub-rating scheme, in which a full-rate occupied channel is temporarily split into two half-rate channels [*Lin96b*] and [*Ivan98*]. One half serves for the existing call and the other half for the arrival handoff request. Therefore the forced termination probability can be improved but at the expense of the temporary degradation of voice quality.

Apart from the handoff as mentioned, which is commonly known as hard handoff, the other is called soft handoff [*Wongd97*]. Hard handoff refers to a definite decision made on whether to hand off or not. Hard handoff is initiated and executed while a user does not attempt to have simultaneous traffic channels communicating with two BSs. Conversely, soft handoff refers to a conditional decision made on whether to hand off or not. During the process of soft handoff, mobile users can simultaneously communicate with more than one BS. The soft handoff procedure is shown in [Figure 2.4], where the active set refers to

the set of BSs involved during handoff, and the discard set is defined as the set which is going to drop out from the active set.

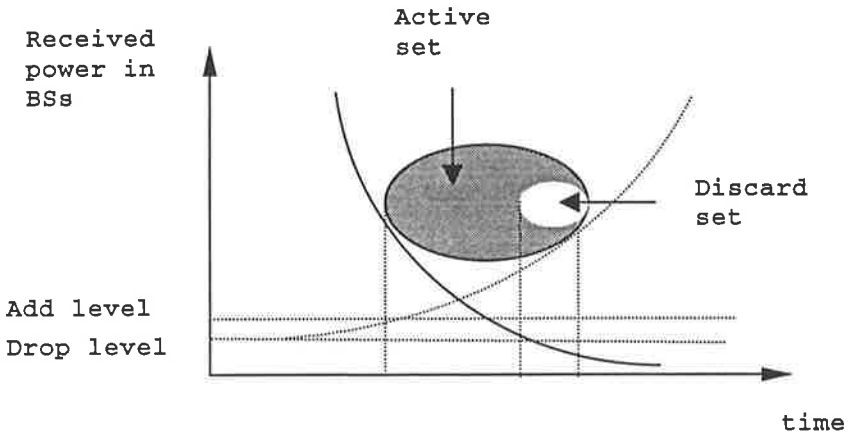


Figure 2. 4: The degradation of soft handoff

Compared with hard handoff, which can only use one of the traffic channels between two adjacent BSs, soft handoff simultaneously uses both traffic channels of two BSs. Therefore, soft handoff is superior to hard handoff in terms of signal quality [*Wongd97*]. Soft handoff has three main advantages. Firstly, it can reduce the flip-flopping effect that commonly occurs during hard handoff. Secondly, soft handoff can eliminate hysteresis margin. Finally, the use of soft handoff can improve uplink capacity. However, soft handoff can bring some disadvantages, such as increased downlink interference and required precise power control algorithms. As a result, extra network resources may be required to support the call process. Soft handoff is exclusively used by CDMA systems, while hard handoff is used in TDMA systems. In particular, baton handoff in future systems may support both soft handoff and hard handoff. However, in this study, our analysis is devoted to hard handoff only because TDMA systems are the main focus.

In summary, in order to have an insight into the impact of handoff call on resource allocation, the inclusion of handoff control studies in traffic performance evaluation is essential. Each handoff call control scheme has its own merits and tradeoff relationships. The choice of these schemes depends upon traffic statistics and design preference. In principle, due to the requirements of control signalling, excessive handoff times should be avoided during network designs. For the fixed guard channel schemes, the improvement of

handoff call failure rate is at the expense of new call blocking probability. Therefore, a more flexible handoff control scheme is proposed in Chapter 4.

2.1.3.3) Handoff Access Scenarios

From the viewpoint of call access, different systems have different handoff call access scenarios. In Digital European Cordless Telephone (DECT), Personal Handyphone System (PHS) and the second generation cordless telephone (CT-2), terminals initiate a request access to the common signalling channel and then gain access to traffic channels. In contrast, the initial call directly gains access to available traffic channels in the Bellcore Wireless Access Communications Systems (WACS) [*Lin94*]. The traffic channels will not be released until either the call is completed or the terminal moves out of the cell.

Three classifications for handoff access control are summarised as follows:

- **Network-controlled Handoff (NCHO).** In this scheme, the network monitors the signal strengths of terminals and then makes a handoff decision once the signal quality deteriorates to lower than a predefined threshold. This scheme is used in Advanced Mobile Phone Services (AMPS), Total Access Communications System (TACS) and Nordic Mobile telephone (NMT) systems.

- **Mobile-controlled Handoff (MCHO).** By using this method, the MS measures signal strength and then makes a subscribing decision to the best BS. MCHO is the highest degree of handoff decentralisation and is adopted by DECT, WACS and PHS systems.

- **Mobile-assisted Handoff (MAHO).** This is a compromise between network-controlled handoff and mobile-controlled handoff. The network asks the mobile terminal to measure signal strength and interference from surrounding BSs. Then the MSC makes the final decision. This is used in GSM systems. In IS-95, MAHO is used in combination with soft handoff [*Tripa98*].

With the increase of decentralised controls from NCHO, MAHO to MCHO, handoff delay decreases from several seconds to milliseconds. However, as a tradeoff, the

measurement information for making a handoff decision also decreases. In this dissertation, only the handoffs associated with MAHO in a GSM-like environment are considered.

In summary, delay and packet loss are seen as the critical factors for seamless handoff in future multimedia services. Fast handoff procedures are required in packet transmission. To improve handoff call dropping probability, call control strategies must be enforced.

2.2 Performance Measurement Methods

As mentioned early, traffic performance is based upon the use of both traffic measurement methods and channel access techniques. In this section, performance measurement methods are described as follows, while channel access techniques are discussed in the next section.

The early studies for integrated services in a wireline network date back to the 1960s [*Gimpel65*]. Because of the discovery of the on-off speech characteristics in [*Brady65*], the time-assigned speech interpolation (TASI) technique is used for the transmission of data packets during the silent speech gaps. With and without considering speech detection, the performance of voice and data in wireline hybrid networks has been analysed in [*Bhat76*], [*Fisch77*], [*Weinst78*], [*Weinst80*], [*Feld82*], [*Sriram83*] and [*Willi84*]. Other integrated studies have appeared in [*Kraime85*] and [*Serres88*].

To enhance mixed traffic performance, queuing rules can be applied. The first one is to use first-in-first-out (FIFO) discipline, which means scheduling the insensitive traffic into buffers according to the order of arrivals. Using the FIFO discipline in integrated services, voice call blocking probability will be increased [*Schwartz87*]. The reason is that the queuing mechanism increases the chance for data traffic to compete with voice calls for the same channel. Another discipline is called pre-emptive priority (PP), which allows voice calls to preempt data packets when voice arrivals find channels full. Meanwhile, the preempted data packets are allowed to queue in accordance with the FIFO discipline. The detailed analysis of these schemes can be found in [*Fischer77*] and [*Schwartz87*]. The advantage of using the PP scheme is to reduce blocking probability for voice calls. However, as a result of tradeoff, packet delay will become prolonged. Compared with the FIFO discipline, the use of the PP scheme would not affect voice call blocking probability because data packets cause no interference to voice calls. Both the FIFO and PP

disciplines are adopted as the queuing rules in this dissertation. Alternatively, other queuing rules, such as last-in first-out (LIFO) or pushout schemes can be used [*Yoon93*].

For integrated services, it is easier to evaluate voice call performance than data traffic because voice calls are slowly varying but with relatively long holding times. The estimation of voice call performance can be derived from the conventional Erlang formula or a variant form. However, it is more difficult to derive the solution for data performance, because data packets are relatively short in length and bursty in nature. More importantly, in order to maximally exploit channel utilisation, data traffic is packetized and its performance is actually dependent upon the instantaneous number of voice users as well as its own offered traffic.

Generally, there exist two kinds of methods in modelling integrated traffic, namely continuous-time models and discrete-time models. For data packet traffic, continuous-time models for integrated services are used in [*Fisher79*], [*Willi84*] and [*Zhang90*]. This method is based on the assumption that the frame length is much smaller than both the service times of voice calls and data calls, while the service time is assumed to be exponentially distributed. In contrast, by using the discrete-time method, the slot structure has to be explicitly considered [*Sriram83*]. In particular, the analysis of the discrete-time model is valid for any general service time distribution. Interestingly, once data loads become high, the difference between these two methods becomes negligible. In the development of traffic analytic algorithm, continuous-time models for integrated services are adopted in this dissertation.

Although simulation can be used to estimate the performance of data packets, the cost of the simulation will increase according to the reduction degree of packet length. In addition, the difficulty of the '*exact*' analytical solution lies in the interaction relationship between voice and data traffic and the discretization of the frame structure [*Garcia82*]. For the sake of simplicity, in order to proceed with the analysis of a large number of channels, approximation approaches are usually adopted. Briefly speaking, the basic idea of these approximation methods is to simplify the two-dimensional (2-D) Markov chain into two one-dimensional (1-D) Markov chains. In other words, the interactive process is approximated as a non-interactive process. Because of such simplicity, the approximation technique is widely adopted in traffic study. Specifically, there exist two approximation approaches for solving data traffic performance, which are described in the following.

Firstly, because the change in the number of voice calls in service will consequently cause an abrupt change to the number of packets, the service rate of data packets varies constantly according to time. As a result, the average performance of data packets will become no better than that under the fixed service rate [*Garcia82*]. In particular, if the period of voice calls remains relatively long, this state can be regarded as quasi-steady. Therefore, the marginal distribution can be used to approximate the behaviour of data packets, which is also known as the quasi-static approximation method. However, this method will become invalid if the instantaneous utilisation exceeds unity and the state of the process becomes unstable [*Garcia82*]. In other words, once an overload period occurs, the use of the quasi-static approximation method becomes unfounded [*Zuker89*].

Solving this problem leads to the second important approximation technique called fluid flow approximation, which can be used while the utilisation exceeds unity [*Garcia82*], [*Gaver82*] and [*Schwartz87*]. Using this approximation method, the analysis for the overload region can be found in [*Gaver82*]. The basic idea of the fluid-flow approach is to treat the service of data packets as a deterministic service. In fact, the fluid flow approximation is based upon the original concept of machine breakdown, in which the breakdown period is equivalent to the overload period and the working period equates to the stable period. In the end, the $GI/G/1$ queue can be used to determine the average waiting time [*Garcia82*]. On the other side, for the underload condition, an ad hoc fluid flow approximation is used in [*Schwartz87*].

As mentioned in the first chapter, because it is difficult to use the conventional GFM in solving complex problems and the conventional iterative methods are also very slow to converge, the quasi-static approximation method is adopted and then developed in Chapter 3, Chapter 4 and Chapter 5. Moreover, because the MAM can converge very fast and provide extremely accurate results for most unstable models [*Latou93*], this method is then developed as an important analytical tool for the analysis of HSD integrated services in Chapter 4.

2.3 Mixed Traffic Performance in Integrated Services

The ultimate aim of wireless communications is to provide a wide range of data services and video services along with voice services. The study of traffic integration needs to take both different QoS and service mechanisms into account, while resource-

sharing methods should be handled according to the different characteristics of each traffic class.

According to different service objectives, quality of service in multimedia communications is classified into user, application, device, and network levels respectively. Because traffic characteristics directly affect network performance, the QoS for both user and network is of importance in performance evaluation. Network performance is seen differently from the viewpoints of operators and users. Integrated services need to provide both users' and networks' QoS with highly efficient resource utilisation. More specifically, the users' QoS refers to the performance parameters, which can be perceived by different users, irrespective of the network-perceived aspects. There are two categories for the users' QoS in integrated systems, i.e., one for real time traffic and the other for non real time traffic [*Gruber83*], which are described in the first chapter.

For performance evaluation, Rappaport and his colleagues have conducted many studies in this area. For example, prioritised handoff traffic has been studied in [*Hong86*]. Subsequently, a priority oriented channel access scheme for low mobility and high mobility users is considered in [*Hong89*]. A single platform supporting multiple calls with multiple handoff attempts is analysed by a multidimentional birth-death process in [*Rapp91*]. Multiple calls within mixed platforms are then investigated in [*Rurzy95*]. Theoretical performance shows trade-offs among blocking, forced termination probability and carried traffic. For the studies of new calls and handoff calls using a guard channel scheme, analyses have appeared in [*Guerin88*], [*Daigle92*] and [*Keil95*]. However, all these studies are limited to voice service only.

Emerging data services are becoming important and inevitably require traffic integration to play a critical role in future wireless networks. Although the aggregate analysis for different traffic classes is essential for determining overall channel performance, the study of the segregation for each service is more important because the distributions of each service determine resource allocation, channel utilisation and even service charging.

As previously stated, apart from the use of measurement methods, performance evaluation in integrated services is subject to the use of channel access methods. In order to transfer data services efficiently, different kinds of access control methods for voice and data integrated services are classified as in [Figure 2.5].

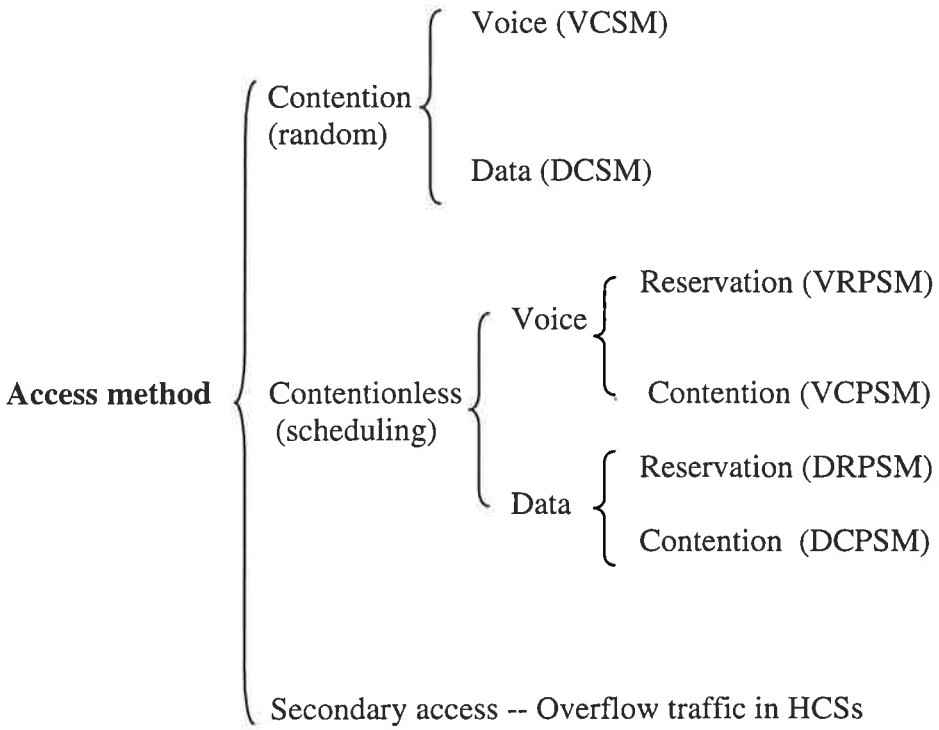


Figure 2. 5: The different channel access methods in IWSs

The contention(less) method can be further classified according to the individual voice and data access methods. For instance, Wong [*Wong92*] investigates the integrated voice and data services using reservation mechanisms in PRMA systems. Later on, integrated services with reservation voice and contention data are studied by Nanda [*Nanda94*]. A protocol called reservation-code multiple-access (RCMA), in which voice services use the reservation mechanism for the spreading code and data services use the contention scheme, is analysed by Tan [*Tan96*]. Recently, Nagatsuka [*Naga98*] analyses the contentionless integrated services that employ a reservation-type packet switching method (RPSM) for data traffic and a circuit switching method (CSM) for voice traffic.

Based upon different system architectures, we review some previous studies in different integrated systems as follows.

2.3.1 Performance Analysis in Time Access IWSs

In contrast to the early investigation of wireline networks, some studies for cellular networks have appeared in [*Stern90*], [*Zhang90*], [*Mitrou93*], [*Li94*], [*Pavlid94b*] and [*Calin98b*]. More specifically, Mitrou [*Mitrou93*] presented a call acceptance algorithm

for voice and data integrated services in a microcell system, in which the delay-blocking relationship is derived from a simulation study. However, mobility and handoff studies are excluded from that study. Pavlidou [*Pavlid94b*] studies a two-dimensional model for the integrated services. Although handoff calls are taken into account, the solution using the conventional GFM is not easy to obtain. Based upon the TIA standard (IS-54), an Enhanced-TDMA (E-TDMA) system for integrated services is analysed in [*Li94*], where traffic variability is not taken into account. Other integrated studies can be found in [*Wilson93*], [*Chang94a*], [*Pavlid94a*], [*Wiese95*], [*Calin97a*] and [*Calin97c*]. More importantly, we observe that these studies only concentrate on the low data rate services. Recently, although some investigations of high data rates have appeared in [*Calin97b*] and [*Calin98*], these studies are limited to the circuit mode services only. In addition, a nonpreemptive priority control policy is used to analyse the performance of HSCSD in [*Calin97b*].

Since the delivery of packet data requires fast and efficient random access, the important issue in packet-switching design is the choice of a suitable MAC [*Wilson93*]. MAC must be designed to share the transmission medium in an efficient and fair manner. As a consequence, the use of different MAC protocols can have significant impact on packet channel allocation.

It is known that CBR mode TDMA is already supported by the present GSM system. In order to provide for efficient transmission, VBR mode can be adopted. The Dynamic TDMA (D-TDMA) system can provide such an operation [*Wilson93*]. In fact, D-TDMA is originally proposed for satellite to ground transmissions [*Falk83*] and [*Gruber83*]. For the D-TDMA protocol, time is organised as a contiguous sequence of TDMA frames. The frame structure are divided into request slots R_q, voice slots N_v and data slots N_d respectively. Furthermore, the request slots are segmented into guard times G_t, overhead and request data. The guard time is used for the propagation delay, where the request data is used for the user ID in voice calls and the address in data packets [Figure 2.6].

Channel access in D-TDMA systems adopts a contentionless mode, which means data packets can occupy the remaining slots left over by voice calls according to some priority orders. For operation in the D-TDMA systems, a mobile terminal initialises a request signal in one of the request slots. If the acknowledgment signal is not received, then corruption will consequently occur. Accordingly, voice calls will retransmit another

request signal in the subsequent frame. In the end, voice calls become lost after a certain number of retries or time out. Data packets are allowed to queue in buffers after packets fail to gain access to channels. In this dissertation, this flexible protocol becomes the focus in the investigation of integrated services.

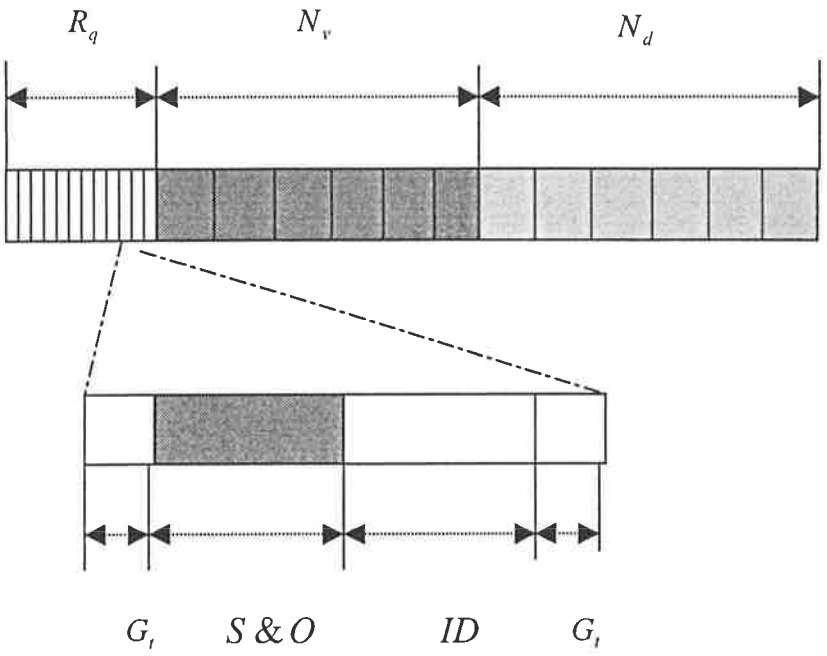


Figure 2. 6: The frame structure of D-TDMA

As previously stated in Section §2.1, because the integrated services intend to provide a wide range of services, the provision of the integrated services poses a severe challenge for network resource allocation. How to utilise the limited resources for different types of users efficiently becomes a key issue. The use of voice calls and data packets dynamically sharing packet data traffic channels can maximise channel utilisation. This is known as Capacity on Demand (CoD) in GPRS [*ETSI364*]. However, taking into account traffic variability and handoff, the trade-off relationships of QoS for different services still remain unclear and need to be further investigated.

As a special example of R-TDMA, the proposed Packet Reservation Multiple Access (PRMA) employs the Reservation-ALOHA (R-ALOHA) protocol, which is first proposed by Crowther in [*Crow73*]. The major feature in PRMA is that the delay of speech packets is bounded in order to guarantee a minimum QoS for speech [*Goodm90*]. In fact, PRMA is

a channel access protocol that combines both TDMA and Slotted ALOHA [*Goodm89*]. In addition, PRMA++, in which slots are further classified as R-slots and I-slots, is known as Advanced TDMA (ATDMA) by some authors [*Urie93*].

PRMA was originally proposed for voice services only. The early studies appeared in [*Goodm89*], [*Goodm90*] and [*Goodm91*]. Because the analytic methods in integrated services are the keys to performance evaluation, many studies have been conducted in this area. Compared with the use of approximate equilibrium point analysis (EPA) in [*Fukuda83*], Wu analyses the integrated voice and data services by a Markov chain technique and claims to have more precise results [*Wu94*]. Meanwhile, Qi uses two approximate methods to analyse the integrated services. The first one is to use a combination of Markov chain and EPA, while the second one adopts the approximate method of marginal distribution. However, it mainly concentrates on the effect of channel error on system performance [*Qi96*]. In addition, Nanda [*Nanda94*] uses elementary catastrophe theory to study the stability of joint voice and data systems.

The other integrated data services into speech transmissions are subsequently investigated in [*Nanda91*], [*Wong92*], [*Wong93a*], [*Nanda94*], [*Wu94*] and [*Qi96*]. In order to combat the instability and improve spectrum efficiency, a non-collision protocol PRMA has been proposed in [*Wen95a*] and [*Ren98*]. However, packet congestion can occur as the number of users increases. Therefore, further investigation is required.

2.3.2 Call Control Procedures in Integrated Systems

Besides TDMA systems, Direct Sequence Code Division Multiple Access (DS-CDMA) is another promising technique in spread-spectrum communications for wireless networks. Because CDMA systems can use the same whole spread bandwidth, resource allocation in CDMA systems is different from TDMA systems and is interpreted as power assignment and code assignment. The comparison of TDMA and CDMA systems can be found from the simulation study in [*Wilson93*]. More importantly, the capacity in CDMA systems is limited by the total multiple access interference (MAI), which includes both intercell interference and intracell interference [*Viterbi93*]. For performance measures in multiuser systems, Erlang capacity, which is defined as the average traffic load in terms of average number of users for servicing, determines blocking probability [*Viterbi93*]. Another important parameter is called loss probability, which results from interference-limited characteristics.

The voice-only service is considered in [*Vieter91*] and [*Taylor91*]. The interference-based CAC is then studied in [*Lui94*] and [*Ishika97*]. From the viewpoint of system capacity, an optimal design must take both loss probability and blocking probability into consideration in integrated services. The study of integrated services has appeared in [*Wilson93*] and [*Naga98*]. However, the impact of variations of traffic loads on performance is not analytically considered. As in TDMA systems, the voice and data integrated services in CDMA systems are required to be dynamic as well as efficient. Accordingly, a common packet data channel (CPDC) is employed to transfer short and infrequent packets in [*Guo96*] and [*Das97*]. In fact, both of them employ spread ALOHA techniques, in which a single code is used for multiusers [*Abram94*]. As a result, the overhead of packet transmissions can be reduced. However, for large or more frequent packets, a dedicated channel needs to be assigned to deliver data transmissions.

CAC is crucial in integrated services in determining the QoS and consequently, quality-based dynamic CAC becomes the focus of this dissertation. This study concentrates on investigating how to smooth the fluctuation of interference through quality-based effective CAC in integrated services. Because of data packet characteristics, resource allocation in the integrated queuing systems needs to include the consideration of data delay.

In order to reduce the premature termination of a progressing call, there exist two methods. One is the prioritised scheme, which has been discussed in Section §2.1. The other one is to use the hierarchical cell structures (HCSs). The deployment of HCSs in wireless systems can satisfy the increased traffic demand and also provide seamless coverage. The increase of capacity depends upon the mixed cell structure and the reuse factor [*Tripa98*]. According to different signalling propagation and terrains, there exist different kinds of cell planning structures. For typical overlaid systems, the main categories of cell structures can be classified into concentric cells, macro-micro cells and global access cells, which are described as follows.

- **Concentric Cells**: In this structure, the underlaid cells have lower irradiated power than the overlaid cells [Figure 2.7.1]. The main feature of this structure is that both cells can share the same BS.

- **Macro-micro Cells**: A macrocell can overlay several small cells [Figure 2.7.2]. This scenario is suitable for the high density or hot-spot areas. However, such a system needs to be properly tuned so as to avoid the ping-pong effects between layers.

- **Global Access Cells**: With the aid of the satellite component, global coverage can be achieved [Figure 2.7.3]. Because the signal has to travel a long distance from the terrestrial segment to the space segment, sufficient signal strength and the tolerable delay are crucial in using this structure.

Apart from the cell structures mentioned above, other approaches, such as partial overlapped cells, can be formed by a mix of these three basic structures.

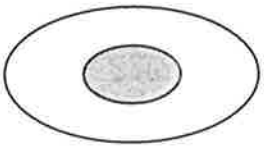


Figure 2. 7. 1

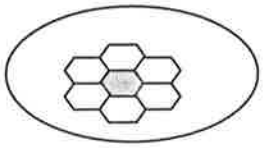


Figure 2. 7. 2

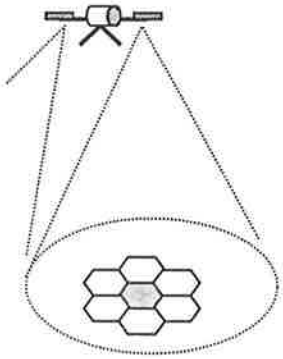


Figure 2. 7. 3

HCSs are able to support both interfrequency handoff and intrafrequency handoff. Traffic flow in HCSs can be either from lower layer to upper layer or vice verse. From the viewpoint of resource allocation, moment matching techniques form the core in the studies of overflow traffic performance. The aim of overflow traffic studies is to reduce the number of handoffs and call dropping probability between layers so as to achieve optimal resource allocation.

The pioneer study about overflow traffic in fixed networks dates back to the early 1950's [*Wilki56*]. Because of the peakedness characteristics of overflow traffic, many papers have conducted such studies under a wireline environment [*Kuczu73*], [*Fred80*] and [*Heffer80*]. Until recently, HCSs have been found useful in applying to the resource-

limited cellular systems. The one-moment matching technique for overflow traffic is used in [*Steele92*] and [*Jabb97*]. However, it is evident that the overflow traffic tends to be bursty if the aggregated arrivals need to be taken into account [*Sriram86*], [*Meier89*] and [*Gusella91*]. Although some recent studies can be found in [*Hu95*], [*Lagran95*] and [*Lagran96*], in which low mobility and high mobility terminals camp on lower and upper layers respectively, the systematic comparison for these modelling methods and how to trade off call handling schemes still remains unclear. In this dissertation, we focus on choosing the adequate moment matching techniques as well as network optimisation methods to enhance overall network performance.

2.4 Summary

In this chapter, we have presented a comprehensive survey of literature in the area of voice-only and voice and data integrated services in dissimilar systems, such as TDMA, R-TDMA, DS-CDMA and HCS systems. Firstly, we observe that better overall network performance and network optimisation can be achieved only if the control strategies can be handled in an efficient manner, especially in an interference-limited environment. Secondly, most of the studies only concentrate on the delivery of low rate integrated services. However, the future wireless data systems need to provide high rate data services for a wide range of emerging data applications, such as wireless Internet and video services. Accordingly, further investigation for high-speed data services is needed. Finally, the study of overflow traffic modelling methods in HCSs is a key to achieve successful global access. Therefore, we focus on the performance of the comparison of modelling methods and optimisation studies. Based upon such considerations, the performance analysis in the multimedia services begins in the next chapter.

Chapter 3

The Performance of LRD in Integrated Services

Radio resource management and integrated services are seen as the keys to the delivery of multimedia services. In this chapter, the multimedia diverse source models and the performance of low rate data (LRD) in integrated services are investigated in detail. Firstly, integrated radio resource management is studied in Section §3.1. Subsequently, an approximation method is developed to analyse the integration of LRD with voice service in Section §3.2. The summary for each section appears at the end of each section.

3.1 Integrated Radio Resource Management for a Voice-only System

As previously mentioned, network radio resource control tasks include channel assignment, admission control and handoff control. Integrated radio resource management aims to optimally trade off system parameters while taking different control aspects into account. As mentioned in the last chapter, if the control strategies are handled in a unified optimal way, better network performance can be expected. In this study, the use of adaptive radio resource management, which includes combining prioritised handoff control and traffic variability, is considered. It shows that call incomplete probability increases according to the increase of traffic variability. Especially, this discernible effect becomes more significant with the increased traffic load. In addition, prioritised handoff control schemes have a significant impact on call incomplete probability. More specifically, in Section §3.1.1, call performance in multiple platforms is investigated. Handoff control policies and timeout models are then discussed in Section §3.1.2. Subsequently, call performance in multiservice loss systems is analysed in Section §3.1.3. Finally, a summary concludes this section.

3.1.1 Call Performance in Multiple Platforms with Multi-Mobility

In order to investigate the influence of traffic variability on call performance, the study of handoff probability needs to be taken into account. Call arrivals are assumed to follow a Poisson process but busy line effect is not taken into consideration. The busy line effect refers to the situation where arrival calls are blocked while the conversation of a destination terminal is still in progress [*Lin96a*]. Lin [*Lin97b*] studied the effect of

terminal mobility on call performance. The excess life theorem of alternating renewal process is used to obtain the probability of handoff times [*Lin94*]. In fact, this is analogous to the inspection paradox theorem, which states that the expected lifetime of a call currently in progress within a renewal process† is longer than the expected lifetime of a typical call [*Goodr88*].

Let the mean call holding time be $1/\mu_v$ and the interarrival time be more than ten times longer than the whole call holding time. Cell sojourn time (CST), is denoted by t_{M_i} with mean $1/\eta$ and t_m denotes the CST in the original cell, where the CST means the duration of a call when a terminal remains in a particular cell before it moves out of the cell or the call terminates. $r_m(t)$ is used to represent the probability density function (pdf) of t_m and $f_m(t_{M_i})$ for the pdf of t_{M_i}. K is the number of handoffs for a call. The effect of cell sojourn time can be shown as in [Figure 3.1].

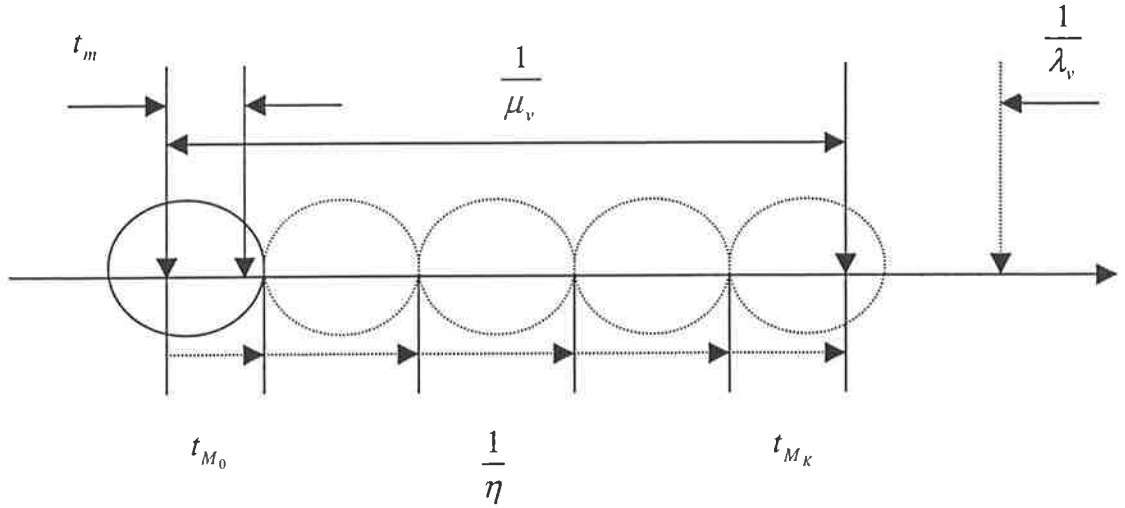


Figure 3. 1: The effect of cell sojourn time

Then we can have the pdf $r_m(t)$ for t_m [*Lin94*]:

$$r_m(t) = \eta \int_{\tau=t}^{\infty} f_m(t) d\tau = \eta(1 - F_m(t)) \tag{3.1}$$

The probability of K handoffs of a call becomes:

† Renew process is a term in renew theory. It states that a process, the variables of which are a sequence of i.i.d. and positive r.v., has a specific distribution function and mean value.

$$\Pr(K) = \Pr(t_M > t_m)\left(\prod_{i=1}^{K-1}\Pr(t_{M,i} > t_{M_i})\right)\Pr(t_{M,K} \le t_{M_k}) \tag{3.2}$$

If call unencumbered time (CUT), which represents the time occupying a channel if no handoff is required, follows an exponential distribution, i.e., $f(t) = \mu_v e^{-\mu_v t}$, we can obtain:

$$\begin{aligned} \Pr(K=0) &= (1-\frac{\eta}{\mu_v}) + \frac{\eta}{\mu_v} f_m^*(s) \quad while \quad k=0 \\ \Pr(K=k) &= \frac{\eta}{\mu_v}(1-f_m^*(s))^2 (f_m^*(s))^{k-1} \quad while \quad k \ge 1 \end{aligned} \tag{3.3}$$

where $s = \mu_v$ and $f_m^*(s)$ is the Laplace-Stieljets Transform (LST) of $f_m(t)$, that is, $f_m^*(s) = \int_{t=0}^{\infty} f_m(t) e^{-st} dt$.

Obviously, the probability of the number of handoffs depends on the variance of the CST distributions. Moreover, the difference of this probability from different distributions, e.g., gamma distribution and exponential distribution, diminishes according to the increased number of handoffs. As a result, the impact of CST distribution on system performance can be described as follows.

3.1.1.1) A Generalisation of CST Distribution

The purpose of using a generalised CST distribution is to accommodate possibly arbitrary mobility patterns and provide more flexibility than the conventional distribution. The CST distribution actually relies on cell physical configurations, mobility patterns and channel propagation errors. For example, the different propagation effects, e.g., small-scale, medium-scale and large-scale effects, can significantly affect the sojourn time. The various effects generate different radio link qualities and thus cause different CST. Accordingly, the study of the general CST distribution is able to provide us with insight into the effect of traffic variability under generalised conditions. In particular, the CST distribution will directly determine channel holding time (CHT), which is expressed by the minimum of the CUT and the CST: $t_{CHT} = \min\{t_{CUT}, t_{CST}\}$. CHT is defined as either the

duration spent in cell before crossing the cell boundary or the time until the channel is abandoned.

The assumption of exponential distribution of CST has appeared in the previous studies [*Hong86*], [*Guer88*] and [*Chang94b*]. However, some recent studies suggest that the CST does not follow exponential distribution and hence the CHT is indeed not exponentially distributed [*Zonoo97*] and [*Fang99*]. Zonoozi [*Zonoo97*] shows that CST can be described by the generalised gamma distribution. Moreover, Lin [*Lin97b*] shows that incomplete probability is a decreasing function of the variance of the CST and the impact of variance on high mobility is more significant than that on low mobility. Two examples of CST distributions are given as below.

3.1.1.2) Examples of CST Distributions

- **Examples of Gamma Distribution of CST**

If the distribution of CST is assumed to be arbitrary, the gamma distribution may be chosen to approximate the shape of the arbitrary distribution [*Zonoo97*]. The reasons for the choice of the gamma distribution are as follows. Firstly, the shape of the gamma distribution varies according to the related parameters. Its mean and variance are not the same as in the case of the conventional exponential distribution. In particular, the exponential distribution is a special case of the gamma distribution while the order of the gamma distribution is equal to one. Secondly, the moments of CST are tractable and easy to compute.

The pdf of a gamma distribution $g_r(t)$ with order r is expressed as:

$$g_r(t) = (r\eta)e^{-r\eta t}\frac{(r\eta t)^{r-1}}{(r-1)!} \quad while \quad t \geq 0$$
$$g_r(t) = 0 \quad while \quad t < 0 \tag{3.4}$$

Then the cumulative distribution function (cdf) $F_r(t)$ of the gamma distribution can be written as:

$$F_r(t) = \int_0^t r\eta e^{-r\eta\tau}\frac{(r\eta\tau)^{r-1}}{(r-1)!}d\tau \quad while \quad t > 0 \tag{3.5}$$

It is known that the mean and variance of the random variable T_t are:

$$E(T_t)=\frac{1}{\eta} \quad , \quad Var(T_t)=\frac{1}{r\eta^2} \tag{3.6}$$

Hence the LST of gamma density function can be expressed by:

$$f_m^*(s)=\left(\frac{r\eta}{s+r\eta}\right)^r \tag{3.7}$$

If we set $s=\mu_v$, then we obtain:

$$f_m^*(\mu_v)=\left(\frac{r\eta}{\mu_v+r\eta}\right)^r \tag{3.8}$$

Finally, the probability of handoff times in formula (3.3) can be rewritten as:

$$\begin{aligned}
&\Pr(K=0)=(1-\frac{\eta}{\mu_v})+\frac{\eta}{\mu_v}\left(\frac{r\eta}{r\eta+\mu_v}\right)^r \quad while \quad k=0 \\
&\Pr(K=k)=\frac{\eta}{\mu_v}(1-\left(\frac{r\eta}{r\eta+\mu_v}\right)^r)^2\left(\frac{r\eta}{r\eta+\mu_v}\right)^{k-1} \quad while \quad k\geq 1
\end{aligned} \tag{3.9}$$

Under the generalised conditions, if t_n denotes new call channel holding time and t_h represents handoff call channel holding time, the expected CHT $E(t_n)$ for new calls can be expressed as [*Lin97b*]:

$$E(t_n)=\frac{1}{\mu_v}(1-\frac{\eta}{\mu_v}(1-f_m^*(\mu_v))) \tag{3.10}$$

However, the expected CHT $E(t_h)$ for handoff calls is given by:

$$E(t_h)=\frac{1}{\mu_v}(1-f_m^*(\mu_v)) \tag{3.11}$$

Therefore, if we assume $f_m^*(\mu_v)\neq\eta/(\mu_v+\eta)$, we can easily obtain this result: $E(t_n)\neq E(t_h)$. This suggests that, once a distribution is with unequal mean and variance,

the expected CHT for new calls and handoff calls can not be treated as the same. This is important because they may have a different impact on overall call performance.

- **Examples of Exponential Distribution of CST**

As a special example of the arbitrary CST distributions, the gamma distribution of CST leads to a simple exponential distribution while $r = 1$. Therefore we can have:

$$f_m^*(\mu_v) = \frac{\eta}{\mu_v + \eta} \quad , \quad while \quad r = 1 \tag{3.12}$$

And then, we obtain:

$$E(t_n) = E(t_h) = \frac{1}{\mu_v + \eta} \tag{3.13}$$

Apparently, the exponential distribution does give a 'smoother' effect on the probability of handoff times than the other orders of the gamma distribution.

The exponential distribution of CST simply leads to the density function $f_{CH}(t)$ of the CHT distribution as:

$$f_{CH}(t) = (\mu_v + \eta) e^{-(\mu_v + \eta)t} \tag{3.14}$$

Although it is simple, this is a useful result in the subsequent analysis.

Taking a prioritised handoff scheme with total channels C and reserved channels C_h into account, we can have the total arrival loads as:

$$\rho_n = \lambda_n E(t_n) + \rho_h \tag{3.15}$$

where handoff loads are $\rho_h = \lambda_{hv} E(t_h)$, λ_n and λ_h denote the arrival rate for new calls and handoff calls respectively.

Therefore, if P_0 denotes a normalising constant, we can have the new call blocking probability P_B as [*Hong86*]:

$$P_B = \sum_{j=C-C_h}^{C} \frac{\rho_n^{C-C_h} \rho_h^{j-(C-C_h)}}{j!} P_0 \tag{3.16}$$

And the handoff failure probability P_{hf} can be written as:

$$P_{hf} = (\rho_n^{C-C_h} \rho_h^{C_h} / C!) P_0 \tag{3.17}$$

In the end, we can have the call incomplete probability P_{nc}:

$$P_{nc} = \frac{\lambda_v P_B + \lambda_{hv} P_{hf}}{\lambda_v} \tag{3.18}$$

As a numerical example, the result for the different kinds of mobility platforms with prioritised handoff schemes in a 500 meters radius microcell is shown as in [Figure 3.2].

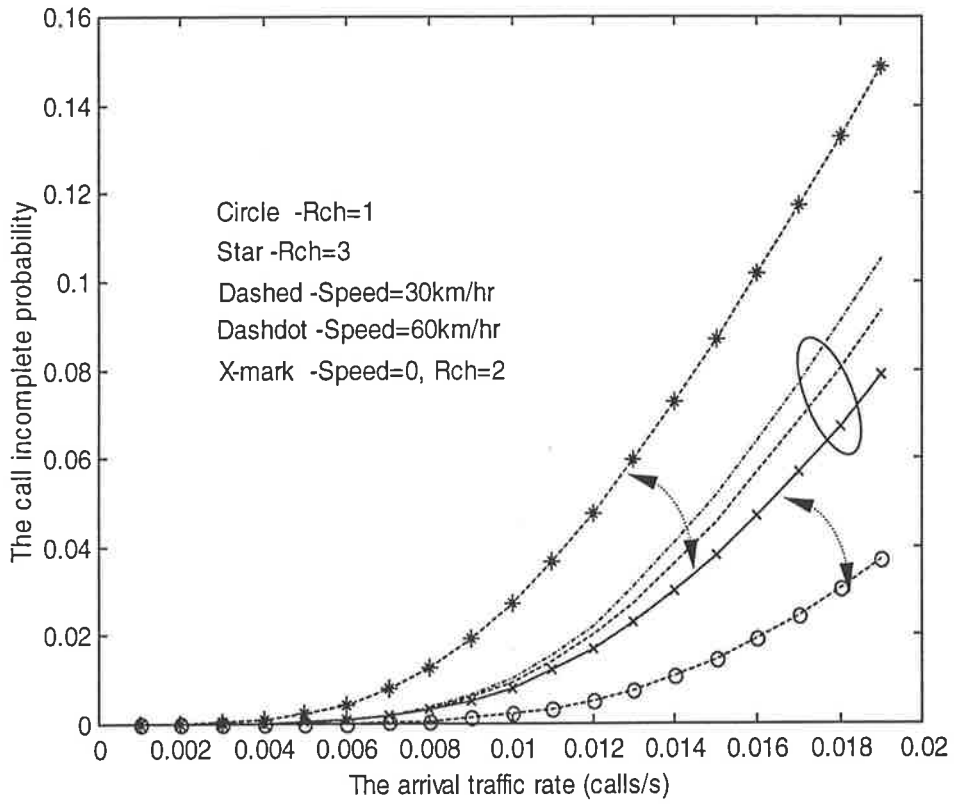


Figure 3. 2: Call incomplete probability

It is known that the new call blocking probability is insensitive to the different mobility patterns. However, from this figure, we observe that the call incomplete probability is sensitive to traffic variability, which is caused by the movement of terminals. For example, if the terminal moves from 0 to 30 km/hr or to 60 km/hr with the input arrival rate of $\lambda_v = 0.012$ calls/s, the incomplete probability increases by 18%, or by 31%

respectively. Therefore the effect of mobility patterns on call performance is indeed significant and thus not negligible, especially in a small cell environment.

If we change the reserved channels in prioritised handoff schemes, we observe that the effect of the number of reserved channels on call incomplete probability is more significant than that of terminal mobility. For example, if the arrival rate is $\lambda_v = 0.012$ calls/s with reserved channels $C_h = 2$ (shown by $R_{ch} = 2$ in Figure 3.2), the incomplete probability increases by 179% for the case of reserved channels $C_h = 3$ and reduces by 69% for the case of reserved channel $C_h = 1$. The reason for this is that the number of reserved channels significantly affects not only the blocking probability of new calls, but also handoff call failure probability.

As a summary for Subsection §3.1.1, the CST distribution has a significant impact on call incomplete probability. Because the distribution of CST is subject to terminal movement and propagation factors, the investigation of the generalised distribution of CST is important. The general distribution may be expanded to Gaussian (Normal) distribution, or even Lognormal distribution. In addition, the study of CST can be applied to location tracking as well as location update [*Lin97a*].

3.1.2 Handoff Call Control Policies

As mentioned above, both priority handoff control schemes and terminal movement have a significant impact on call performance. In fact, there exist different handoff control strategies, which can be used flexibly according to different requirements of QoS. Meanwhile, coupled with queuing rules, there are many alternatives as well. For example, a non-reservation scheme called prioritised pushout is proposed by Yoon [*Yoon89*]. The first-in first-out (FIFO) with a head-of-line (HOL) priority scheme is used for handoff calls, whereas the LIFO or FIFO discipline applies to originating calls. This prioritised pushout scheme without guard channel can reduce the originating call blocking probability and exhibit better performance than the guard channel one [*Yoon93*]. In addition, the LIFO pushout scheme can yield better waiting time distribution for the originating calls than the FIFO pushout scheme does. However, by the use of such schemes, the addition of signalling needs to be taken into account in actual network implementation.

Except for the pushout scheme and the pre-emptive priority scheme, which has been mentioned in Chapter 2, the other alternative schemes can be described as follows.

3.1.2.1) A Handoff Delayed Model

The prioritised handoff scheme for wireless networks is originally proposed by Posner [*Posne85*]. It is found that the use of the priority handoff scheme can reduce handoff call blocking while paying a small penalty for new call blocking probability. The FIFO discipline with unlimited queuing mechanism is subsequently analysed in [*Hong86*], [*Guer88*], and [*Lin94*]. In addition, Tekinay [*Tekina92*] considers the case of infinite storage rooms without terminal movement. As a result, the use of infinite queue can introduce long average waiting time. Long delays can cause impatient customers to drop out of services. Because long waiting times are undesirable to users, it is necessary to improve delay performance by investigating waiting time distribution.

Taking terminal movement and handoff into account, the standard waiting time distribution [*Tekina92*] can be modified as below.

In order to proceed to the analysis, priority is given to handoff calls over originating calls. In this study, the service rate is assumed to be $\mu_{CH} = \mu_v + \eta$ and also the rate of waiting time in queue $\mu_Q = \mu_{CH}$, while the arrival rate is $\lambda_1 = \lambda_v + \lambda_{hv}$. If the FIFO discipline is applied to queuing, then the waiting time distribution can be evaluated as follows.

The probability of a call that does not need to wait in queue is an arrival customer sees less than C users already in service:

$$P(W_Q = 0) = \sum_{j=0}^{C-1} P_j = \sum_{j=0}^{C-1} \left\{ P_0 \frac{1}{j!} \left(\frac{\lambda_v + \lambda_{hv}}{\mu_v + \eta} \right)^j \right\} \tag{3.19}$$

The cdf of unconditional waiting time can be written as:

$$P(W_Q \le t) = P(W_Q = 0) + F(t \mid W_Q > 0) P(W_Q > 0) \tag{3.20}$$

Hence the cdf of waiting time under the condition of positive waiting time becomes:

$$F(t \mid W_Q > 0) = \sum_{k=0}^{\infty} F(t \mid j = C + k) P(j = C + k \mid j \ge C) \tag{3.21}$$

Therefore,

$$F(t \mid W_Q > 0) = \sum_{k=0}^{\infty} \left\{ \int_0^t g_{k+1}(\tau) d\tau \bullet (1 - \frac{\lambda_{hv}}{C\mu_Q})(\frac{\lambda_{hv}}{C\mu_Q})^k \right\} \tag{3.22}$$

where $g_{k+1}(\tau) = C\mu_{CH} e^{-C\mu_{CH}\tau} \frac{(C\mu_{CH}\tau)^k}{k!}$ represents a gamma distribution with service rate $C\mu_{CH}$ and j represents the number of users in a system. Because there are already k customers in the queue ahead of the last arrival, it has to wait until $(k+1)$ customers finish their services before entering the server. Each service is assumed to have an identically, independent distribution (i.i.d.). Thus it can be simplified as:

$$F(t \mid W_Q > 0) = (1 - \frac{\lambda_{hv}}{C\mu_Q}) \int_0^t \left\{ \sum_{k=0}^{\infty} \frac{(C\mu_{CH}\tau)^k}{k!} C\mu_{CH} e^{-C\mu_{CH}\tau} (\frac{\lambda_{hv}}{C\mu_Q})^k \right\} d\tau \tag{3.23}$$

Namely,

$$F(t \mid W_Q > 0) = (1 - \frac{\lambda_{hv}}{C\mu_Q}) \int_0^t C\mu_{CH} e^{-C\mu_{CH}(1-\frac{\lambda_{hv}}{C\mu_Q})\tau} d\tau \tag{3.24}$$

Hence we can have:

$$F(t \mid W_Q > 0) = 1 - e^{-C\mu_{CH}(1-\frac{\lambda_{hv}}{C\mu_Q})t} \tag{3.25}$$

The conditional probability that an arrival exceeds a waiting time t for service becomes:

$$P(W_Q > t \mid W_Q > 0) = e^{-C\mu_{CH}(1-\frac{\lambda_{hv}}{C\mu_Q})t} \tag{3.26}$$

Then we can simply write down the probability of waiting time in a queue as:

$$P(W_Q > 0) = \sum_{j=C}^{\infty} P_j = C(C, \rho_h) \tag{3.27}$$

where $\rho_h = \frac{\lambda_{hv}}{C\mu_Q}$.

The waiting time distribution can be shown as:

$$F_{W_Q}(t) = P(W_Q \le t) = 1 - C(C, \rho_h) e^{-C\mu_{CH}(1-\rho_h)t} \tag{3.28}$$

Finally, the unconditional waiting time distribution becomes:

$$P(W_Q > t) = C(C, \rho_h) e^{-C\mu_{CH}(1-\rho_h)t} \tag{3.29}$$

If the mean service time is assumed to be 3 minutes†, the cell size is assumed to be equal to 10 km and total channels are $C = 8$, we can have the conditional waiting time distribution as shown in [Figure 3.3]. Note that the relationship between the mean CST and the average speed can be found in Section §4.1.

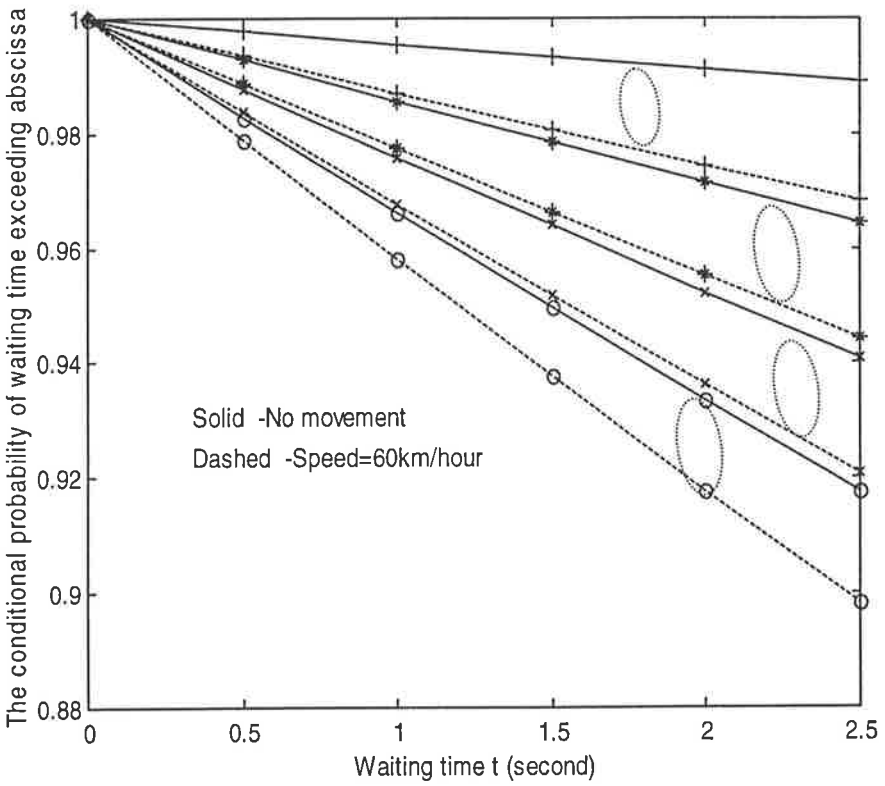


Figure 3. 3: Waiting time distribution with terminal mobility

† In some literature, the average call holding time is assumed to be only two minutes. We use both of the assumptions in this dissertation.

From this figure, we observe that a large number of arrivals incur longer delay time and terminal mobility has a discernible effect on the waiting time distribution, especially when the terminal movement increases. Faster mobility actually reduces the time of waiting. Such reduction increases according to the increased allowable waiting time in the queue. For example, if terminals move from 0 to 60 km/hr, waiting time distribution reduces to 0.42%, 0.84%, 1.26%, 1.68% and 2.1% as the waiting time increases from 0.5s, 1s, 1.5, 2s and 2.5s respectively. This suggests that increased mobility tends to have more impact on the calls having longer waiting intervals. In essence, in order to reduce the waiting probability, either handoff arrival loads or the CHT can be reduced. As another result, although we do not show it in the figure, both smaller cell size and faster movement lead to handoff calls having less chance of waiting.

3.1.2.2) A Handoff Reneging Model

Besides the delayed model, another handoff control scheme is to use the reneging model, in which calls can become degraded according to a specified defected rate in the queue. This situation occurs while the terminal enters handoff areas, where a radio link signal from the current BS degrades because of terminal movement. Once waiting time exceeds the maximum queuing time, the call is regarded as lost.

McMillan [*McMil93*] studies a M/G/1 finite queue with impatient users, in which the variance of service time distribution is taken into account. Lin [*Lin94*] has investigated a defection model in a M/M/1 infinite queue and then derives forced termination probability for handoff calls. In this study, the arrival rate is assumed to be: $\lambda_1 = \lambda_v + \lambda_{hv}$. The service rates are assumed to be: $\mu_{CH} = \mu_v + \eta$ and the rate of waiting time in queue $\mu_Q = \mu_{CH}$, where η represents the rate of CST. The handoff forced termination probability is derived as follows.

Considering an arrival in $(C + j)th$ state at time t, the Markov chain jumps to the next state at time $t + t_j$. It is known that the maximum queue time T_Q is the minimum of the degradation time and the CST. If the interval T_j for the jth arrival is bigger than the maximum queue time T_Q, the jth call will be dropped eventually. Then the dropping probability can be shown as [*Lin94*]:

$$P(T_j > T_Q \mid_{C+j}) = \int_{t_j=0}^{\infty} \int_{t_{j-1}=0}^{\infty} \cdots \int_{t_0=0}^{\infty} \int_{t=0}^{t_0+t_1+\cdots+t_j} \prod_{k=0}^{j} (C\mu_{CH} + k\mu_Q) e^{-(C\mu_{CH}+k\mu_Q)t_k} \mu_Q e^{-\mu_Q t} dt dt_0 \cdots dt_{j-1} dt_j \tag{3.30}$$

Subsequently, we can have:

$$P(T_j > T_Q \mid_{C+j}) = \int_{t_j}^{\infty} \int_{t_{j-1}=0}^{\infty} \cdots \int_{t_0=0}^{\infty} \prod_{k=0}^{j} (C\mu_{CH} + k\mu_Q) e^{-(C\mu_{CH}+k\mu_Q)t_k} (1 - e^{-\mu_Q(t_0+t_1+\cdots+t_j)}) dt_0 \cdots dt_{j-1} dt_j \tag{3.31}$$

Hence,

$$P(T_j > T_Q \mid_{C+j}) = 1 - \int_{t_j=0}^{\infty} C\mu_{CH} e^{-(C\mu_{CH}+j\mu_Q+\mu_Q)t_j} dt_j = \frac{(j+1)\mu_Q}{C\mu_{CH} + (j+1)\mu_Q} \tag{3.32}$$

Apparently, the forced termination probability P_F is the summation of all j in queue:

$$P_F = \sum_{j=0}^{\infty} P(T_j > T_Q \mid_{C+j}) P_{C+j} = \sum_{j=0}^{\infty} \frac{(j+1)\mu_Q}{C\mu_{CH} + (j+1)\mu_Q} P_{C+j} \tag{3.33}$$

Because we already assume $\mu_{CH} = \mu_Q$, we can have the simple form as:

$$P_F = \sum_{j=0}^{\infty} \frac{j+1}{C+j+1} P_{C+j} \tag{3.34}$$

where the equilibrium probability can be obtained from:

$$\begin{aligned} P_j &= P_0 \frac{1}{j!} \left(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}}\right)^j \quad while \quad 0 \le j \le C \\ P_j &= P_0 \frac{(\lambda_v + \lambda_{hv})^C}{j! \mu_{CH}^j} \lambda_h^{j-C} \quad while \quad j > C \end{aligned} \tag{3.35}$$

For the prioritised handoff queuing scheme, we can not simply substitute it with $P_F = P_B$, because it is only valid for the nonprioritised loss system [*Hong86*]. We have to use this algorithm shown as in [Figure 3.4].

Then the handoff rate becomes:

$$\lambda_{hv} = \frac{\eta(1-P_B)\lambda_v}{\mu_v + \eta P_F} \tag{3.36}$$

In order to have the optimal cutoff parameter for C_h, it is necessary to maintain a minimum value for the cost function [*Hong86*] and [*Chang94b*]. The cost function $f_{cost}(C_h)$ is denoted as:

$$f_{cost}(C_h) = (1-\alpha)P_B + \alpha P_F \tag{3.37}$$

where α $(0 \le \alpha \le 1)$ is a weighting factor, which depends upon the significance of P_B and P_F.

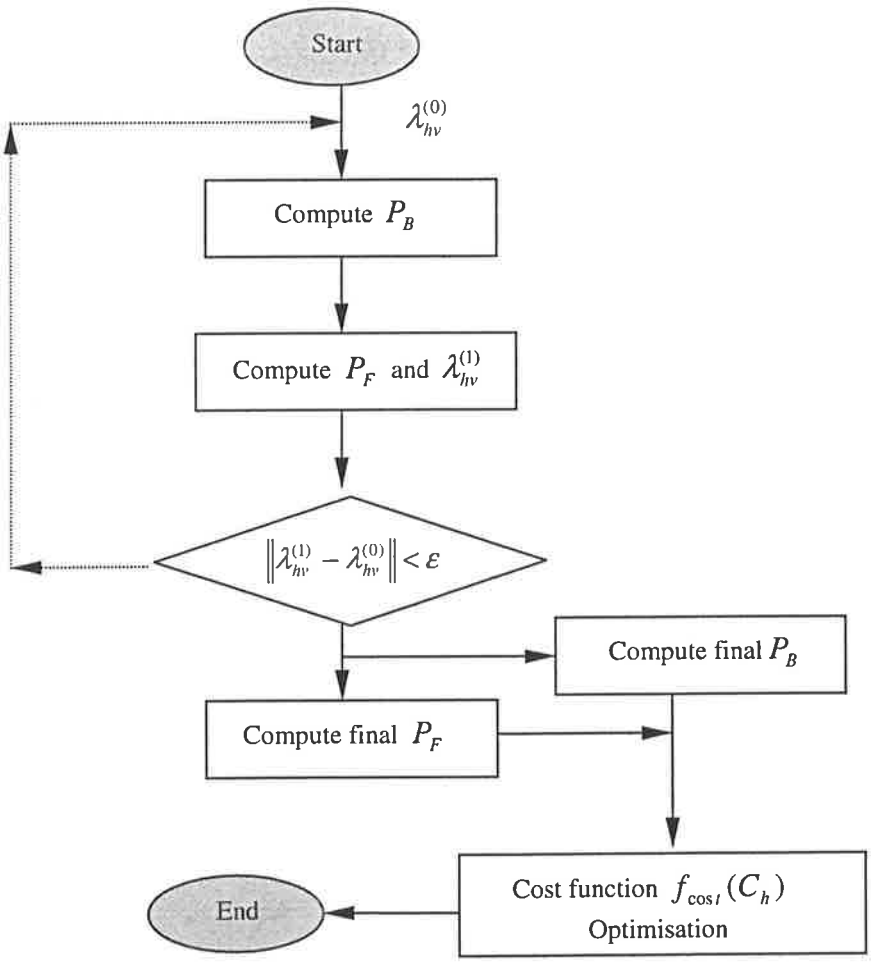


Figure 3. 4: Handoff rate determination flow chart

In the end, the mean waiting time in queue can be derived as:

$$E(T_Q) = \frac{E(N)}{\lambda_{hv}(1-P_F)} = \frac{\sum_{j=0}^{\infty} jP_{C+j}}{\lambda_{hv}(1-P_F)} \tag{3.38}$$

Generally, a backlog function is used to represent waiting time distribution [*Klein75*]. The backlog function can be used to represent the interval time to completely empty all users who arrive prior to a test time. It is easy to find that this actually forms a continuous state Markov process with some discrete jumps. Although the busy period distribution is independent of queuing disciplines, the distribution of waiting time indeed relies upon the rules of service order. Let $f(w)$ denote the stationary density function of the virtual waiting time in the rule of FIFO, we can have the normalised relationship for waiting time distribution as:

$$f(w=0) + \int_{0^+}^{\infty} f(w)dt = 1 \tag{3.39}$$

For the FIFO discipline, the backlog function is equivalent to the virtual waiting time of users. However, under the rule of LIFO, this is not the case.

Comparing the delayed model with the reneging model, we observe that the equilibrium probability in the delayed model is always no more than that in the reneging model, that is, $P_{de}(c,a) \le P_{reg}(c,a)$. In other words, the penalty paid for the use of the reneging model is an increase of blocking probability.

3.1.2.3) A Handoff Fixed Timeout Model

The aforementioned reneging model belongs to the cases of random variable timeout scheme [*Lin94*]. The maximum allowable waiting time of an arrival call must be larger than the timeout random variable if the call is admitted to servers. Otherwise, the arrival call becomes lost. On the other hand, another type of timeout controls is called constant timeout. For the handoff request queued scheme mentioned earlier, the fixed timeout scheme can be used to discard the excessive long delayed customers.

For a system with random variable timeout, the basic principle of a Markov process can still be applied. However, for a system with fixed constant timeout, a Markov chain is no

longer applicable [*Gned68*]. The state of the system at time $(t+\Delta t)$ depends not only upon the number of calls j in system at time t, but also the waiting time of the calls prior to t. For instance, one cycle of a regenerative process is shown as in [Figure 3.5].

Here we consider an unlimited queuing system with C servers queuing system following the FIFO discipline. Let $\xi_j(t)$ represent the virtual waiting time reaching jth server while the arrival is at time t. Therefore we can form a virtual waiting time vector $\xi(t) = \{\xi_1(t), \xi_2(t), \ldots, \xi_j(t), \ldots, \xi_C(t)\}$. From this diagram [Figure 3.5], we find that the arrivals will become lost if the instant time $\xi_j(t)$ of backlog function is bigger than the limited constant time τ. Therefore the virtual waiting time $\xi(t)$ can actually construct a Markov process because the state of $\xi(t)$ at $(t+\Delta t)$ only depends on the state of $\xi(t)$ at time t. Here we are going to derive the forced termination probability under the assumption of constant timeout constraint.

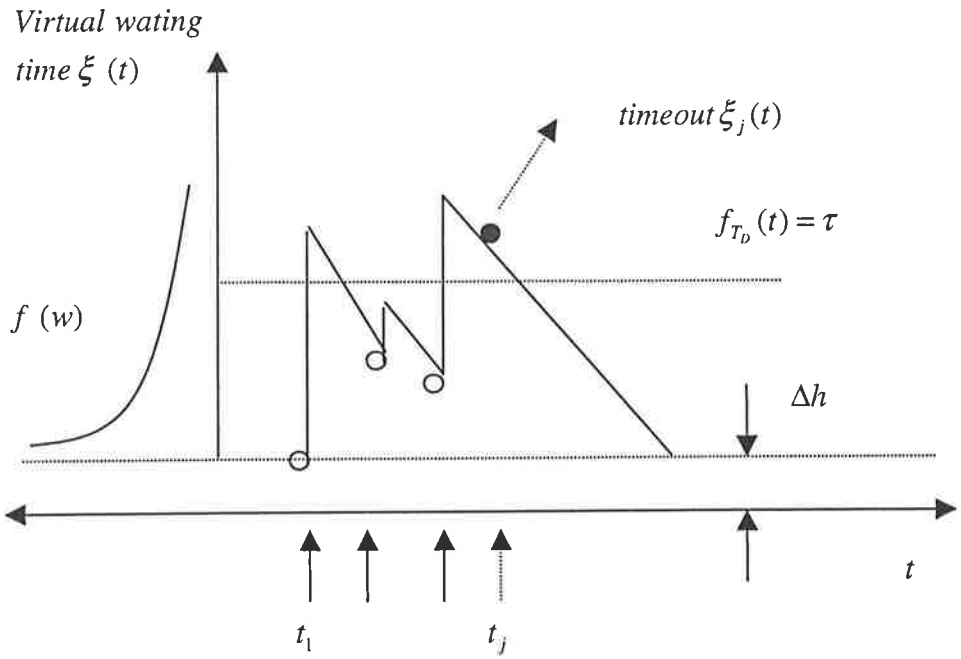


Figure 3. 5: A fixed time-out system

Firstly, we assume that new call and handoff call arrival rates are λ_v and λ_{hv} respectively. These calls can gain access to all servers C. Once the arrivals find all servers

full, only the handoff arrivals are allowed to wait in the queue. Meanwhile, the CHT is equivalent to $\mu_{CH} = \mu_v + \eta$.

A test customer is assumed to arrive with a virtual waiting time $\xi_j(t)$ reaching the *jth* server at time t. Let X_i represent the time periods when j servers are busy until all the calls before the test customer finishes. If we denote $\xi_1(t) < x_1$, $\xi_2(t) < x_2, \cdots, \xi_j(t) < x_j$, then $p(x_1, x_2, \cdots, x_C)$ is used to represent the pdf of the virtual waiting time $\xi(t)$ while all C servers are busy. Following the pdf for virtual waiting time in [*Gned68*], we can modify it as:

$$p_C(x_1, x_2, \cdots, x_C) = p_0 \frac{(\lambda_v + \lambda_{hv})^C}{C!} e^{-\mu_{CH}(x_1+x_2+\cdots+x_C)+\min(\lambda_{hv}\tau+\lambda_{hv}x_1+\lambda_{hv}x_2+\cdots+\lambda_{hv}x_C)} \tag{3.40}$$

The blocking probability P_B for constant timeout is when the minimum of the all virtual waiting time exceeds τ:

$$P_B = P(\min(\xi_1, \xi_2, \cdots, \xi_C) > \tau) = P(\xi_1 > \tau, \xi_2 > \tau, \cdots, \xi_C > \tau) \tag{3.41}$$

That is,

$$P_B = \int_{\tau}^{\infty}\int_{\tau}^{\infty}\cdots\int_{\tau}^{\infty} p_C(x_1, x_2, \cdots x_C)dx_1dx_2\cdots dx_C \tag{3.42}$$

Then, we have:

$$P_B = P_0 \frac{1}{C!}\left(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}}\right)^C e^{-(C\mu_{CH}-\lambda_{hv})\tau} \tag{3.43}$$

where the normalised constant P_0 is:

$$P_0^{-1} = \sum_{j=0}^{C-1}\frac{1}{j!}\left(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}}\right)^j + \frac{1}{C!}\left(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}}\right)^C \frac{\lambda_{hv}e^{-(C\mu_{CH}-\lambda_{hv})\tau} - C\mu_{CH}}{\lambda_{hv} - C\mu_{CH}} \quad while \quad \lambda_{hv} \neq C\mu_{CH}$$

$$P_0^{-1} = \sum_{j=0}^{C-1}\frac{1}{j!}\left(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}}\right)^j + \frac{1}{C!}\left(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}}\right)^C (1+\lambda_{hv}\tau) \quad while \quad \lambda_{hv} = C\mu_{CH} \tag{3.44}$$

For the probability of an arrival who needs no waiting, we can have:

$$P(W_Q = 0) = \sum_{j=0}^{C-1} P_0 \frac{1}{j!}(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}})^j \tag{3.45}$$

In addition, we already know:

$$P(W_Q \le t) = P(W_Q = 0) + F(t \mid W_Q > 0)P(W_Q > 0) \tag{3.46}$$

That is,

$$P(W_Q \le \tau) = P(W_Q = 0) + P\{\min(\xi_1, \xi_2, \cdots \xi_C) < \tau\} \tag{3.47}$$

Therefore the forced termination probability P_F can be shown as:

$$P_F = 1 - \{P(W_Q = 0) + P(\min(\xi_1, \xi_2, \ldots, \xi_n) < \tau\} \tag{3.48}$$

In the end, we can have the final result as:

$$P_F = 1 - \sum_{j=0}^{C-1} P_0 \frac{1}{j!}(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}})^j - \frac{C}{C!}(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}})^C \frac{(1 - e^{-(C\mu_{CH} - \lambda_{hv})\tau})\mu_{CH}}{C\mu_{CH} - \lambda_{hv}} P_0 \quad while \quad \lambda_{hv} \neq C\mu_{CH}$$

$$P_F = 1 - \sum_{j=0}^{C-1} P_0 \frac{1}{j!}(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}})^j - \frac{C}{C!}(\frac{\lambda_v + \lambda_{hv}}{\mu_{CH}})^C \mu_{CH} \tau P_0 \quad while \quad \lambda_{hv} = C\mu_{CH} \tag{3.49}$$

These formulas can be used to study the performance of the timeout model with constant time constraint.

As a summary for Subsection §3.1.2, we find that different handoff control schemes have different contributions on call performance. The choice of control scheme depends on specific traffic handling preference and requirement of QoS.

3.1.3 Performance of Multiservice in Loss Systems

As mentioned early, in a wireless environment, traffic variability and handoff traffic must be taken into account. Therefore, the conventional multidimensional birth and death

process and approximation techniques derived in wireline networks need to be reconsidered.

Using a FCA, the product form for equilibrium probability holds for a voice-only multicell system once the state space is truncated in a loss system [*Everitt89a*]. Here we present a traffic model with explicit mobility modelling based upon the extension from the result in [*Everitt94*]. This result is not only applicable to negative exponential channel holding time, but also to a generalised service time distribution.

For example, there are a total of $m\ (1 \le i \le m)$ cells with n_i calls in cell i. The overall system has available channels C. The offered load in cell i can be expressed by $\rho_i = (\lambda_{ni} + \lambda_{hi})/(\mu_{ni} + \eta_i)$, where η_i represents the rate of CST in cell i. Then we can have:

$$\pi_{\infty}^{(\bar{n})} = \pi_{\infty}^{(\bar{0})} \prod_{i=1}^{m} \left(\frac{\lambda_{ni} + \lambda_{hi}}{\mu_{ni} + \eta_i} \right)^{n_i} \frac{1}{n_i!} \quad , \quad A\bar{n} \le \bar{c} \tag{3.50}$$

where A is an appropriate matrix corresponding to each cell, that is, $\bar{n} = (n_1, n_2, \cdots, n_i, \cdots, n_m)'$ and $\bar{c}$ is a vector with channel numbers.

For a multiservice circuit-switched loss system, we can map it into a tree-type topology according to the fixed routing assumption [*Kelly91*]. For instance, a class j can be viewed as voice or data service. A *rth* type service of class j is equivalent to different kinds of call services in that class. Class j requires C_j channel units for transmissions. Then this can be easily shown as in [Figure 3.6].

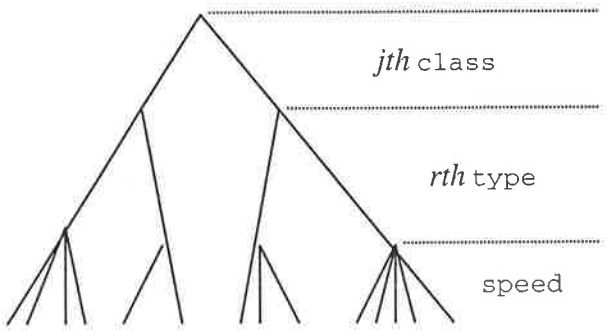


Figure 3. 6: A tree type structure for a loss cellular network

As a result, the structure gives rise to a product form solution for equilibrium probabilities. The blocking probability for *jth* class is the summation from the blocking states in this *jth* class. If we assume that class j totally requires C_j resource $(1 \le j \le J)$, which can be expressed by a vector $C = (C_1, C_2, \ldots, C_j, \cdots, C_J)$. The active number of *rth* type service of Class j can be denoted by $n_r\ (1 \le r \le R)$. Thus the vector, $n = (n_1, n_2, \ldots, n_r, \cdots, n_R)$, where $r \in \Re$, shows the number of active calls in the system, The arrival process is assumed to follow a Poisson process with rate λ_r and the service process corresponds to identically distributed CHT with mean rate μ_r. The *rth* type service of j class is allowed to use A_{jr} $(A_{jr} \in Z_+)$ channels.

Then the equilibrium probability distribution $\pi_\infty^{(n)}$ can be denoted by a R-dimensionally truncated birth and death process as:

$$\pi_\infty^{(n)} = \frac{1}{G(C)} \prod_{r=1}^{R} \left(\frac{\lambda_r}{\mu_r} \right)^{n_r} \frac{1}{n_r!}, \quad n \in S(C) \tag{3.51}$$

The state constraint is given by:

$$S(C) = \left\{ n \in Z_+^{\Re} : An \le C \right\} = \left\{ n : \sum_r A_{jr} n_r \le C_j \right\} \tag{3.52}$$

The normalised constant $G(C)$ is:

$$G(C) = \sum_{An \le C} \prod_{r=1}^{R} \left(\frac{\lambda_r}{\mu_r} \right)^{n_r} \frac{1}{n_r!}, \quad n \in S(C) \tag{3.53}$$

If we use $e_r = (0, 0, \ldots, 1, 0, \ldots, 0)^T$ to express the *rth* type of calls in the system, we can use this formula to calculate the blocking probability P_{B_r} for the *rth* type service:

$$P_{B_r} = 1 - \frac{G(C - Ae_r, R)}{G(C, R)} \tag{3.54}$$

Although this has a neat and simple look, it is not trivial to calculate the normalised constant $G(C)$ for the large types of class services, which require a large number of

channel units. In order to obtain a solution, the convolution algorithm, the fast Fourier transform-based algorithm, and the generalised Kaufman/Roberts algorithm can be employed, which can be found in [*Tsang90*]. In order to simplify the computation, the other approximation methods, such as reduced load approximation, knapsack approximation and Pascal approximation can be adopted [*Chung93*]. As an example, for the reduced load approximation, the basic idea is to treat blocking occurring independently from link to link and therefore the offered load in a link is reduced according to the blocking on other links.

The equilibrium probability distribution of a two-class circuit access loss system can be shown as in [Figure 3.7].

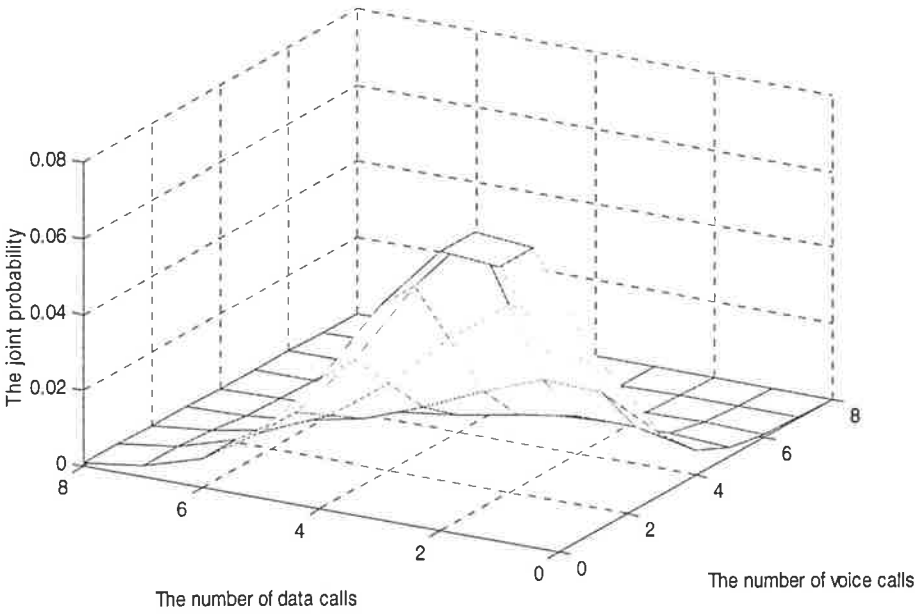


Figure 3. 7: The probability distribution for a two-class service

In this study, new voice call arrival rate is equal to 0.0333 calls/s, while handoff calls arrive at a rate of 0.0083 calls/s. Voice services are assumed to have a speed of 30 km/hour, while data terminals are stationary and have an arrival rate of 0.0167 calls/s. Both data and voice services are assumed to have a mean duration of two minutes.

From this figure, we observe that the distributions for both types of traffic are asymmetric and dependent upon both offered traffic loads and terminal mobility.

3.1.4 Summary

In Section §3.1.1, we study the effect of traffic variability and handoff priority schemes on call performance. An important observation from this study is that, although call blocking probability is insensitive to the distribution of CST, the probability of call forced termination can be significantly affected by the variance of CST. Taking cell size and terminal speed into account, we observe that call incomplete probability increases according to the increase of terminal movement speed. Note that the actual CST data must be collected from field trials and then fitted into each type of distribution, but this is beyond the scope of this study and needs to be further investigated.

In Section §3.1.2, we have studied the impact of different handoff control schemes in various traffic models on call performance, such as the handoff delayed model, the reneging model and the timeout model. We show that fast mobility has a significant effect on call waiting time distribution in the delayed model. The high handoff loads are likely to cause long delays. Smaller cell size and faster movement lead to handoff calls having less chance of waiting. In addition, the forced termination probability is derived for the constant timeout system.

Finally, in Section §3.1.3, call performance for multiservice in loss wireless systems is investigated. From this study, we observe that the call performance in loss wireless systems is tractable and can be solved by the specific methods as aforementioned. In contrast, the performance of multiservice in queuing system is more challenging and thus it is investigated from the next section.

In conclusion, we observe that call performance is subject to the use of different call control schemes. Integrated radio resource management can become more effective once the individual tasks, like admission control, handoff control, and mobility control, can be taken into overall consideration. Better overall performance can be obtained through such optimal procedures and their trade-offs.

3.2 An Approximate Method for Integrated LRD Services

As mentioned in the last chapter, there are many approximation methods that can be used to study the behaviour of integrated services. Because data services have a wide range of applications, different traffic source models have different statistical characteristics. A tractable approximation approach is developed to analyse the integration of voice calls and low rate data (LRD) packets in a prioritised wireless system. The effect of priority handoff schemes on packet delay is investigated. It shows that the channel utilisation in mixed services is subject to the use of priority schemes as well as traffic variability. More importantly, packet performance depends on packet message lengths, offered loads and channel conditions. The following section is organised as follows. Introduction begins in Section §3.2.1. Three kinds of source traffic models are depicted in Section §3.2.2. An approximation approach is proposed in Section §3.2.3. Subsequently, results are discussed in Section §3.2.4. Finally, we conclude with a summary.

3.2.1 Introduction

Global Systems for Mobile (GSM) is a digital cellular standard developed by the European Telecommunications Standards Institute (ETSI) that is now adopted worldwide. At present, GSM supports circuit data services like short message services. In order to support bursty data services and also minimise the change of existing infrastructure, it is proposed that General Packet Radio Service (GRPS) will be used to deliver the connectionless mode data services over connection mode voice services [*ETSI360*].

From a network point of view, integrated services must be achieved by the efficient sharing of spectrum. In the design of integrated systems, one of the fundamental problems is how to satisfy the different QoS requirements for a variety of traffic classes while maintaining the efficient use of resources.

From the result shown in [Figure 3.8], we observe that the result of a more stringent QoS requirement for single voice service is lower channel utilisation. Channel utilisation will not achieve the maximum value of unity even if the available channels are large enough. For instance, although channel utilisation shows improvement according to the increased channels with the constraint of 2% blocking probability, the utilisation only reaches about 87% when the total channels are equal to 100.

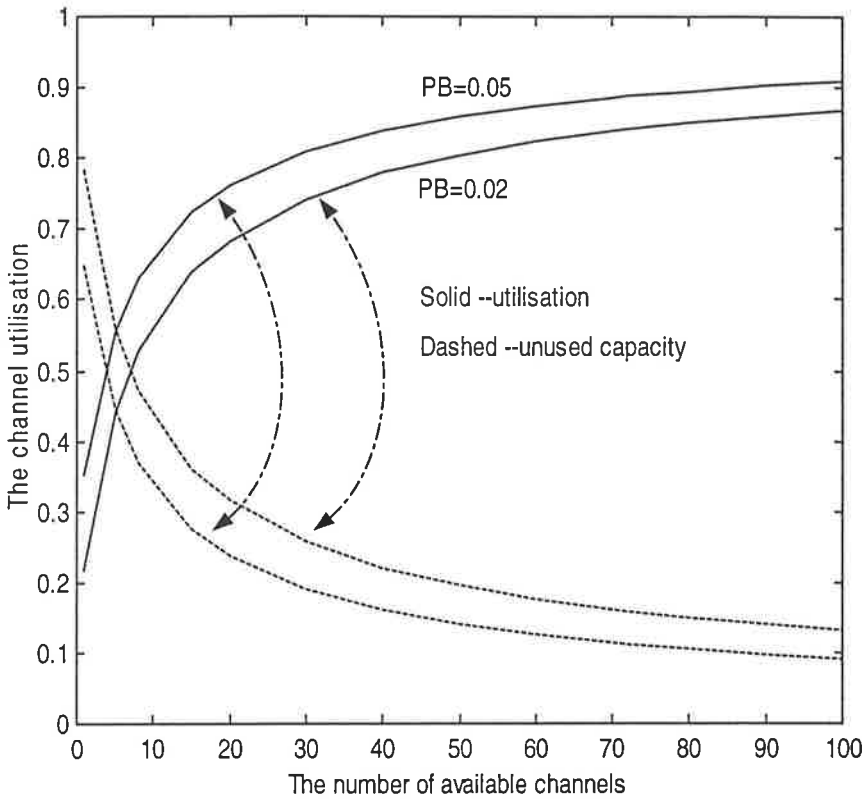


Figure 3. 8: Channel utilisation for voice calls

As a result, the basic idea of GPRS is to exploit the leftover capacity by voice calls for the delivery of data operation, e.g., a low-speed data rate of 9.6 kb/s in the present GSM systems. In order to share physical channels dynamically, the CoD concept is introduced and network operators are able to decide the number of Packet Data Traffic Channels (PDCHs) according to the actual GPRS traffic demands.

Briefly, the main elements in GPRS consist of gateway GPRS support node (GGSN) and serving GPRS support node (SGSN). SGSN is used to transmit data packets to the BSC, whereas GGSN is used to extend the data packets to the outside data networks. Analogous to the terms developed for circuit mode in GSM, there are two kinds of logical channels in GPRS. One is called packet control channels (CCHs) and the other is packet traffic channels (TCHs). From the purpose of traffic study, TCHs are the focus in this investigation.

During the operation, a slotted ALOHA reservation scheme is used in GPRS packet transmissions, which is similar to PRMA systems. Terminals initiate a data packet by

lodging a random access request in one of the bursts called packet random access channel (PRACH) on the uplink. Note that there is either one phase or two-phase access that can be adopted [*ETSI364*]. In the one phase channel access, a terminal initially lodges a request to the BS on the PRACH. Subsequently, the network will initiate an immediate assignment and then channel resources are assigned. In contrast, two-phase access means that a terminal has to lodge two requests separately, one for packet channel request and the other for packet resource request. The random access subchannel is then determined by the uplink state flags (USFs). More specifically, if USFs are equal to 'FREE', the mobile terminal can send a random request to the PRACH. Otherwise, as a response to random access, USRs are set to 'R' for channel reservation [*Turina96*].

The investigation of packet performance in GPRS can be found in [Brasch97] and [Cai97]. In addition, Sarker [*Sarker97*] proposes to use a Consistent–TDMA (CTDMA) system, in which the number of data terminals are restricted to control the problem of stability. However, neither mobility nor prioritised handoffs, which is important to a wireless network, has been included in these studies.

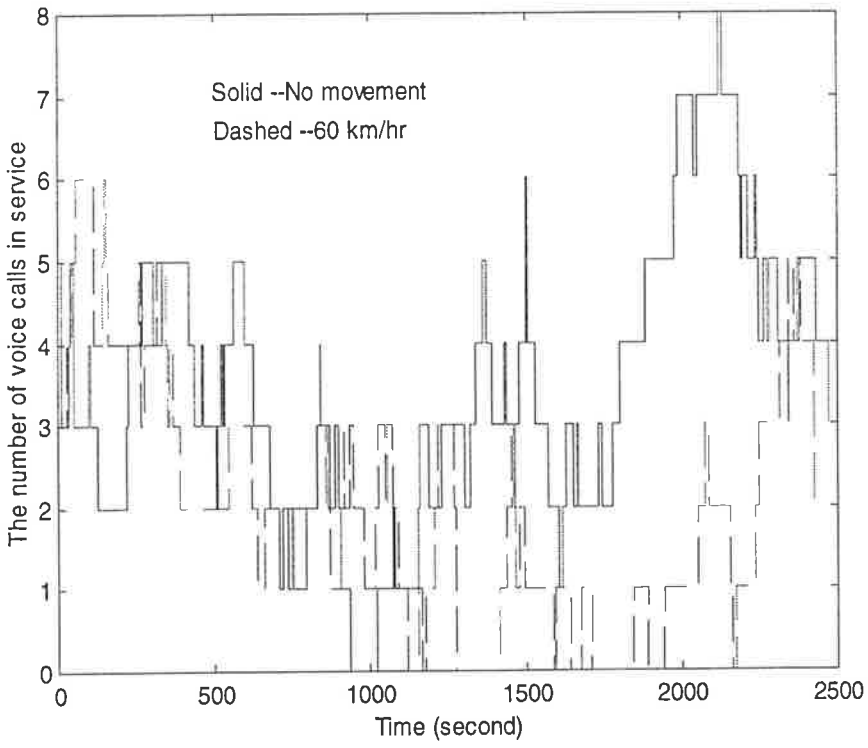


Figure 3. 9: Simulation result for a voice system

From the simulation study [Figure 3.9], in which the total channels are assumed to be equal to $C = 8$ and the required blocking probability is 2%, it is shown that channels can not be used all the time. For example, we observe that there is a big difference for channel utilisation between 1500 seconds and 2500 seconds. As a result, terminal movement has a significant effect on channel utilisation according to the time when movement increases. In other words, apart from satisfying the required stringent QoS, we have to take terminal movement into account during traffic study.

With the control access for data terminals, both the prioritised handoff scheme and terminal movement are taken into account in this study. For simplicity, the approximation method is used for evaluating the quantitative performance. Beside single slot operation, the other important characteristic of GPRS is the employment of multislot services for high-rate data packet transmission. However, in this section, analysis is only focused on low rate data operation. In Chapter 4, the investigation of multiple data slots is followed.

Although there exist different standards for the present systems, the GSM-like environment is used as an example for analysis in this study. However, it is worthwhile to mention that this analytical approach is generic and can be easily applied to other TDMA systems as well. Because traffic characteristic affects traffic performance, diverse traffic source models need to be introduced as follows.

3.2.2 Multimedia Traffic Source Models

As a matter of fact, different data applications have diverse traffic generators. The assumption of using source models directly affects queuing performance. Some short-range dependent Markovian and autoregressive (AR) models are introduced as follows. In addition, traffic models, which are not analytically tractable, can only be used to generate traffic traces by simulation studies [*Adas97*].

- **Voice Traffic Source Model**

First, speech source traffic can be characterised as an on-off model [*Brady69*]. Both talking periods and silent periods are generated by exponential distributions. If the talking active voice numbers are assumed to be large enough, such as $(n \geq 25)$, a simple birth and death process can be used to approximate active states [*Weinst78*].

In this study, voice calls are assumed to follow a Poisson process with mean rate λ_v. Meanwhile, voice handoff calls are assumed to arrive at rate λ_{hv}. The service rate of handoff calls is equivalent to that of new calls: $\mu_{hCH} = \mu_{vCH}$. The on and off model is adopted for the studies of R-TDMA systems in Chapter 4, DS-CDMA and HCS systems in Chapter 5 respectively.

- **Diverse Data Traffic Source Models**

Anagnostou [*Anag96*] uses a multilevel model for bursty data source traffic, which include connection level, action level and transmission level [Figure 3.10]. Although this model can better capture the bursty behaviour of data packets, it will drastically increase the complexity of evaluation due to multi-state, especially for the integration with voice service.

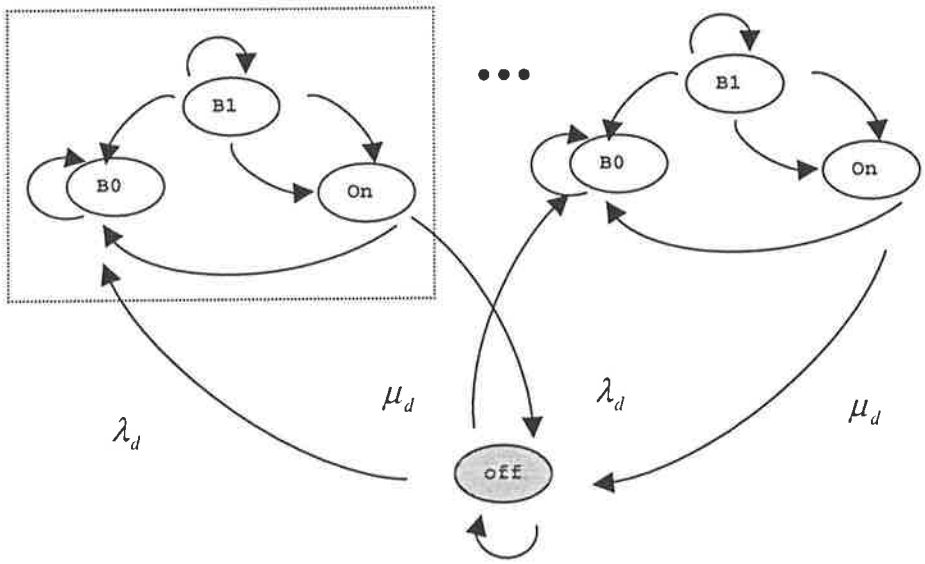


Figure 3. 10: A multilevel data source traffic model

The specific data source models depend on the types of data applications. For instance, a truncated Cauchy distribution can be used for the e-mail traffic, with a maximum message size of 10 kbytes [*Cai97*] and [*Brasch97*]. A simulation study of 100,000 samples is fitted into the statistical result shown as in [Figure 3.11]. In this study, we observe that the simulation result is consistent with analytical result.

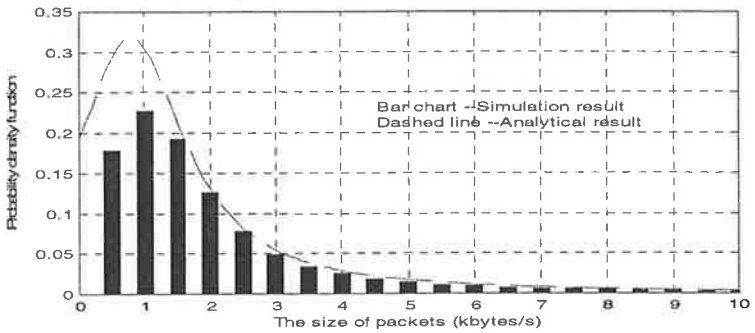


Figure 3. 11: A simulation of the e-mail application

Another important example is a World Wide Web (WWW) source traffic model, which is shown in [Figure 3.12].

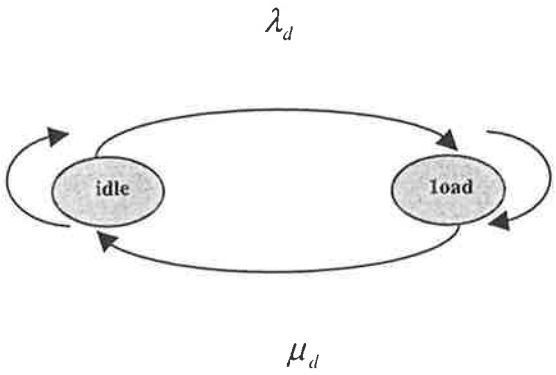


Figure 3. 12: A WWW traffic model

In this model, the idle time between two successive documents is assumed to be an exponential distribution. The number of WWW objects per document follows a geometrically distributed, whereas the size of a WWW object follows a Pareto distribution. The WWW traffic model is regarded as a heavy tailed distribution†. The actual parameters of the different distributions must be obtained from the traffic measurement [*Hoff98*].

The simulation study and the theoretical result can be shown as in [Figure 3.13]. From the figure, we observe that both results are in a good agreement. The probability of packet

† A probability distribution $P(y)$ is heavy tailed if it satisfies $P(Y > y) \approx L(y)y^{-l}$ for $y \to \infty$ and $l > 0$ with $L(y)$ being a slowly varying function.

sizes decrease according to the increase of packet sizes. Meanwhile, it is worthwhile to mention that the uplink and downlink traffic in WWW services is asymmetric.

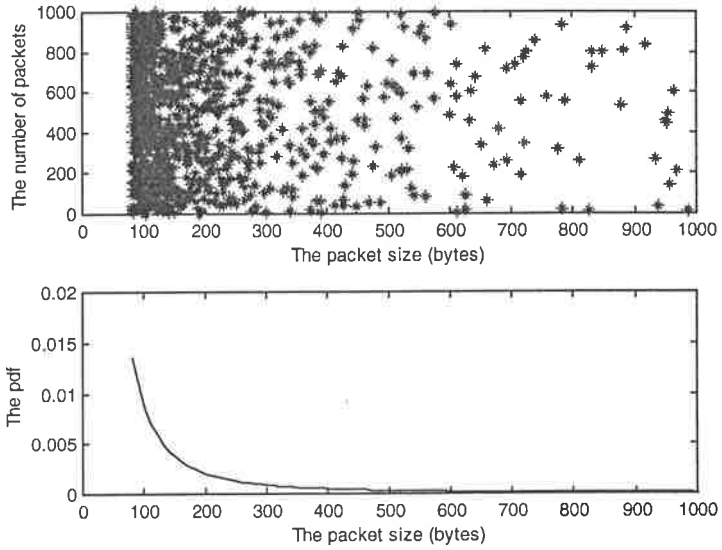


Figure 3. 13: The packet size for a WWW model

- **Video Traffic Source Model**

Video traffic can occupy a multiple of traffic channels in time multiple access systems. The characteristics of video traffic vary greatly depending on the different applications as well as coding schemes. It is known that there are two kinds of coding schemes used in video service. One is CBR coding and the other VBR. For the CBR coding scheme, a buffer can be used to smooth the bit rate variability in the video encoder output. On the other hand, because the VBR coding scheme needs to deliver the maximum bit rate, buffers can be installed in both the encoder and decoder ends if the maximum channel bit rate exceeds the maximum encoder output [*Bout97*] and [*Boutyu97*]. In particular, a prime characteristic of VBR coding is that the change of scene can cause a large abrupt change in bit rate.

In order to obtain the autocorrelation between frame sequences, the AR(p) model of order p can be used to represent their relationships [*Nomura89*]:

$$X(n) = \sum_{i=1}^{p} \phi_i X(n-i) + N_0(n) \tag{3.55}$$

where $X(n)$ represent the bit rate in n^{th} frame, $N_0(n)$ is the Gaussian white noise, and ϕ_i is the coefficient with real number as shown in [Figure 3.14].

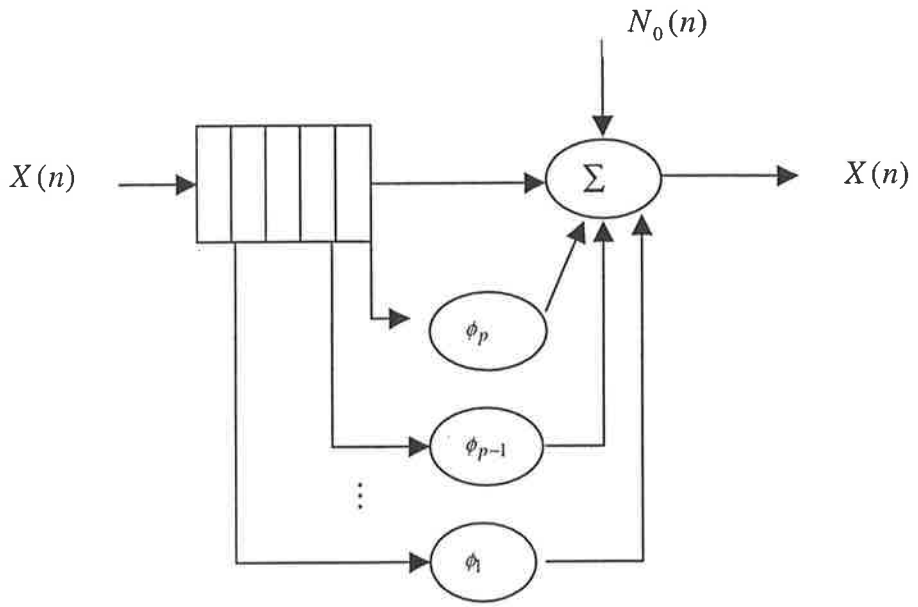


Figure 3. 14: A multilevel data source traffic model

Thus a video source can be approximated by a continuous fluid flow model, where the output bit rate within a frame period can be represented by the first order AR (1) model:

$$I(n) = \phi_1 X(n-1) + N_0(n) + I_0 \tag{3.56}$$

where I_0 is average frame size.

A sample of video traffic can be shown in [Figure 3.15]. From the figure, we observe that successive frames within a video scene are strongly correlative and the change of variance leads to the change of frame bit rate.

As a summary, from the studies we have showed previously, we observe that different types of data applications have different source models. The queuing performance for multimedia traffic is subject to using the type of source traffic as well as buffering. In order to obtain a tractable solution in our sequent analysis in the following, we adopt the simple assumption of exponential distribution as the arrival and service processes.

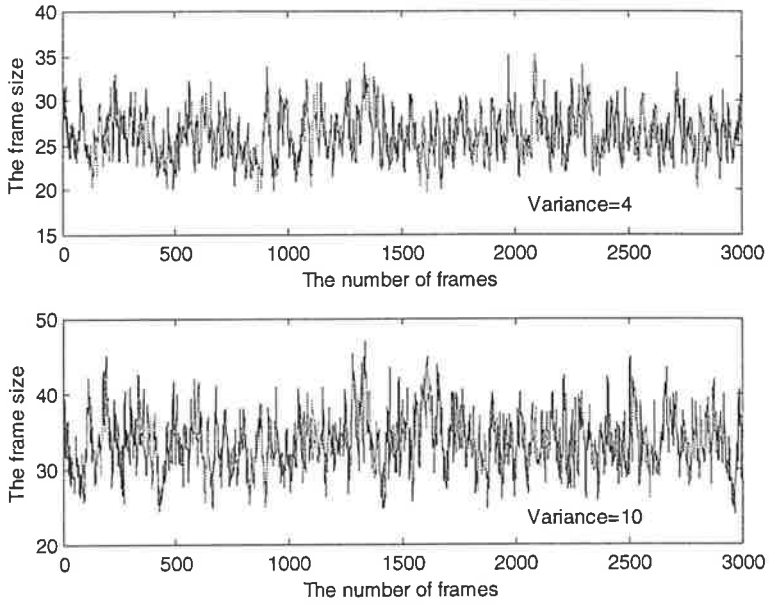


Figure 3. 15: A sample of video source traffic

In this simulation, we assume that Gaussian distribution is with mean $m = 0$. *The average frame size is* $I_0 = 12.8$ *and the coefficient equal to* $\phi_1 = 0.88$.

3.2.3 Performance of LRD Integrated Services

As mentioned in Chapter 2, due to terminal movement, a service breakdown for handoff calls can occur if there are no idle channels in the next cell. Based upon the assumption of the source model as previously stated, for simplicity, data packets are assumed to follow Poisson arrivals with rate λ_d and handoff calls with rate λ_{hd} respectively.

Under call admission control (CAC) policy [*Calleg95*], integrated services can operate under different permitted regions so that different requirements of QoS can be satisfied. In order to guarantee the QoS, an efficient call handling scheme becomes essential. This is true not only for a single service, but also for integrated services. The channel access and queuing discipline are organised as follows:

1) Certain guard channels are exclusively retained for handoff voice calls and data packets only. The CHT of voice calls is $\mu_{vCH} = \mu_v + \eta$ and the CHT of data calls is $\mu_{dCH} = \mu_d + \gamma$.

2) Whenever channels, except for the reserved channels, are available, voice calls or data packets are served by the order of FIFO basis. However, voice calls have priority over data packets and have a privilege to pre-empt data packets while data packets are being served and there are no available channels.

3) In order to maintain the stability of operation, only N data terminals are permitted to gain access to the unoccupied channels provided all traffic channels are not occupied by voice terminals. One time slot per terminal per frame is also assumed.

Data users in service are represented by i and voice users in service by j respectively. The queuing model can be depicted as shown in [Figure 3.16].

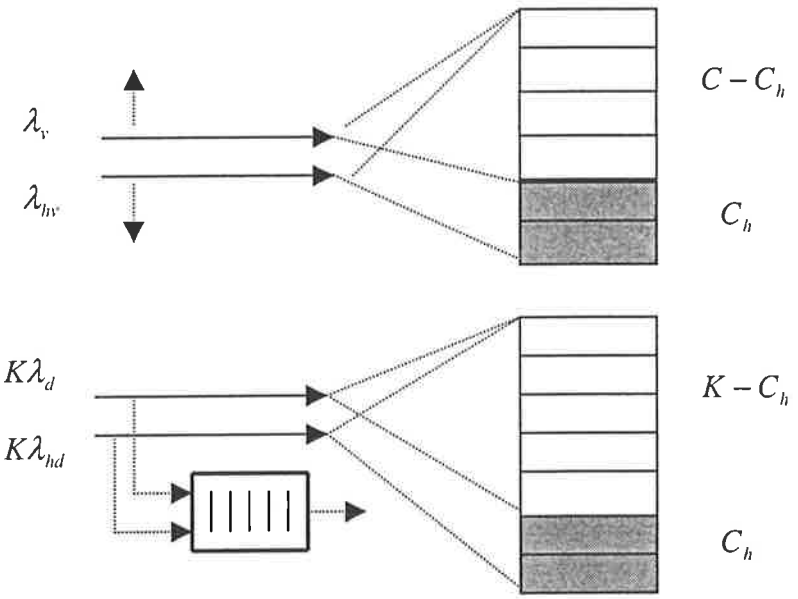


Figure 3. 16: The integrated model

In order to exploit channel utilisation, data packets can occupy all the slots $K = \overline{N} - j$, which are left over by voice calls. Because $\overline{N}$ is the average number of slots in a frame and j is the number of slots occupied by voice calls, K is a random variable (r.v.).

Therefore, the available slots for data packets vary according to the number of active voice users. The analysis is then approximated by marginal distributions.

As a solution for voice calls, the equilibrium probability is easily expressed by:

$$P_j = P_0 \frac{(\lambda_v + \lambda_{hv})^j}{j!\mu_{vCH}^j} \quad while \quad 0 \le j \le (C - C_h) \tag{3.57}$$

$$P_j = P_0 \frac{(\lambda_v + \lambda_{hv})^{C-C_h} \lambda_{hv}^{j-(C-C_h)}}{j!\mu_{vCH}^j} \quad while \quad (C - C_h + 1) \le j \le C \tag{3.58}$$

Then the originating voice call blocking probability can be determined by:

$$P_{Bv} = \sum_{j=C-C_h}^{C} P_j \tag{3.59}$$

Consequently, the probability of attempt failure for handoff calls becomes: $P_{hfv} = P_C$.

For the analysis of low rate data, it is similar to the results derived for voice calls except for the consideration of the queuing effect. The important performance measures, like equilibrium distribution, the mean number of data packets, and throughput, can be derived respectively [*Yu97*].

The equilibrium distribution of data packets in service is:

$$P_i = P_0 \binom{K}{i} \frac{(\lambda_d + \lambda_{hd})^i}{\mu_{dCH}^i} \quad while \quad 0 \le i \le (K - C_h) \tag{3.60}$$

$$P_i = P_0 \binom{K}{i} \frac{(\lambda_d + \lambda_{hd})^{K-C_h} \lambda_{hd}^{i-(K-C_h)}}{\mu_{dCH}^i} \quad while \quad (K - C_h + 1) \le i \le K \tag{3.61}$$

If the total available queue length can be expressed by $(N-K)$, the data packets in queue can be shown to be:

$$P_i = P_0 \binom{N-K}{N-i} (i-K)! \frac{(\lambda_d + \lambda_{hd})^{i-C_h} \lambda_{hd}^{C_h}}{\mu_{dCH}^K \prod_{l=1}^{i-K} (K\mu_{dCH} + l\mu_Q)} \quad while \quad (K+1) \le i \le N \tag{3.62}$$

...be derived as below:

$$\frac{(\lambda_d+\lambda_{hd})^{K-C_h}\lambda_{hd}^{i-(K-C_h)}}{\mu_{dCH}^{i}}$$

$$\left[\sum_{i=K+1}^{N}\binom{N-K}{N-i}(i-K)!\frac{(\lambda_d+\lambda_{hd})^{i-C_h}\lambda_{hd}^{C_h}}{\mu_{dCH}^{K}\prod_{l=1}^{i-K}(K\mu_{dCH}+l\mu_Q)}\right]^{-1} \tag{3.63}$$

Subsequently, the total mean number of time slots during channels, including reserved channels, can be calculated by:

$$E_D=\sum_{i=0}^{K-C_h}iP_i+\sum_{i=K-C_h+1}^{K}iP_i \quad while \quad 0\le i\le K \tag{3.64}$$

Hence the throughput of data packets S_D during random access becomes:

$$S_D=\frac{1}{K}\left(\sum_{i=0}^{K-C_h}P_0\binom{K}{i}\frac{(\lambda_d+\lambda_{hd})^i}{\mu_{dCH}^i}+\sum_{i=K-C_h+1}^{K}iP_i\right) \quad while \quad 0\le i\le K \tag{3.65}$$

The marginal average number of delayed data packets $\overline{D_{D,K}}$, which refers to the data packets in the queue becomes:

$$\overline{D_{D,K}}=\sum_{i=K+1}^{N}(i-K)P_i \tag{3.66}$$

After taking the voice traffic into consideration, the joint average number of data packets $\overline{D_{V,D}}$ can be shown as:

$$\overline{D_{V,D}}=\sum_{j=0}^{C}P_j\overline{D_{D,(\overline{N}-j)}} \tag{3.67}$$

Finally, according to Little's law, the marginal average waiting time in the queue is:

$$\overline{W}=\frac{\overline{D_{D,K}}}{\lambda_d+\lambda_{hd}}=\frac{1}{\lambda_d+\lambda_{hd}}\left(\sum_{i=K+1}^{N}(i-K)P_i\right) \tag{3.68}$$

3.2.4 Results Discussion

In this study, we assume that there are 1, 2, and 3 Transmitters/Receivers (TRXs) in systems with two reserved channels respectively. The total number of data packets is 40. Without considering the dynamic load of voice calls, we assume that the maximum leftover channels for packet transmission of the 3 TRXs equal to 8, 16, and 24. Then we have the result as shown in [Figure 3.17], in which channel transmission rate is 9.6 kb/s. Handoff data arrivals are equivalent to 10% of new data arrivals. Meanwhile, 1.5 kbytes is assumed to be the average message length. As a comparison, the solid lines show the results for the case of 2 kbytes message length.

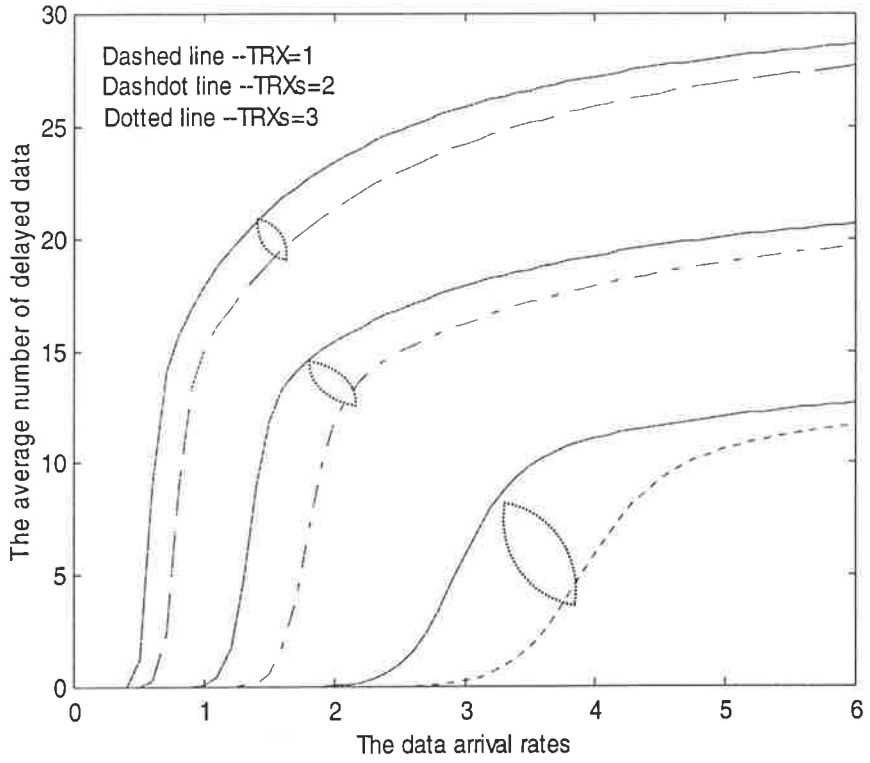


Figure 3. 17: The delay of data packets

From the figure 3.17, firstly, we observe that the control of data arrivals effectively restricts the unbounded data delay. Under a certain average delay number constraint, packet arrival loads must increase if data message length becomes shorter. Secondly, the use of a prioritised handoff scheme has little effect on the average number of data packets

while the arrival rates stay at both the low end and high end of arrivals. The conclusion for this is that, if the arrival rates are low, there is no need to adopt the priority scheme. On the other hand, if the arrival rates become high, the use of the control scheme becomes invalid due to overload. Finally, packet performance is subject to the addition of the number of TRXs and the average number of delayed data packets largely depends on the statistics of voice calls.

3.2.5 Summary

For Chapter 3, the key contribution lies in Section §3.2. In this section, various multimedia source traffic models are investigated. The performance of integrated services is subject to the use of traffic source models and queue length for specific data applications. For the integration of low rate data packets into voice services, we propose and develop a simple approximation method in a prioritised integrated environment. From the analysis, we observe that firstly, channel utilisation can be improved by the use of integrated packet-based data services into voice services. More importantly, channel performance for data packets is subject to the use of control schemes. Secondly, packet performance is determined not only by the average length of message, the offered loads, but also by channel conditions, which are determined by the volatility of voice traffic in the system. Finally, the drawback for the approximation method is that it is applied to a light traffic load condition only. Therefore, a more efficient and accurate modelling method for the integrated services, especially for the integration of HSD, becomes essential and is subsequently developed in the following chapter.

Chapter 4

Analysis of HSD in Integrated Services

In the previous chapter, the performance of low rate data in integrated services is analysed. However, high-speed data (HSD) is seen as the key for the successful delivery of complex data applications, such as wireless Internet and video services in the near future. Both high-speed circuit-switched data (HSCSD) and GPRS will support higher data rates than current GSM data services. In this chapter, two kinds of switching mechanisms for these high-rate data services are investigated in detail. More specifically, in Section §4.1, HSD services in circuit mode are analysed. Subsequently, HSD packet services are analysed in Section §4.2. Finally, reservation-TDMA, that is, PRMA integrated systems, are studied in Section §4.3. Summaries for each section are included.

4.1 Analysis of HSCSD in Priority Queuing Systems

In Chapter 3, an approximation approach was developed for integrating low rate data into voice services. In this section, a systematic MAM is proposed to analyse the integration of circuit-mode high rate data into voice services. For traffic studies, the important performance measures, such as voice distribution and data delay, are analysed in detail. In order to fine tune system performance, a hybrid reservation scheme is proposed to enhance voice call performance. As a result, we show that this flexible scheme outperforms the previous scheme, where the guard channel number is fixed during the system operation.

This section is organised as follows. The background is explained in Subsection §4.1.1. The assumptions for the model are described in Subsection §4.1.2. A hybrid reservation scheme is then proposed in Subsection §4.1.3. Then the problem is formulated in Subsection §4.1.4. Subsequently, a traffic model in the presence of mobility is investigated in Subsection §4.1.5. The results and summary conclude this section.

4.1.1 Introduction

Future PWC is expected to cater not only for voice conversations, but also data and moving image services. Even if voice service maintains its dominant position in the short-term, data and video services are rapidly emerging and are expected to become the leading

services in wireless systems. In order to deliver a requirement for high QoS, high bit rate transmissions must be provided by new transport mechanisms. HSCSD is theoretically designed to deliver data rates at about 115 kb/s, which is much higher than the present basic data rate [*Ojan98*]. The reason that HSCSD uses circuit mode is to provide a mixture of different data rate services by simply exploiting the existing physical layer structure.

In the standardisation of GSM phase 2+, circuit mode HSCSD and packet mode GPRS are considered as the prominent candidates to deliver high data rate services [*Cai97*]. In order to minimise the physical change for the present GSM systems, HSCSD service, which is assumed to employ paralleled channels by using a multislot structure, can be implemented in the existing infrastructure. In other words, HSCSD enables the users to use multiple TCH/F simultaneously for a single connection, in which traffic channels are allocated to HSCSD according to the rule of traffic channel increments. Examples include high-speed file transfer or low bit rate video applications. On the other side, GPRS is used for the delivery of bursty wireless access to the Internet, which is subsequently investigated in the next section.

Although HSCSD have been specified in [*ETSI234*] and [*ETSI334*], channel allocation issues and call control policies still remain open and need to be further investigated. The objective of this study is to pursue new call control policies, which can be used to enhance system performance. Although a CAC policy is proposed in [*Nagh96*], it has to be used in a distribution manner. Recently, a dynamic handoff control strategy is used in [*Rama99*]. However, as a result of tradeoff, the use of this policy might result in higher blocking probability for new calls. In addition, some previous studies of HSCSD can be found in [*Calin97b*] and [*Calin98*], in which a nonpreemptive priority queuing policy is employed and a finite queue is presumed. Meanwhile, a DNAmaca tool is used to solve the Markov chain in [*Calin97a*] and [*Calin97c*]. In contrast to a limited waiting mechanism, an infinite queuing buffer is assumed in this study. Subsequently, we propose a preemptive priority policy for the queuing system. More importantly, the performance of high rate data services is analysed by an efficient MAM with various loads. This method is not only suitable for the analysis of HSCSD, but also it is generic enough to pertain to other TDMA systems.

4.1.2 The Model Descriptions

In order to proceed with the analysis, we describe the assumptions for integrated services as follows:

1) For voice conversation, new arrivals follow a Poisson process at a rate λ_v. Meanwhile, handoff voice calls are denoted by another arrival rate λ_{hv} from adjacent cells. Accordingly, $\lambda_1 = \lambda_v + \lambda_{hv}$ is used to represent the aggregate arrival rate for voice calls. The available channels are C with the guard channels C_h, which are exclusively used for handoff calls. The service rates of voice and data calls are represented by μ_{vCH} and μ_{dCH} respectively. Similarly, data arrivals are confined to a Poisson process with a total mean rate $\lambda_2 = \lambda_d + \lambda_{hd}$, where λ_{hd} represents the arrival rate of data handoff calls. The channel allocation scheme can be shown as in [Figure 4.1].

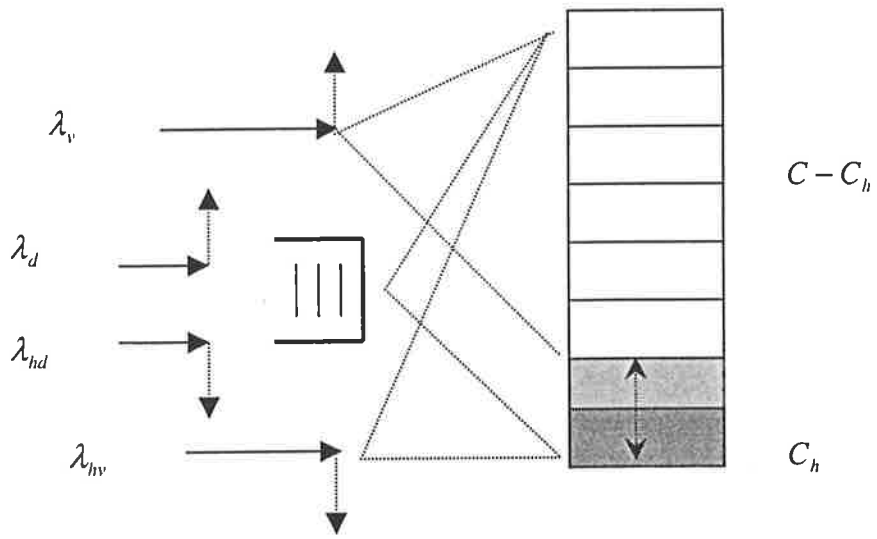


Figure 4. 1: HSCSD system queuing model

2) Upon arrival, it is assumed that handoff voice calls have priority over data calls. The queuing discipline is governed by the preemptive priority policy. That is, only handoff voice calls are allowed to preempt data calls while arrival handoff calls find all channels full. The preempted data calls return to the queue and wait for the next

available channel. Data calls can be scheduled in infinite buffers while channels are unavailable temporarily.

3) The number of voice calls is represented by j and the data calls by i respectively. The use of a multislot structure in HSCSD is assumed to be consecutive. Although a non-consecutive structure can be used, for the sake of simplicity, we are not going to discuss it in this study.

4.1.3 The Hybrid Reservation Scheme (HRS)

Since voice and data calls are allowed to contend for the same channel simultaneously in a random access manner, a 2-D Markov chain can be used for performance analysis. In order to favour handoff voice calls, a prioritised channel reservation scheme is adopted. Due to the use of multislot, the transmitted data bit rate can vary from one time slot up to a maximum number of n . The analytical method used in the study is that data call services will not start until the n data slots in a row are successfully seized [Figure 4.2].

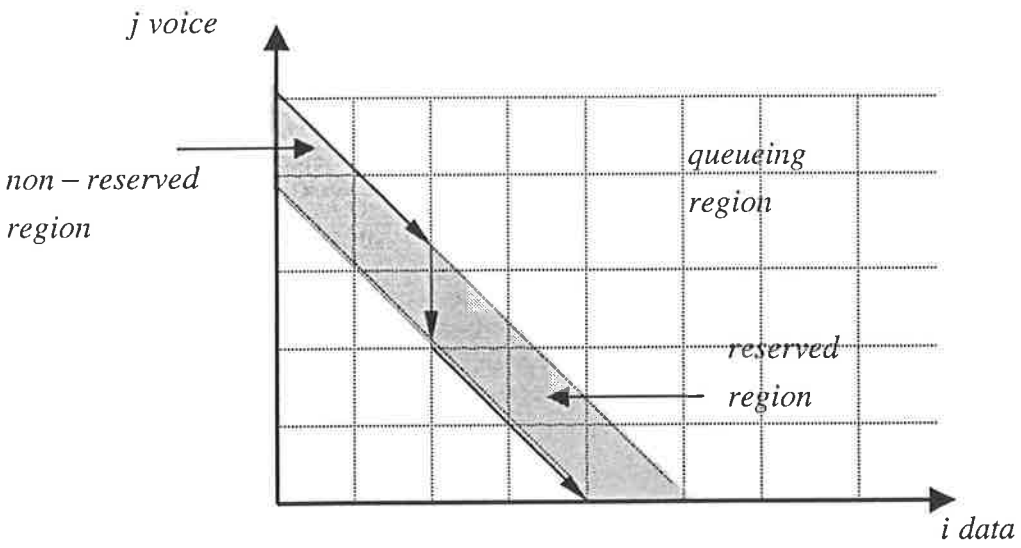


Figure 4. 2: The HRS for integrated services

In this study, based upon the reservation on demand (RoD) principle, we propose a new sharing resource scheme called hybrid reservation scheme (HRS). The number of reserved traffic channels is allowed to vary according to voice traffic loads. In other words, the

number of reserved traffic channels is determined by the demand of the voice arrivals. As a special case, this scheme can extend to the inclusion of a non-reservation scheme. As shown in [Figure 4.2], the flowing arrow shows the changing direction of reserved channels. The chosen criteria of the flowing direction will be dynamically determined by an optimal condition of the mixed traffic loads. In this study, we assume that the flowing direction comes from the number of channels C to $(C-C_h)$.

4.1.4 Problem Formulation

Based upon the aforementioned assumptions, the state space can be regarded as a two-dimensional Markov chain with state space $\Omega=\{(i,j):1\le j\le C, i\ge 1\}$. The state space is then divided into subsets, known as levels. More specifically, level i refers to $l(i)=\{(i,j):1\le j\le C\}$ for $i\ge 1$ and level 0 is called the absorbing state 0, whereas j refers to the phase of the state. In fact, there is a homogeneous structure along the levels.

In order to proceed with the analysis, we can construct an infinitesimal generator Q for the continuous time process with the block-partitioned form as:

$$Q=\begin{bmatrix} A_{01} & A_{02} & & & & & \\ A_{10} & A_{11} & A_{12} & & & & \\ & \cdots & \cdots & \cdots & & & \\ & & A_{r0} & A_{r1} & A_{r2} & & \\ & & & A_2 & A_1 & A_0 & \\ & & & & A_2 & A_1 & A_0 \\ & & & & & \cdots & \cdots \end{bmatrix} \tag{4.1}$$

where $r=\lfloor C/n \rfloor$ denotes the maximum data calls that can be served simultaneously.

Constructing the matrix of one step transition, we can show that A_{i2} and A_0 are scalar matrices with all elements of λ_2. A_{i0} and A_2 are the diagonal matrices with the dimensions of $(C+1)\times(C+1)$. We also have $A_{i1}=Q_{i1}-A_{i0}-A_{i2}$, where Q_{i1} is a tridiagonal matrix with the same dimension as A_{i2} and A_{i0}. In addition, we have $A_1=Q_0-A_0-A_2$.

Then we also have:

$$\begin{aligned}
&Q_0(0,0) = -\lambda_{hv}, \\
&Q_0(j, j+1) = \lambda_{hv}, && j = 0, 1, \cdots, C-1, \\
&Q_0(j, j-1) = j\mu_v, && j = 1, 2, \cdots, C, \\
&Q_0(j, j) = -\lambda_{hv} - j\mu_v, && j = 1, 2, \cdots, C, \\
&Q_0(j, l) = 0, && j \neq l, j \neq l \pm 1, l = 0, 1, \cdots, C, \\
&Q_0(C, C) = C\mu_v.
\end{aligned} \tag{4.2}$$

Then we are able to evaluate the equilibrium probability π_∞ if the continuous time Markov process is assumed to be irreducible:

$$\pi_\infty Q = 0 \quad , \quad \pi_\infty e = 1 \tag{4.3}$$

where e denotes a column vector of ones.

According to the Theorem 4.1 as described in [*Latou98*], for an irreducible continuous time quasi-birth-death (QBD) with transition matrix Q, there always exists a matrix-geometric invariant measure.

Therefore, the stationary distribution π_∞ has a matrix-geometric form and is given by:

$$\pi_\infty = [\pi_\infty^{(0)}, \pi_\infty^{(1)}, \cdots, \pi_\infty^{(r-1)}, \pi_\infty^{(r)}, \pi_\infty^{(r+1)}, \cdots] \tag{4.4}$$

Hence,

$$\pi_\infty = [\pi_\infty^{(0)}, \pi_\infty^{(1)}, \cdots, \pi_\infty^{(r-1)}, \pi_\infty^{(r)} \cdot R, \pi_\infty^{(r+1)} \cdot R^2, \cdots] \tag{4.5}$$

In fact, formula (4.2) comprises a set of linear equations for the solution of state space. We also note that the transition rates Q depend on state probabilities π_∞, whereas the state probabilities depend on the transition rates. In order to decouple these two parameters, there exist many traditional iterative methods, for example, Jacobi iterative method, Guass-Seidel method, or successive-over-relaxation (SOR) method for the solution of linear equations [*Rapp96*]. In addition, the bisection method and the successive substitution method for implicit nonlinear equations can be used to solve this problem as well. However, the difficulty of using these methods is the excessively computational time if the traffic loads become heavy and the states are large. In order to compute the recurrent states efficiently, we exploit an efficient algorithm for the solution of matrix R [*Latou93*].

And if the Markov chain is positive recurrent, we have:

$$\pi_{\infty}^{(r+k)} = \pi_{\infty}^{(r-1)} R^{k+1} \quad while \quad k = 0, 1, 2, \ldots \tag{4.6}$$

In addition, we can also construct another three matrices G, R and U for the continuous time Markov process. These matrices are characterised as the minimal non-negative solution of the non-linear equations. Their physical meanings can be found in [*Latou93*]. Therefore, matrix R has spectral radius less then 1 if the Markov chain is recurrent:

$$0 = A_0 + RA_1 + R^2 A_2 \tag{4.7}$$

In addition, matrix G is the minimal non-negative solution to the matrix quadratic equation:

$$A_2 + A_1 G + A_0 G^2 = 0 \tag{4.8}$$

Another substochastic matrix U can be shown as:

$$U = A_1 + RA_2 \tag{4.9}$$

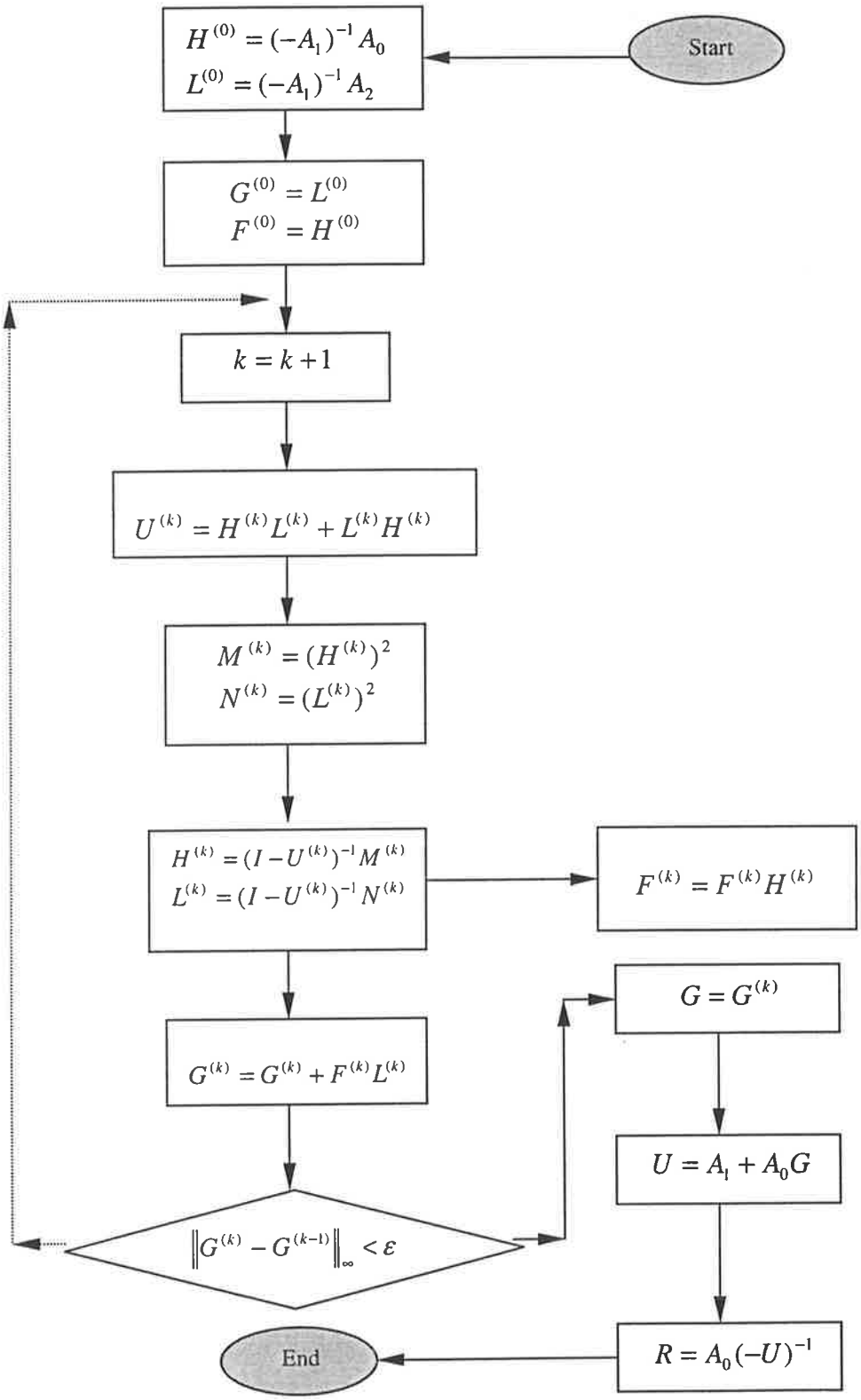
After some manipulations, these matrices can be changed into:

$$\begin{aligned} G &= (-U)^{-1} A_2 \\ U &= A_1 + A_0 G \\ R &= A_0 (-U)^{-1} \end{aligned} \tag{4.10}$$

In fact, the Quasi-birth-death (QBD) processes are the general case of the simple classical birth-death process of $M/M/1$. We find that this model belongs to the QBD with a complex boundary. In order to have the solution, we need to evaluate the rate matrix R. If the process is ergodic, the logarithmic reduction algorithm (LRA) can be adopted to solve matrix R iteratively, where ε denotes the stopping criterion of the iteration [Algorithm 4.1].

If $L(\varepsilon)$ denotes the number of iterations for the satisfaction of the stopping criterion, then the number of iterations for the computation G only approximates to $\lfloor \log_2 L(\varepsilon) \rfloor$. Therefore, by using the LRA, the matrices can converge more rapidly [*Latou93*]. This has been verified by our experiments.

Logarithmic Reduction Algorithm 4.1:



In order to maintain the stable condition of the queuing system, it must satisfy this condition:

$$\frac{\lambda_v + \lambda_{hv}}{\mu_{vCH}} + n(\frac{\lambda_d + \lambda_{hd}}{\mu_{dCH}}) < C \tag{4.11}$$

where n is the maximum number of data calls that can be accommodated.

4.1.5 Incorporating with Terminal Mobility

As mentioned above, performance evaluation needs to take spatial traffic variability into account. However, the study of cellular modelling poses some formidable challenges, which lies in the difficulties of the incorporation of terminal mobility into communication traffic. Communication must be maintained even during vehicle movement while the dedicated channel changes from one cell to another. For nonuniform traffic, the analysis is complicated and is usually time-dependent. Because we only concentrate on long term performance measures, uniform vehicular distribution in cells is assumed and then the time-independent analysis method is adopted.

The moving terminals considered in this study are confined to two assumptions:

1) The mobile terminals and the spatial traffic distribution are uniformly distributed over a specific cell.

2) The mobile terminals have a constant speed under the consideration of randomised moving direction.

The average number of handoffs in a call lifetime can be calculated from the ratio of the mean call holding time $1/\mu_v$ to the mean cell sojourn time $1/\eta$ as η/μ_v. Based on the assumptions as mentioned above, we use V for the average vehicle speed, L for the perimeter of the cell, R_c for the cell radius and S for the area of the cell. If the direction of the moving terminals is uniformly distributed within the cell $[0, 2\pi]$, it is easy to show that the average cell crossing rate is given by $R_s = VL/\pi S$. Because the channel holding time in the cell is either the unencumbered duration or cell sojourn time, whichever is less, it can be shown that the mean channel holding rate is:

$$\mu_{vCH} = \mu_v + \frac{2V}{\pi R_c} \tag{4.12}$$

In the meantime, because we already use j to denote the number of voice calls and i for the number of data calls, the average number of data calls that sojourns in the system can be expressed as:

$$S = \sum_{j=0}^{C}\sum_{i=0}^{\infty} i\pi_{\infty}^{(ji)} = \sum_{j=0}^{C}(\pi_{\infty}^{(j0)} + \pi_{\infty}^{(j1)}(I-R)^{-2}) \tag{4.13}$$

where I is a unit matrix and R is the rate matrix.

In the end, the data call delay in the system can be calculated from: $D = \dfrac{\sum_{j=0}^{C}\sum_{i=0}^{\infty} i\pi_{\infty}^{(ji)}}{\lambda_2}$.

4.1.6 Discussion of the Results

As a numerical example, we assume that voice traffic can only have one traffic time slot and data calls can occupy three contiguous time slots. The total available traffic channels are equal to eight. The mean service time of both voice and data calls lasts for three minutes. The handoff arrival rates of voice and data calls are equal to ten percent of the originating arrival rates.

First, if voice and data call traffic loads are kept constant and the number of reserved channels changes from $R_{ch}=1$ to $R_{ch}=3$, we can have this result showing the relationship of new voice blocking probability and data call offered loads (Erlangs) as in [Table 4.1].

Table 4. 1: The effect of reserved channels on voice blocking probability

ρ_d	$P_{R_{ch}=1}$	$P_{R_{ch}=2}$	$P_{R_{ch}=3}$	$\frac{P_{R_{ch}=2}}{P_{R_{ch}=1}}-1$	$\frac{P_{R_{ch}=3}}{P_{R_{ch}=1}}-1$
0.001	0.0058	0.0184	0.0509	217%	177%
0.101	0.0252	0.0507	0.1005	101%	98%
0.201	0.052	0.0865	0.1487	66%	72%
0.301	0.0851	0.1254	0.1962	47%	56%

We observe that the new voice blocking probability distribution versus data arrivals will significantly increase according to the addition of reserved channel numbers. Although the larger data loads lead to less degradation, the degradation is still quite significant. For instance, the blocking probability increases by 47% while the reserved channels increase from one to two. This suggests that the number of reserved channels must be optimally chosen even if a certain number of prioritised channels can be used in favour of the forced termination probability.

For the RoD scheme, we have the results as shown in [Figure 4.3].

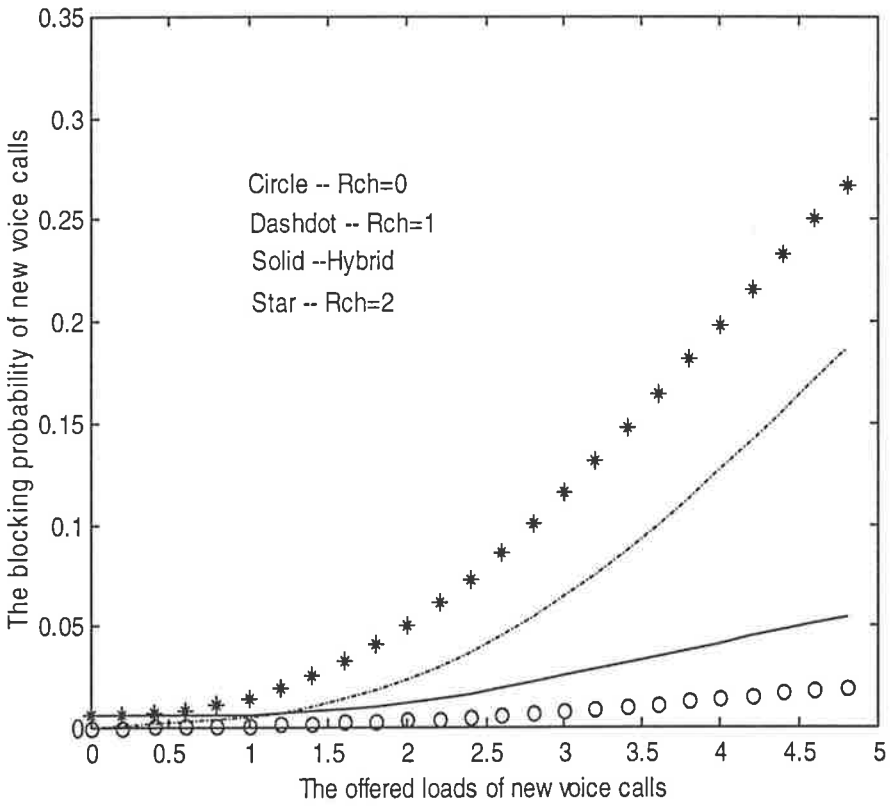


Figure 4. 3: The performance of hybrid reserved scheme

We compare this result with the different fixed prioritised schemes. The worse case is when the number of reserved channels is equal to two. This lies in the large number of reserved channels restraining new call from gaining access. In this study, while voice arrivals are lower than 1.2 Erlangs, the new call blocking probability by using HRS is no better than that by using a reserved channel $R_{ch} = 1$. This is because this hybrid scheme

actually uses $R_{ch} = 2$ while the voice arrivals are low. However, the blocking probability shows significant improvement while the new voice loads are higher than 1.2 Erlangs. For instance, while the arrival load is equal to 2.6 Erlangs, the new voice call probability is 4.5% by using $R_{ch} = 1$ and 8.6% by using $R_{ch} = 2$. The new voice call probability achieves near 2% with the same arrival load while there is no reserved channels. As a result, the RoD scheme exhibits more robustness for traffic performance than using the conventional scheme.

Once voice terminal mobility is taken into account under the fixed scheme, we have the result shown as in [Table 4.2].

Table 4. 2: The effect of mobility on new voice call blocking probability

ρ_v	P_{B0}	P_{B30}	P_{B60}	$1 - \frac{P_{B30}}{P_{B0}}$	$1 - \frac{P_{B60}}{P_{B0}}$
0.001	0.0218	0.0218	0.0218	0%	0%
0.101	0.0227	0.0226	0.0225	0%	0%
0.201	0.0252	0.0247	0.0243	2%	2%
0.301	0.0288	0.0278	0.0269	3%	3%
0.401	0.0333	0.0316	0.0303	5%	4%
0.501	0.0384	0.036	0.0342	6%	5%
0.601	0.0439	0.0409	0.0385	7%	6%

From this table, not surprisingly, we observe that the new voice blocking probability, e.g., 2%, is not changed according to the increased terminal speeds from 0 to 30km/hour and then to 60 km/hour while the voice loads are equal to 0.2 Erlangs. Even if the traffic loads are increased, the blocking probability barely changes. Therefore, the new call loss properties are insensitive to terminal mobility.

4.1.7 Summary

For the study of the integration of HSCSD into voice service, the main contribution of this section is that we propose an efficient MAM to quantitatively analyse the prioritised

traffic schemes with the incorporation of spatial traffic variability. The impact of terminal mobility and guard channels on the performance of voice and data calls has been analysed under various loads. New call blocking probability is found insensitive to terminal mobility. The stable condition of the queuing system under the reservation scheme is subsequently discussed. The analysis of voice distribution and data delay is subsequently derived. Voice performance exhibits an oscillatory phenomenon because the multislot data packets initially compete for the same traffic channel with voice calls. As mentioned in Chapter 2, the use of prioritised handoff schemes can improve handoff failure probability at the expense of increased call blocking probability. In order to optimise handoff call performance, HRS is then proposed. As a conclusion, comparing the HRS with the conventional prioritised handoff scheme, we conclude that this flexible scheme, which relies on channel measurements, is robust enough to enhance overall voice call performance. In addition, this scheme improves the performance of integrated services in terms of flexibility as well as effectiveness.

4.2 Data Traffic Distributed Control Scheme in HSD services

In the last section, the performance of HSCSD services has been studied. In contrast, packet mode data services are investigated in this section. A MAM is developed to analyse the integrated high data bit rate into voice calls. This method can achieve quadratic convergence compared to the conventional spectral methods. The voice call distribution, data packet throughput, delay and waiting time distribution are derived subsequently. A new priority scheme and a voice coding rate control scheme are proposed to mitigate the packet congestion. It shows that larger packets cause longer latency and terminal mobility becomes an indispensable element in influencing both voice and data performance. This section is arranged as follows. The background is introduced in Subsection §4.2.1. Subsection §4.2.2 describes the assumptions used in the model. In Subsection §4.2.3, the matrix-analytic method is used to analyse the prioritised integrated services with the consideration of mobility. We then propose a multipriority scheme in Subsection §4.2.4 followed by a discussion of the results in Subsection §4.2.5 and some conclusions in Subsection §4.2.6.

4.2.1 Introduction

As aforementioned, the evolving and future PWC systems are expected to support a wide range of new services, such as high-speed WWW and video services. As these services are introduced, traffic volume will increase significantly. Because the characteristics of voice and data require different service mechanisms, some efficient control strategies may have to be enforced in order to provide for the stringent QoS requirements.

In general, service integration is achieved by the dynamic and efficient sharing of limited resources. In a previous study of integrated voice and data services, a movable control scheme was employed for a wireless indoor system [*Zhang90*]. It shows that a two-dimensional Markov chain for voice and data channel access closely matches simulation results. An Enhanced-TDMA system, based on the North American digital cellular standard (IS-54), has been analysed in [*Li94*]. Recently, another analysis of voice and data integrated service has appeared in [*Calin97a*]. However, these analyses have only been conducted for single voice and single data slot cases and do not include terminal

mobility and high data rates, which can occur with the full exploitation of the timeslot structure.

In order to provide a high data rate based on the present systems, there are two methods for increasing transmission rates. The first one is to use paired up multiple carriers simultaneously, which is analysed in [*Serres88*]. In contrast, the second method is to exploit the time multiplexing structure by stringing contiguous slots together. The latter can be used in the environment of high-speed data packet cellular systems [*Honk94*] and [*Hamal95*]. Compared to the wireline situation, the time frame structure is more difficult to analyse due to the correlation of time slot arrangements. As mentioned in last section, HSCSD use multislot configuration, in a similar way, high rate data packet can also use multislot structure [*ETSI502*]. However, for the use of circuit mode HSCSD services, although the advantage is of the isolation for different traffic classes, which can eliminate the interaction of different traffic sources, the signalling channels eventually limit the channel utilisation [*Bohm96*]. In addition, the drawbacks of circuit-switched networks with fixed capacity and long set-up delay make them unsuitable for the bursty data packet transmissions. Therefore, packet mode data services are investigated in this section. Packet switching has apparent advantages for the transmission of bursty data services. For instance, with the present technology, GPRS is going to expand the present 9.6 kb/s data rate in GSM to about 170 kb/s theoretically as described in [*Ojan98*].

From the simulation results shown in [*Cai97*], multislot data packets system can have the advantages of high throughput and low delay over a conventional single slot system. Although some other simulation studies of multislot performance can be found from [*Cai97*], [*Wang95*], and [*Turin96*], a more general analytical framework for such integrated access is rather complex and needs to be further developed under prioritised integrated environment [*Jabb96*]. Our study is motivated on this basis. In our context, a high-speed data channel is defined as the number of stringed slots that are required to achieve the specified high data rate. The performance of low rate voice calls and high data rate is analysed by a MAM under an efficient algorithm with various traffic loads. Our approach is also different from Serres's method [*Serres88*]. Serres's method only pertains to the case of wireline systems. Although a randomised control scheme is used by another recent analysis [*Kaya98*], prioritised traffic and terminal mobility are excluded in that study.

In addition, in order to increase the capacity of the cellular network, a small cell configuration is often used to increase capacity. As an adverse effect, the frequent movement of terminals across the cell boundaries can significantly increase the signalling loads for network control. Thus the problem of the volume of control information exceeding the controller processing capacity becomes a growing concern [*Meier94*]. To overcome this problem, we propose a priority policy to mitigate overload problem, which is intended to operate in a decentralised fashion. This scheme is effective and novel because it can select data packets by classes rather than through the use of the conventional rejection or discard mechanisms. Our approach is generic and can be applied to other slotted systems as well. The objective of this study is to investigate the interactive relationships of voice and data and to improve data transport under congestion conditions.

4.2.2 The Model Descriptions

In the analysis of voice call and data packet integrated services, traffic performance largely depends on the use of multiple access techniques and the method of scheduling time slots. As a result, these techniques can significantly influence both terminal complexity and system architecture.

The following assumptions allow us to obtain a tractable solution:

1) Voice calls are assumed to arrive at a rate λ_v with exponential distribution. Similarly, handoff voice calls can be represented by another arrival rate λ_{hv} from adjacent cells. We use $\lambda_1 = \lambda_v + \lambda_{hv}$ to represent the aggregate arrival rates.

2) The number of total available capacity is assumed to be C channels with a specific reserved channel C_h, which is exclusively designed for handoff voice calls. That is, handoff calls are always given priority over new voice calls in the areas with frequent handoffs.

3) Data packet arrival process is approximated by Poisson process with a mean arrival rate λ_2. Data message length is exponentially distributed with average length M kbits. The actual structure is consistent with the descriptions detailed in [*Zhang90*], [*Wilson93*], and [*Falk83*]. To overcome the difficulty of a possible fragmented packet burst, we allow the fragmented burst to transmit across a frame boundary

[*Falk83*]. Moreover, voice calls are allowed to have preemptive priority over data packets, once an arriving voice call finds all channels full.

4) From our assumption, the mobility of data terminals and handoff arrivals of data packets are not included because of the nature of relatively short packet length and the ability of retransmissions due to channel distortion or transmission errors. A similar assumption can be found from [*Wang95*]. Additionally, the service rates of voice and data packets are shown by μ_1 and μ_2 respectively. The number of voice calls is shown by j and data packets by i respectively.

Therefore, we can adopt a two-dimensional canonical setting to represent the mixed traffic state spaces as shown in [Figure 4.4].

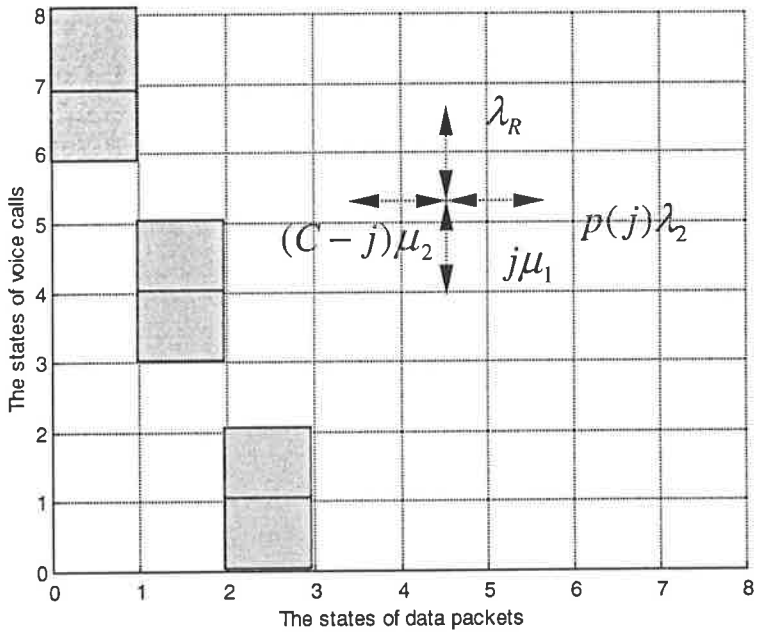


Figure 4. 4: The interactive state spaces for voice and data packets

The shadows show the time slots for reservation with arrival rate λ_{hv}. That is, $\lambda_R = \lambda_{hv}$ if reservation exists, otherwise $\lambda_R = \lambda_1$.

In addition, the channel unit in our study is defined as one of the available time slots designed for available traffic access. Due to the use of multislot arrangements, data packet bit rates can vary from the original single time slot up to a maximum number n. In

practice, the voice calls channel access usually adopts a time-out model, but in our analysis, voice calls are simply assumed to be lost if all channels are full.

4.2.3 Analysis with Explicit Mobility Modelling

- **Mobility Model**

As mentioned in last subsection, mobility is seen to be at the heart of PWC. The inclusion of terminal mobility in a traffic model is very important because it can determine actual cell sojourn time. As a consequence, network congestion may occur due to frequent handoffs. More importantly, terminal movement can affect handoff dropping probability and reduce voice call QoS [*Hong86*]. Although some studies for voice-only exist in [*Hong86*] and [*Lin94*], we investigate how the terminal mobility of voice calls affects data packets and how to enhance packet performance in this study.

In order to proceed with the analysis of the integrated services, the explicit mobility model is assumed to be similar to the one as already described in Section §4.1.5. Because channel holding time in a cell is the minimum of call holding time and cell sojourn time, then the mean channel holding rate becomes:

$$\mu_{vCH} = \mu_v + \eta = \mu_v + (2V / \pi R_c) \tag{4.14}$$

Therefore, we can proceed to the mathematical analysis as below.

- **Mathematical Definitions and Analysis**

In order to analyse the joint voice and data performance, many approximation techniques can be used, but here we employ the matrix technique. Then we can form a one-step transition rate matrix Q as shown below [*Ramas85*]:

$$Q = \begin{bmatrix} A_{01} & A_{02} & & & & & & \\ A_{10} & A_{11} & A_{12} & & & & & \\ & \cdots & \cdots & \cdots & & & & \\ & & A_{r0} & A_{r1} & A_{r2} & & & \\ & & & A_2 & A_1 & A_0 & & \\ & & & & A_2 & A_1 & A_0 & \\ & & & & & \cdots & \cdots & \end{bmatrix} \tag{4.15}$$

where $r = \lfloor C/n \rfloor$ denotes the maximum multislot data packets in servers and $\lfloor x \rfloor$ represents rounding down to the nearest integer.

We can now calculate the equilibrium probability π_∞ once the Markov process is assumed to be irreducible:

$$\pi_\infty Q = 0 \quad , \quad \pi_\infty e = 1 \tag{4.16}$$

where e is a one column vector.

By constructing the sub-matrices, we can show that A_{i2} and A_0 are the dimension $(C+1)\times(C+1)$ scalar matrices with all diagonal elements λ_2. A_{i0} and A_2 are the diagonal matrices with the diagonal elements $(C-j)\mu_2$, for $j = 0,1,\ldots,C$. The matrices of A_{i1} are tridiagonal with order $(C+1)$. Then we can have $A_{i1} = Q_{i1} - A_0 - A_2$. The elements of Q_{i1} are given as below:

$$\begin{aligned}
&Q_{i1}(0,0) = -\lambda_R \\
&Q_{i1}(j, j+1) = \lambda_R \ , \ j = 0,1,\cdots,C-1 \\
&Q_{i1}(j, j-1) = j\mu_v + (2jV)/(\pi R_c), \\
&\qquad\qquad j = 1,2,\cdots,C \\
&Q_{i1}(j, j) = -\lambda_R - j\mu_v - (2jV)/(\pi R_c), \\
&\qquad\qquad j = 1,2,\cdots,C \\
&Q_{i1}(C, C) = C\mu_v + (2VC)/(\pi R_c) \\
&Q_{i1}(j, l) = 0, \ \ j \neq l, j \neq l \pm 1, l = 0,1,\cdots,C
\end{aligned} \tag{4.17}$$

where the meaning of λ_R is already defined as in Subsection §4.2.2.

In addition, $A_1 = Q_0 - A_0 - A_2$, where Q_0 is similar to the formula (4.17) without the consideration of reservation.

In order to compute the recurrent states, an improved logarithmic reduction algorithm can be employed [*Latou93*]. The technique used here is different from directly solving the linear equations. We employ a more efficient method by solving the matrix quadratic equations. In general, as described in Subsection §4.1.4, in order to have the solution of the transition rate matrix, many algorithms that are based on a modification of Newton's method have been developed for quadratic convergence. Here we use an efficient method to solve the rate matrix R rather than by the conventional spectral decomposition method.

By using this method, matrix R has a spectral radius less then 1 if the Markov chain is recurrent. Therefore, we can have this relationship for the solution of rate matrix R:

$$0 = A_0 + RA_1 + R^2 A_2 \tag{4.18}$$

and hence

$$\pi_{\infty}^{(r+l)} = \pi_{\infty}^{(r-1)} R^{l+1} \quad while \quad l = 0, 1, 2, \ldots \tag{4.19}$$

Therefore, we can use the logarithmic algorithm as derived in the last section.

Additionally, in order to maintain the stable condition of the queuing system, it must also satisfy:

$$\frac{(\lambda_v + \lambda_{hv})}{(\mu_v + (2V)/(\pi R_c))} + n\sum_{i=0}^{r} i\pi_{\infty}^{(ji)} < C \tag{4.20}$$

The average number of data packets sojourning in the system can be calculated from:

$$S = \sum_{j=0}^{C}\sum_{i=0}^{\infty} i\pi_{\infty}^{(ji)} \tag{4.21}$$

We also can use $b = \lfloor (C - j)/n \rfloor$ to show the integer number of data packets in services and thus we have $r = b_{\max}$.

- **The Performance of Data Packets**

In fact, it can be found that this model is the special case of the QBD waiting time distribution with a complex boundary. Because such a stationary queue distribution is a type of modified matrix-geometric, we can exploit the modified matrix-geometric technique developed in [*Ramas85a*]. Actually, the waiting time distribution of the QBD can be seen as the time until absorption in the Markov process with the infinitesimal generator. The key difference of this numerical analysis for the waiting time distribution is the avoidance of the use of differential equations, which can lead to a more complicated analytical procedure as shown in [*Neuts81*].

For waiting time distribution, we have to keep track of the number of calls in the system and the elapsed time since the last arrival. Because the data packets are only allowed to queue, the waiting time of k data packets in the queue can be characterised as

the k-stage Erlangian arrival time distribution. However, the queue length distribution is not a Markovian process. If we only consider that a phase arrival occurs after the total data packets k have completed service, this model still belongs to the classes of quasi-birth-and-death (QBD). Finally, because data packets are only allowed to queue, packet waiting time can be viewed as the time of the process until reaching absorption state. The infinitesimal generator Q_w can be shown as below [*Ramas85*]:

$$Q_w = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ A_2 & A_0 + A_1 & 0 & \cdots & 0 \\ 0 & A_2 & A_0 + A_1 & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots \end{bmatrix} \tag{4.22}$$

Let $W(t)$ denote the data packet waiting time distribution in a queue following the FIFO discipline. In practice, packet delay is constrained by retransmission. If the maximum allowable data packet holding time is denoted by $D_{s\max}$, the probability that a multislot data packet arrival in an arbitrary time has to wait longer than time $D_{s\max}$ $(D_{s\max} > 0)$ can be calculated from:

$$W(D_{s\max}) = \sum_{m=0}^{\infty} d_m e^{-\theta D_{s\max}} \frac{(\theta D_{s\max})^m}{m!} \tag{4.23}$$

where

$$H_0 = I, \quad H_m = H_{m-1} P_1 + R H_{m-1} P_2 \quad while \quad m \geq 1$$

$$d_m = \pi_{c-1} (I - R)^{-1} R H_m e \quad while \quad m \geq 0$$

and

$$\theta = r(Max(-(A_0 + A_1)_{jj}))$$

$$P_1 = \frac{1}{\theta} (A_0 + A_1) + I \quad , \quad P_2 = \frac{1}{\theta} A_2$$

Moreover, even the higher moment of waiting time distribution can be also obtained through this study. The h^{th} moment of long term waiting time can be computed by:

$$W^{(h)} = h\theta^{-h} \sum_{m=0}^{\infty} d_m \frac{(m+h-1)!}{m!} \quad while \quad h \geq 1 \tag{4.24}$$

In general, higher moments are used to characterise packet variability, which is important for bursty packet traffic.

4.2.4 The Multiple Priority-Based Scheme (MPBS)

The adaptive control scheme proposed here is able to operate in a distributed manner, in which the controllers in each base station can be used to sense the state information of voice calls. Then they can make decisions to sift the classifications of packets individually based upon the collected information. In other words, the controllers can behave alone within different cells and thus control the admission of packets by priority class.

In order to exploit the leftover capacity of voice calls, data packets are assumed to be time-insensitive and able to queue. However, as an adverse effect of queuing, data packets can incur long delay once the input loads become high. Usually, the conventional non-priority control mechanism is to discard the older packet [*Nanda94*]. However, some significant messages may become lost due to the enforcement of time-out control policies or congestive conditions. Therefore, such a non-prioritised strategy will cause the loss of some important messages when input loads are beyond a certain threshold.

The method proposed here is different from the one used in [*Goodm89*] and the one adopted in dynamic channel as well [*Everitt89*]. Goodman adopts the permission probability p and then uses the permission probability as an important design parameter. Obviously, this can prevent users without permission from gaining access to traffic channels. Similarly, Everitt suggests only a fixed number of calls at each cell. In fact, these methods belong to the case of explicit rejection control schemes. In contrast, our proposal is to classify the ranking of data users and then try to restrain the low priority ones. As a result, this can increase flexibility and add some degree of fairness to data packets. In addition, although Williams [*Willi84*] used flow control scheme for data packets, the threshold of data packet throttling must be measured according to packet traffic loads. This may be difficult to implement if data offered loads are fluctuating and data lengths are relatively short in nature.

In order to alleviate the congested behaviour of data packets, a flexible priority admission policy named the Multiple Priority-Based Scheme (MPBS) is proposed in our study. The basic idea is to keep the short-term arrival loads as low as possible while meeting the minimum data throughput as required.

Using this policy, the population of data users is classified into k classes $(k \geq 1)$ according to different priorities from low to high levels as well as variations of voice traffic. The lower priority classes have less privilege for obtaining service than the higher priority classes.

For instance, the process can be represented with the state space $\Omega = \{(i, j) : 0 \leq j \leq C, i \geq 0\}$ under our assumptions as before. The k-class prioritised data packets arrival matrix Λ_d in the uniform admission regions can be shown as below:

$$\lambda_d = \begin{cases} p_k \lambda_2 & 0 \leq j < \left\lfloor \frac{C}{k} \right\rfloor \\ p_{k-1} \lambda_2 & \left\lfloor \frac{C}{k} \right\rfloor \leq j < \left\lfloor \frac{2C}{k} \right\rfloor \\ \cdots & \cdots \\ p_1 \lambda_2 & (k-1)\left\lfloor \frac{C}{k} \right\rfloor \leq j < C \end{cases} \tag{4.25}$$

If we choose priorities from high to low, then we can use $p_1(j) \leq p_2(j) \leq \cdots \leq p_k(j)$ to represent the orders of priorities in a decreasing chronological order.

For classification, the number of classes can be chosen according to the significance of data packets coming from each stream. For implementation, our control scheme can be applied with the method of *tagged* customers for the variable priority classes. The tagged customers can be indicated by packet headers, which is used to inform the base stations. Therefore, a f-bit header can maximally represent up to $k = 2^f$ level priorities. Then the base station can make a decision to accept or discard data packets with the marked priority classes based on computing voice traffic channel occupancy. Once data packets are discarded by the base station controller, the base station will send a negative acknowledgment to inform the terminals so that the data packets can be retransmitted. This can be shown as in [Table 4.3].

Table 4. 3: Priority classes for the MPBS

Header	Priority ranking
00	Highest priority
01	Second priority
10	Third priority
11	Lowest priority

4.2.5 Results Discussion

In order to verify the result from the analytical solution, we can show the simulation result as in [Figure 4.5].

In this simulation, we assume that the number of reserved channels is $C_h = 0$ and data packets are delivered in a 9.6 kb/s channel. The new voice call arrival rate and handoff call arrival rate are 0.02 calls/s and 0.002 calls/s respectively. In the end, we observe that the simulation result is in good agreement with the analytical result.

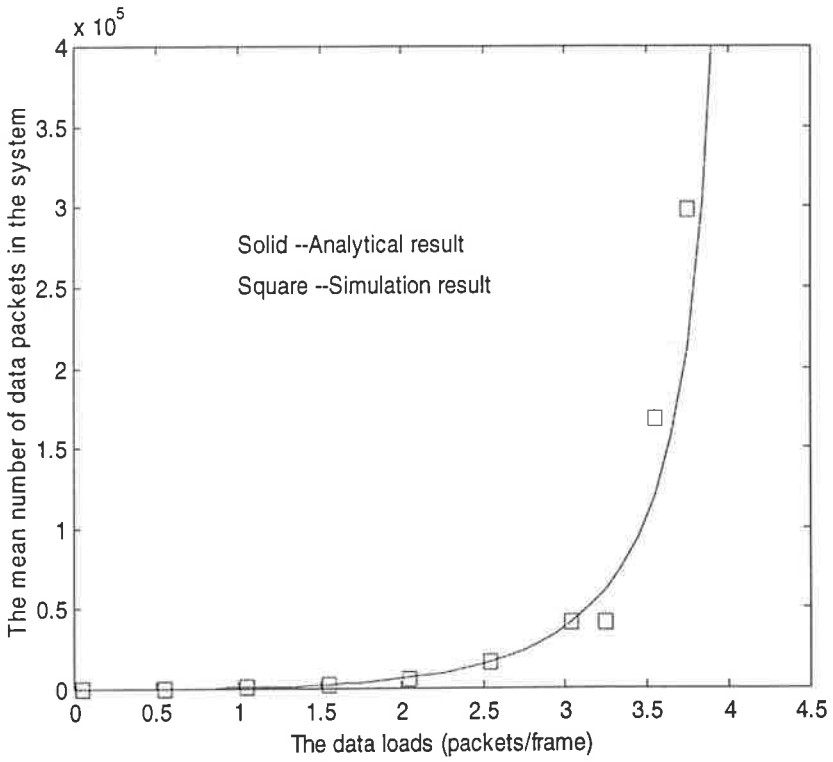


Figure 4. 5: The simulation result for data packets

As a numerical example, we assume parameters similar to those used in the present GSM system. However, our approach is generic and can be easily applied to other slotted systems. We assume that there are eight time slots in a frame. Each frame duration equals to 4.615 ms. Voice traffic can only occupy one traffic time slot and data packets can have as many as three contiguous time slots. In brief, the system parameters used in our numerical study can be listed as [Table 4.4]. First, we study the influence of voice terminal movements on the performance of data packets. We use the parameters as shown in [Table 4.4].

Table 4. 4: Typical parameters

Item	Symbol	Value
Reservation (Unit)	C_h	2
Voice duration (mins.)	$1/\mu_v$	3
Handoff rate (calls/s/user)	λ_{hv}	0.002
New call rate (calls/s/user)	λ_v	0.02
Cell radius (km)	R_c	10
Voice terminal speed (km/h)	V	0, 30, 60

In addition, data transmission rate is assumed to be 28.8 kb/s and average packet message length is $M = 1$ kbits. Because the use of priority reservation can apparently improve the forced termination probability [*Hong86*], we assume that the voice calls can have C_h=2 reserved channels as shown in [Figure 4.6].

For instance, we increase the speed of voice terminals from a stationary state to 30 km/h and then 60 km/h. During the periods of light loads, such as offered loads smaller than 1.5 Erlangs, the changing of average data packets due to users' movements can be negligible. However, with the increase of data loads, the mean number of data packets can have a dramatic increase. For example, while the offered load of data packets is equal to 3 Erlangs, the average number of data packets can drop 48% even if the terminal speed only changes from 0 km/h to 30 km/h. Therefore, we find that a specific highly loaded cell of

voice calls with high mobility can gain some benefits to the transmission of data packets because more data packets can be simultaneously served at that time.

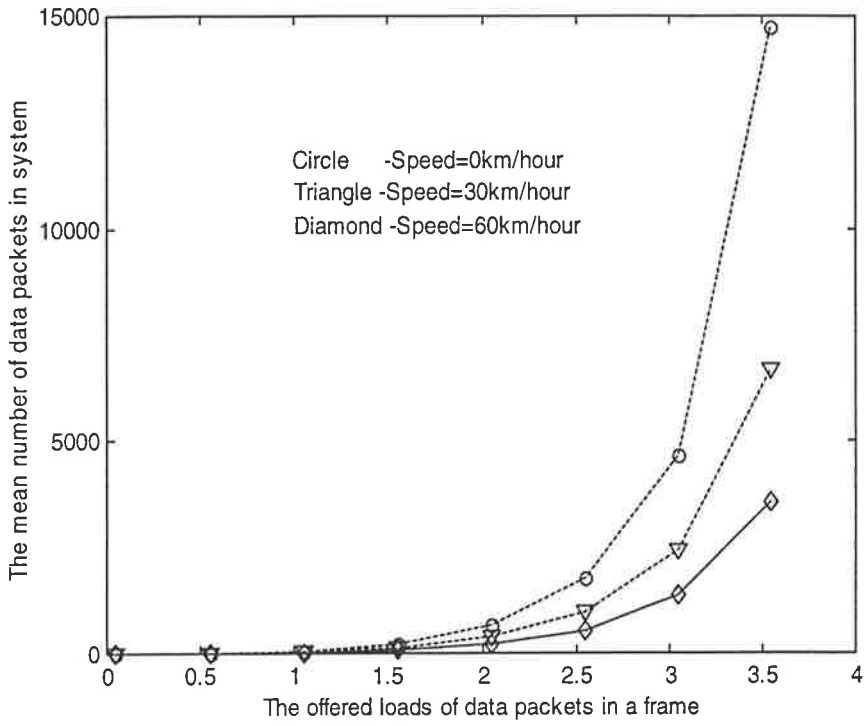


Figure 4. 6: The mobility effect on multislot data packets

This may also show that there exists an apparent correlative relationship between voice terminal mobility and data packet delay under the mixed environment. The reason for this is that the movement of voice terminals causes the reduction of the channel holding time. Thus the more data packets that can be served, the less the delay that occurs.

Accordingly, we plot the use of the control scheme as shown in Figure 4.7.

For the two-level scheme, we adopt the diagonal vector of the arrival matrix as $\Lambda_d = \lambda_2 diag[1,1,1,1,1,0.8,0.8,0.8,0.8]$. The three-level arrival matrix has $\Lambda_d = \lambda_2 diag[1,1,0.8,0.8,0.8,0.6,0.6,0.6,0.6]$. If the packet arrival loads are 2 Erlangs, we observe that the use of two-level MPBS can force delay packets to drop up to 58% and 86% for the three-level MPMS. Similarly, if arrival loads become 3 Erlangs, we can find that there is a reduction rate of 62% for the use of two-level MPBS and 90% for the three-level MPBS. Obviously, the improper choice of coefficients of the arrival matrix can drive data packets to have low throughput below the requirement. Be aware that we do not intend to address the optimal value for the coefficients here. This leaves for further study.

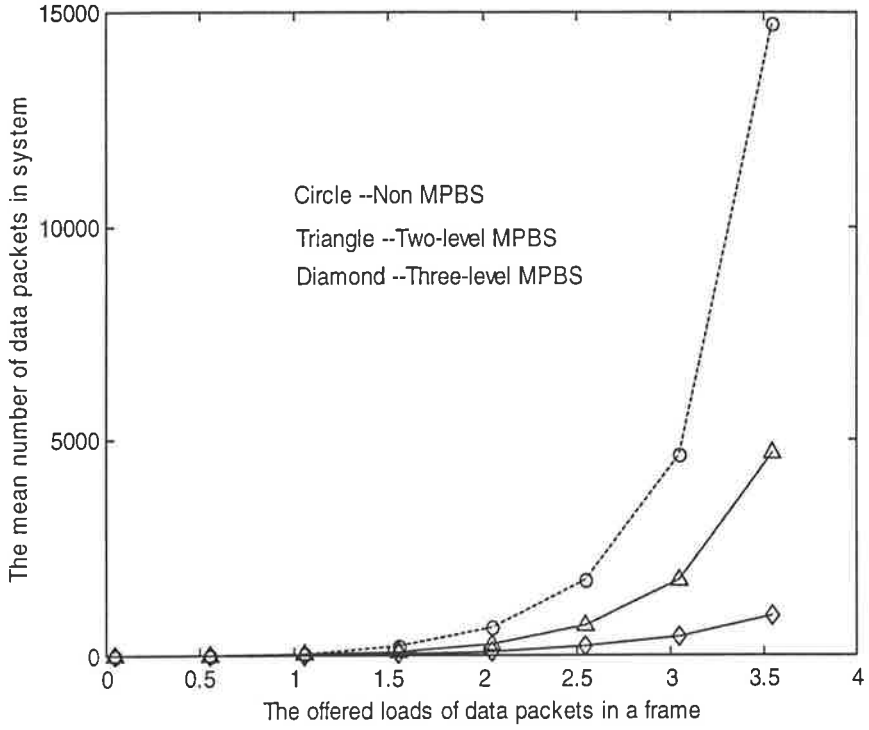


Figure 4. 7: The delayed data packets with MPBS

In practice, the choice of the number of priorities depends on the actual voice traffic and the required QoS for data packets. Thus this scheme is rather flexible.

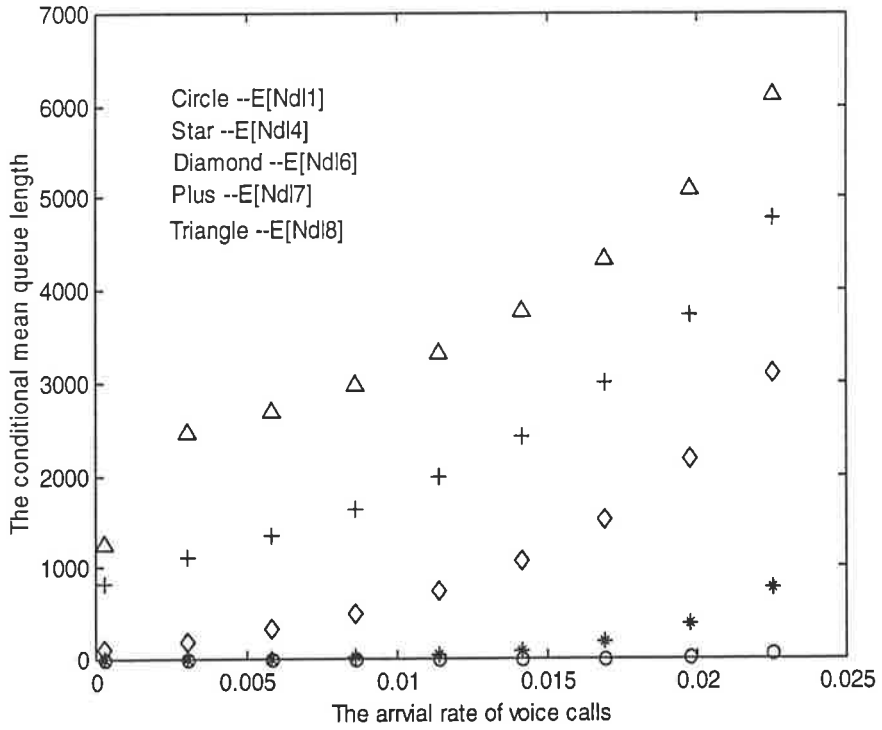


Figure 4. 8: The influence of voice calls packet length

As a conclusion, The higher packet loads can yield better packet reduction rate. Moreover, it suggests to us that the classifications of prioritised data packets can yield lower data packet delay in congested regions and thus enhance the performance of data packets in future PWC environments, when the input loads of data packets become high.

Finally, conditioned on the voice calls, we can plot the expected queue length of data packets in [Figure 4.8].

It shows that the queue length can easily accumulate with the increase of the number of voice calls. In order to relieve the congestion of the conditional queue length, we assume that the voice coding rate can be reduced to half as shown in [Figure 4.9].

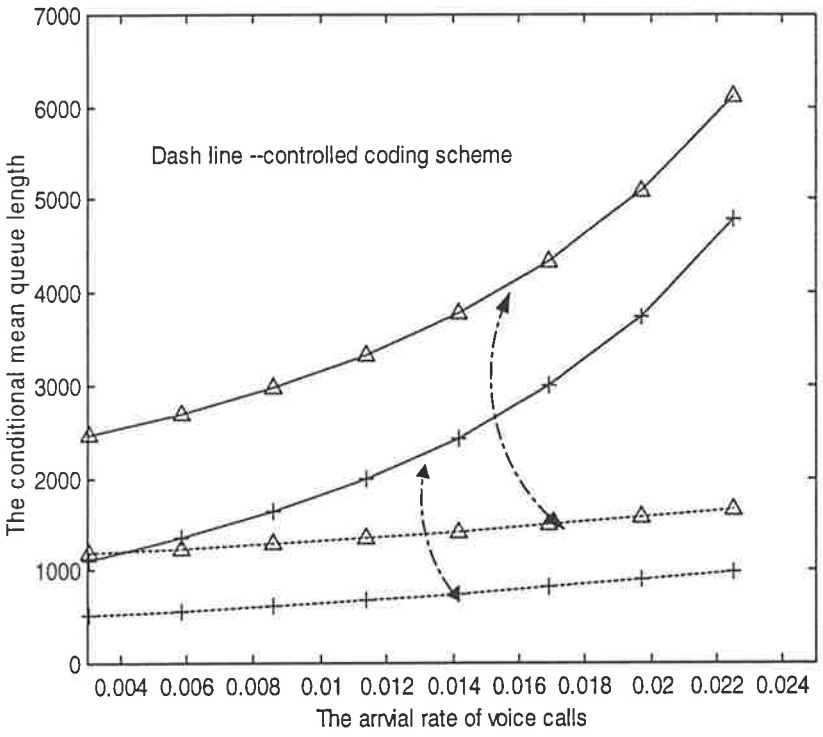


Figure 4. 9: The use of voice controlled coding scheme

The arrival data packets are assumed to be 60 packets/s. With the dashed lines, it shows that the conditional queue length can be decreased when compared with the original solid line.

4.2.6 Summary

Future cellular systems are envisioned to support mixed traffic, and ultimately multimedia services. However, a mixture of voice and data requires novel service

mechanisms that can guarantee quality of service. In order to transfer HSD, multislot channel allocation is seen as a favoured solution to the present systems with the least compromise to circuit-switched services.

The main contribution of this section is that we propose an efficient MAM to analyse the integration of HSD services with voice service in a frame structure. This method achieves quadratic convergence compared to the conventional spectral methods. Mobility is also considered in a prioritised cellular environment where frequent handoff has the potential of degrading data performance. The voice call distribution, data packets throughput, delay and waiting time distribution are derived. Moreover, the other contributions are that a new multiple priority-based distributed control algorithm and a voice rate control scheme are proposed to mitigate the queuing congestion of data packets.

We show that larger data packets incur longer latency and voice terminal movement becomes a decisive factor in the influence of both voice and data packets with increasing traffic loads. As a conclusion, we have found that the average data packets will be reduced proportionally to the increase of the terminal movement. Therefore, it seems that high mobility may appreciably affect data performance with an increase in traffic loads. However, it is worthwhile to investigate the overall effect of mobility if the additional signalling loads are taken into account. More importantly, a flexible priority control policy and the reduction of voice coding rate are proposed to enhance data performance without significantly degrading voice services.

4.3 Performance of R-TDMA in Integrated Services

Multi-service is seen as the heart of future multimedia studies. As the number of users increases, the problem of channel efficiency in integrated services becomes salient. The main purpose of the PRMA protocol is to increase channel efficiency and system capacity while operating in a wireless full packet mode. The slot-limited nature of PRMA systems requires that channel access strategies and packet queuing performance need to be further investigated. Based on the MAMs developed in the previous sections, the performance of PRMA in integrated queuing systems is analysed in this section. It is known that the maximum speech packet time D_{max} can directly determine speech quality, while there are no severe limits on the maximum queuing time of data packets [*Frullo94*]. However, from users point of view, an unbounded packet delay is undesirable. Moreover, because the study of packet waiting time can be used to determine the retransmission of data packets and their effect on the provision of best effort QoS, the investigation of packet waiting time distribution is of importance. As a benchmark for the evaluation of data packets, the maximum allowable data packet queuing time is numerically derived. It shows that, firstly, the behaviour of data performance is subject to voice load condition. Secondly, the waiting time distribution is largely dependent upon the ratio of call duration to packet message length and input traffic loads. Moreover, a prioritised queuing policy is proposed to control packet congestion in overload regions without degrading voice packet services.

This section is organised as follows. Subsection §4.3.1 contains the introduction. System operation and some design problems that were encountered are presented in Subsection §4.3.2. Subsequently, the performance of a voice system is discussed in Subsection §4.3.3. Channel access strategies are then investigated in Subsection §4.3.4. Accordingly, queuing performance is analysed in Subsection §4.3.5. Finally, a summary appears at the end.

4.3.1 Introduction

Integrated wireless systems are expected to unify diverse communication sources and deliver information to designated end users. The fully packetized mechanism is seen as a suitable candidate, which is able to harmonise the wired backbone Broadband ISDN (B-ISDN) or Broadband ATM (B-ATM) networks [*Goodm89*]. On the other side, although the present Local Area Networks (LANs) and Metropolitan Area Networks (MANs) can

provide high rate data transmissions, they are unable to meet the requirements of terminal mobility as well as call handoff. Meanwhile, Wireless LANs (W-LANs) can only provide high rate data services within a limited area. Similarly, although ATM is another strong candidate for high rate data transmissions, it needs to be modified into Wireless ATM (W-ATM). Therefore, it is essential to consider the performance of fully packet-based wireless networks.

The conventional centralised control structure means that all control functions, such as terminal registration, location determination and channel allocation, are concentrated in a controller centre [*Meier94*]. However, if traffic loads become high in a small cell, the control centre is unable to afford the heavy burden of traffic signals. As the demand for wireless integrated services increases, this conventional architecture is found to be inadequate to meet such emerging demand. Instead, the proposed distributed control in packet-based networks allows small processing units to handle these tasks individually. Meanwhile, fast resource assignment and fast handoff can be realised with greater ease.

The PRMA protocol, employing packet transmission technique, was originally proposed by Goodman in the early 1990's [*Goodm89*], [*Goodm90*], and [*Goodm91*]. The PRMA system actually combines random access with time division access. The reason for using the reservation mechanism is to achieve high throughput and high channel efficiency. With the increasing traffic demands, it is of vital importance to exploit spectrum efficiency gains. Specifically, the efficiency exploitation depends on many techniques, such as channel coding, modulation, and cell layout, etc. As far as the statistics of traffic sources are concerned, a means to improve spectral efficiency without a significant increase in terminal complexity is to exploit the on and off characteristics within a conversation. After taking the idle periods into account, channel utilisation becomes more efficient during packet access. This is usually termed as the gain from statistical multiplexing. In particular, once the technique of low fill-in or hangover duration is used, the multiplexing gain is expected to become better in terms of data delay [*Sriram83*].

The PRMA systems are seen differently from the conventional TDMA systems, in which a channel is dedicated for a whole call until the termination of a conversation. Because of the long set-up time and the dedicated channel for the whole call, the circuit mode TDMA systems are seen as inefficient for the transmission of bursty data traffic [*Mitrou93*]. In contrast to the TDMA systems, the talkspurt in the PRMA system is only

allowed to occupy the dedicated channel. During the silent period, the dedicated channel has to be given up for the transmission of data packets. The talkspurt and silent states can be detected by a voice activity detector (VAD) [*Brady69*]. For instance, an active packetized voice call is classified as a talkspurt and a silent period shown as in [Figure 4.10].

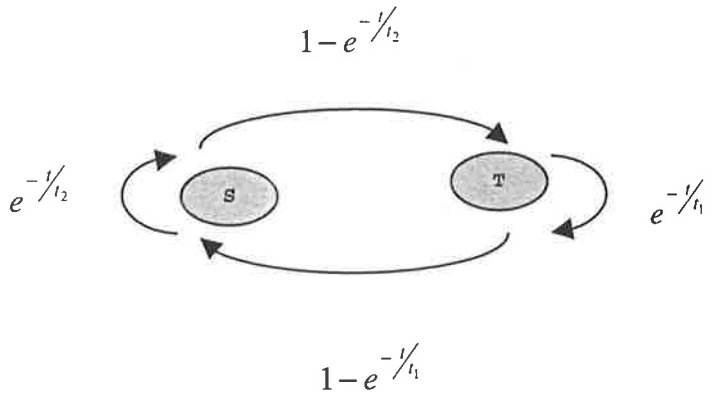


Figure 4. 10: The two-state speech model

Once a fast detector is used, a more precise multi-state can be considered. However, for simplicity, a model with two simple changeover states following negative exponential distributions is used. Similar assumptions have been adopted in [*Goodm91*]. If we use t_1 (seconds) to denote the mean duration of a talkspurt and t_2 (seconds) to represent the mean duration of a silent gap, the time proportion of a talkspurt in an assigned time slot can be expressed by $P_{BUSY} = t_1/(t_1 + t_2)$. This is commonly known as the voice activity factor (VAF) α_v. Similarly, the probability of a silent period in a dedicated channel can be easily shown to be $P_{IDLE} = t_2/(t_1 + t_2)$. Thus the transition probability P_{TS}, which transits from a talkspurt to a silent period, becomes $P_{TS} = 1 - e^{(-1/t_1)}$. Correspondingly, the transition probability P_{ST} from a silent to a talkspurt is $P_{ST} = 1 - e^{(-1/t_2)}$.

It is known that spectrum efficiency is an important issue for integrated services. The PRMA efficiency is defined as the number of conversions per channel [*Goodm91*]. In essence, the efficiency of PRMA depends upon the statistics of speech talkspurt and silence. As an example, without considering the overhead, if the channel rate is r_c kb/s and the source rate r_s kb/s, the equivalent number of TDMA channels per PRMA frame

simply becomes: $N_c = r_c / r_s$. Compared with TDMA systems, the PRMA efficiency in terms of conversations per channel can be expressed by:

$$\eta_{0.01} = \frac{M_{0.01}}{N_c} = \frac{M_{0.01} r_s}{r_c} \tag{4.26}$$

where $M_{0.01}$ represents the maximum number of conversations constrained by speech packet dropping probability, e.g., $P_{drop} \leq 0.01$.

As a consequence, the upper bound on the PRMA efficiency becomes: $\eta_{0.01} = \frac{1}{\alpha_v} = 2.35$. This means that PRMA can achieve the result of $1 < \eta_{0.01} \leq 2.35$. Goodman [*Goodm91*] shows that PRMA achieves the efficiency of 1.64 conversations per channel. In addition, Bout shows that the multiplexing gain for packet voice is achievable from 1.8 to 2.3 by using some advance techniques [*Bout97*]. Therefore, PRMA efficiency is found to be better than the conventional multiple access techniques, such as the conventional TDMA or FDMA systems.

As mentioned in Chapter 2, many studies have been conducted for PRMA systems [*Wu94*], [*Nanda94*] and [*Qi96*]. In contrast, our analytical study is different. A MAM is developed to analyse the queuing performance at both call level and packet level for the PRMA integrated services. Subsequently, different channel access strategies are then compared. Consequently, the waiting time distribution is derived and a priority scheme is proposed to ease the congestion condition.

4.3.2 The Operation and Design of the PRMA Protocol

The operation of the PRMA system can be described as follows. A speech traffic source is assumed to generate exactly one packet per frame. Each packet comprises an information field for the message, a control field for the address, synchronisation flags and parity check sequences. Slots are further classified into available slots and reserved slots. During channel random access, the first voice packet (FVP) competes against data packets (DP) for gaining access to the traffic channels with probability p_{permv} [*Nanda94*]. The reason of using a permission probability for the previous colliding users to contend for the next time slot is to prevent consistent collision [*Hanzo94*]. In order to contend for the available channels, two criteria must be satisfied. The first one is that there must exist

available slots. The next one is that a terminal has to hold a permit. If the FVP is successful, speech terminals can reserve a sequence of slots in the succeeding frames until the termination of the call. In practice, if the colliding packets arrive at a BS with different signal levels, the BS can be designed to detect packets according to the strongest signal. This is usually regarded as capture effect. Similarly, data terminals can compete for the available slots with another permission probability p_{ermd}. A scheme, in which a successful data terminal has no right to reserve slots for data packets, is proposed by Nanda [*Nanda94*]. This has been modified and leads to another scheme called Integrated PRMA (IPRMA) by Wong [*Wong92*] and [*Wong93a*], in which data packets are allowed to use the reservation mechanism.

Since the PRMA performance is sensitive to permission probability, the permission probability is an important design parameter. If the permission probability chosen is too low, mobile users have to wait a long time and eventually drop the packets when a timeout limit is reached. This can lead to low throughput. Generally, the tolerable timeout limit of speech packets relies on several factors, such as encoding, decoding, interleaving, and propagation delay, etc. On the other hand, if the permission probability chosen is too high, the subsequent congestion can result in high packet dropping probability due to collision. Therefore the choice for an optimal permission probability is crucial in order to provide a required QoS. Specifically, the occurrence of permission events is seen differently from terminal to terminal, or statistically independent from each other. Voice terminals are usually assigned with higher permission probability than data terminals.

In addition, a voice terminal is said to be in an active-state if it initiates a call and then starts a conversation. Conversely, the terminal is in an inactive-state if the terminal is in the thinking or standby states. This assumption is based upon a call level only. We use l to denote the number of users being idle states, which can be illustrated in [Figure 4.11].

According to the $M/M/m/m$ queue, we can have the probability P_l for l idle voice users as:

$$P_l = \binom{M}{l}\left(\frac{\rho_v}{1+\rho_v}\right)^l\left(1-\frac{\rho_v}{1+\rho_v}\right)^{M-l} \tag{4.27}$$

where the offered load is $\rho_v = \lambda_v / \mu_v$ and M is the maximum number of channels.

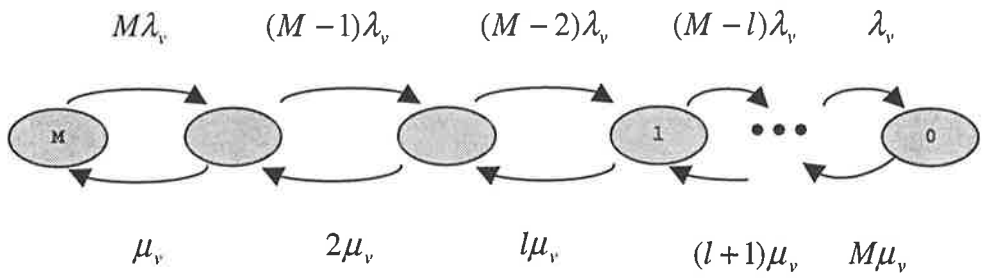


Figure 4. 11: The state transition diagram for voice calls

The result for idle users can be shown in [Figure 4.12]. We find that the analytical result is in good agreement with the simulation result. In particular, we also observe that the probability P_l is solely determined by the offered loads ρ_v rather than the reserved states or contention states.

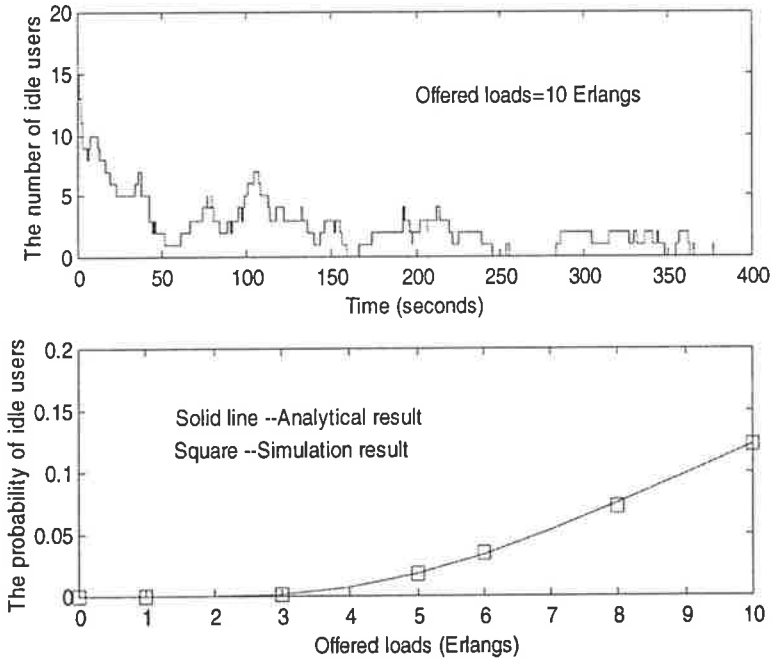


Figure 4. 12: The voice idle users in PRMA

Since speech is packetized, packet dropping probability at a packet level needs to be taken into account. Therefore, another important guideline in design is to consider the QoS at both call level and packet level. Packet dropping probability is defined as the probability

that a PRMA terminal discards the packets that wait longer than a threshold $D_{\max}$ while the redundant packets are stored in buffers by the FIFO rule:

$$P_{drop} = \Pr\{D \geq D_{\max}\} \tag{4.28}$$

Apparently, if the limit threshold $D_{\max}$ is equal to zero, the packet dropping probability is equivalent to the actual waiting time distribution in the queue. Namely,

$$P_{drop}\Big|_{D_{\max}=0} = \Pr\{D \geq 0\} \tag{4.29}$$

In particular, the normalised throughput η_{TP} is exactly equivalent to the equilibrium reserved time slot probability: $\eta_{TP} = r$ [*Nanda91*].

For example, the packet dropping probability is usually set to be 1% because voice quality will become unsatisfactory if it becomes higher [*Goodm90*]. As a consequence, the maximum number of simultaneous conversations is eventually restricted. Moreover, because voice packets require prompt transmission, long delayed packets will be discarded if waiting time exceeds a threshold. Therefore the allowable delay becomes an important design parameter in the PRMA system design. This leads to the investigation of data packet delay in Subsection §4.3.5.

Finally, another important characteristic in a wireless network is propagation complications. In general, a mobile radio channel is characterised according to the statistical propagation models, while the channel parameters are estimated by stochastic variables. As far as a macrocell is concerned, the small-scale effect of the short-term fluctuations is usually caused by multipath fading [*Frullo94*]. Multipath can lead to a rapid fluctuation of the signal amplitude and phase for a fast moving terminal. For a medium-scale effect, shadowing fading can be represented by a lognormal distribution, while a free space loss is used to represent a large-scale effect [*Linnar93*]. In contrast, as far as a microcell is concerned, the free space loss can be used to approximate the fading of line-of-sight (LOS). Street corner needs to be taken into account in the case of non-LOS propagation [*Frullo94*]. Wong [*Wong93b*] shows that the deteriorated channel condition has a significant effect on the packet dropping rate. In conclusion, PRMA exhibits robust performance only in the condition of least channel errors.

4.3.3 The Performance of a Voice System

Taking a slow detector into account, the state space Ω of a PRMA system can be simply denoted as:

$$\Omega = \{M_{sil}, M_{con}, R_0, R_1, \cdots, R_{N-1}\} \tag{4.30}$$

where M_{sil} denotes the number of silent terminals, M_{con} for the number of contending terminals and R_i for the number of terminals in the *ith* reserved state [Figure 4.13].

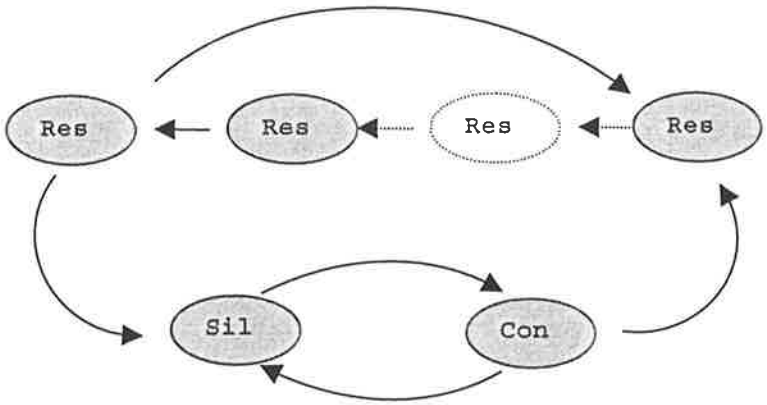


Figure 4. 13: The state space of speech subsystem model

As an example, if $M_{(s)} = 2$ and $N_{(s)} = 20$, the total number of state spaces can be as large as $2^{N_{(s)}} M_{(s)}^{\ 2} = 4{,}194{,}304$. Obviously, this prohibits us from using the exact Markov analysis [*Nanda91*]. Alternatively, approximate techniques can be used.

The use of PRMA and IPRMA can lead to two problems. One is that packet collision prevents us from fully exploiting channel utilisation. The other is the congestion caused by the instability under heavy load conditions [*Ren98*]. Specifically, the cause of the unstable problem is that the outflow from the contending state decreases, while there is an increased inflow towards the contending state. One of the solutions for preventing the instability is to use a timeout model for speech packets. After a certain fixed period, the speech packets will be automatically dropped. The non-linearity of the maximum number of simultaneous conversations leading to instability can be found in [Figure 4.14].

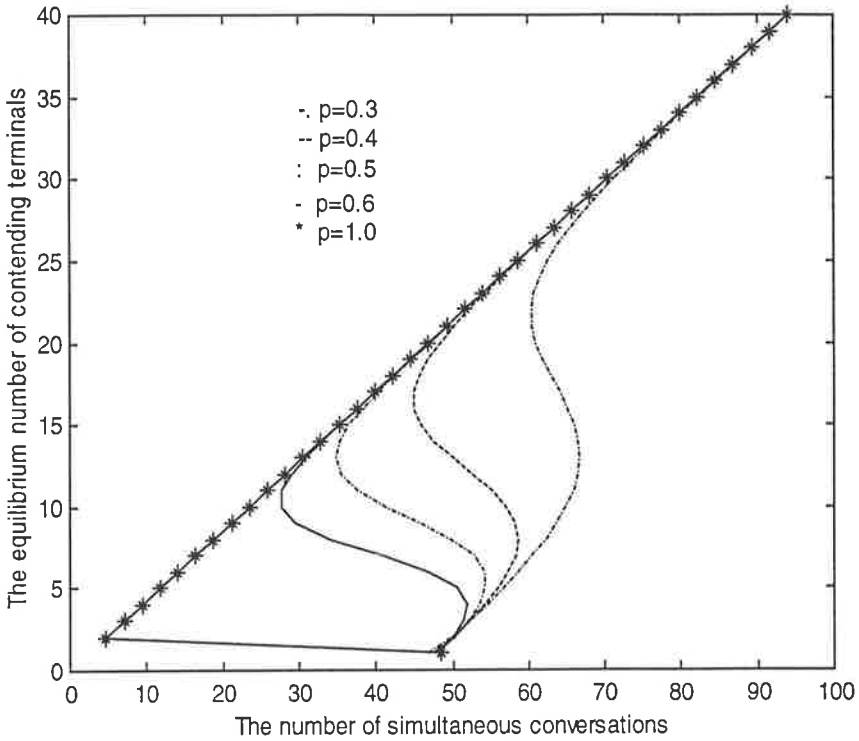


Figure 4. 14: The system stability with different permission probabilities

From the figure, we observe that, firstly, there exist multiple equilibrium points under high traffic loads. The PRMA protocol will lead to instability under high load conditions. Secondly, the larger permission probability p_{erm} is more likely to cause the system to become unstable or congested for a certain number of simultaneous conversations. As the permission probability increases, the critical point for stability of PRMA logarithmically decreases. In particular, the system has a convergent limit while the permission probability is equal to one. Therefore the instability problem becomes so important that multiple equilibrium points must be avoided in system designs.

4.3.4 Channel Access Strategies in the NC-PRMA Protocol

In the integrated services of PRMA, speech terminals discard voice packets that encounter excessive delay [*Goodm89*]. In contrast, data packets, which are denied with services, can be scheduled in a queue. The momentary states may become overloaded while the transient service rate is smaller than the transient arrival rate. If congestion occurs in the queue, this will cause long delay and the analysis of the static performance

becomes invalid. As a consequence, the long waiting time can hamper network performance. In order to overcome the instability problem and improve spectrum efficiency, a modified PRMA protocol called non-collision PRMA (NC-PRMA) is proposed in [*Wen95a*] and [*Ren98*]. Because the NC-PRMA protocol adopts a contentionless mode for channel access, it can efficiently avoid the instability caused by the contention and have better spectrum utilisation efficiency than the conventional PRMA protocol.

In particular, NC-PRMA protocol is different from the IPRMA proposed by Wong [*Wong93a*]. Because high data packet volume can easily exacerbate the speech dropping probability, Wong suggests that a certain number of idle slots are reserved for speech packets in order to enhance the chance access for speech packets under high traffic load condition. The problem arising in the IPRMA is that the efficiency of the idle slots can not always be guaranteed while the speech loads are low and the data loads are high. This requires that the sliding window size needs to be constantly changed according to voice traffic.

The frame structure of NC-PRMA protocol can be described as in [Figure 4.15].

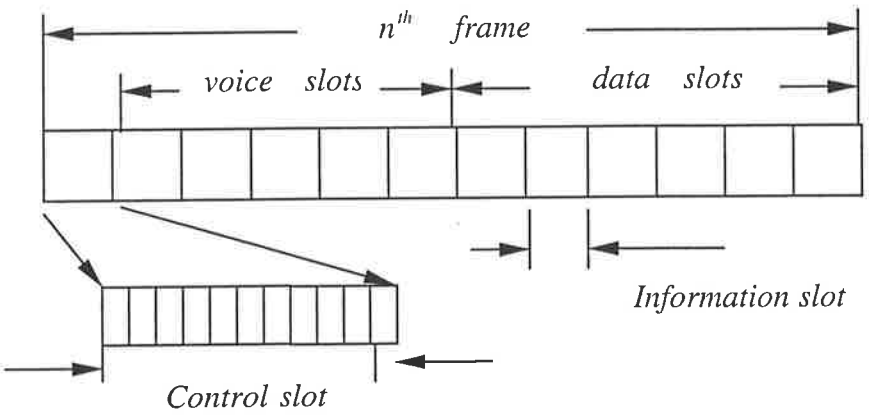


Figure 4. 15: The uplink structure of NC-PRMA protocol

Each frame consists of one control slot and N information slots. The control slot can be further divided into M_{ini} minislots, where M_{ini} is the maximum number of the simultaneous conversations. During the operation of NC-PRMA protocol, the terminal initiates a request packet to the BS by the control channel. After successfully receiving the request message, the BS assigns an information channel and control minislots to the MS.

The control minislots are used for slot assignments, power control and other control purposes. Intuitively, the NC-PRMA protocol structure is analogous to the Dynamic TDMA (D-TDMA) protocol proposed in [*Wilson93*]. However, they are applied to different systems.

As a comparison, throughput in PRMA and NC-PRMA is shown in [Figure 4.16].

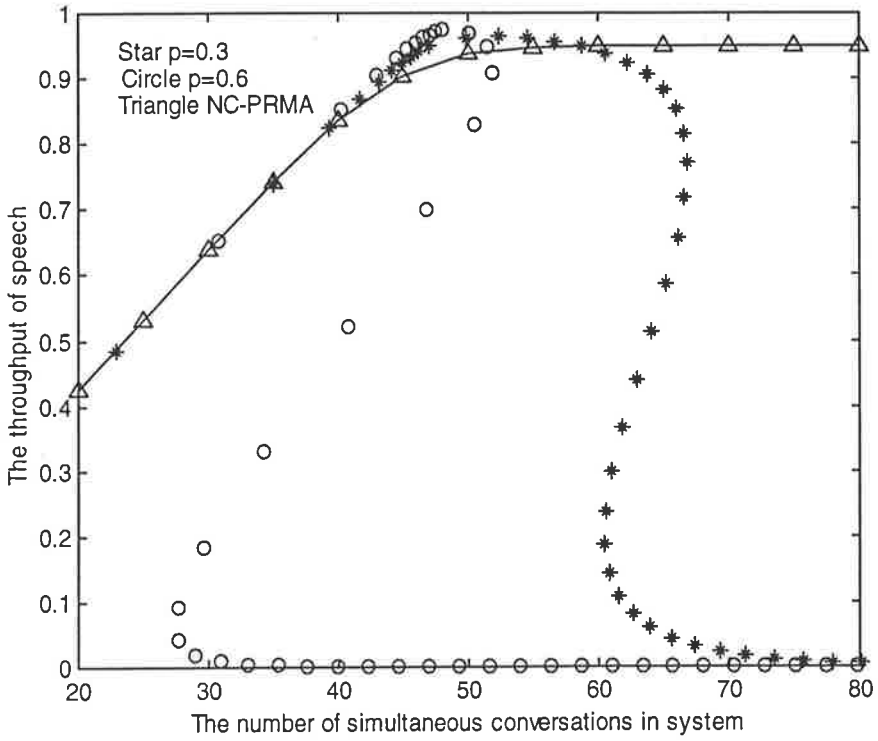


Figure 4. 16: The speech throughput in PRMA and NC-PRMA

We observe that NC-PRMA can have better performance than the conventional PRMA when the number of simultaneous conversations increases.

Here we use the parameters, which are similar to those used in [*Goodm90*]. That is, the talkspurt is assumed to be with a mean duration $t_1 = 1\,s$ and the silent period with a mean duration $t_2 = 1.35\,s$, then the VAF becomes $\alpha_v = \frac{t_1}{t_1 + t_2} = 0.43$. The channel transmission rate $r_c = 270$ kb/s operates in a 200 kHz bandwidth. The speech coding rate is $r_s = 16$ kb/s. The packet size is 552 bits with 472 bits for information and $H = 80$ bits for header. Therefore, a speech terminal, which generates 472 bits, needs the duration of

$T = 29.5$ ms per frame. Because the speech packet has a maximum delay of $D_{\max} = 32$ ms, the maximum integer number of slots that terminals can contend within one frame becomes:

$$N = \left\lfloor \frac{D_{\max} r_c}{r_s T + H} \right\rfloor \tag{4.31}$$

Therefore, we easily have $N = 15$ slots/frame. This is similar to a GSM-like environment.

As another example, if we assume that the same speech code rate is $r_s = 16$ kb/s and the channel transmission rate can operate as high as $r_c = 672$ kb/s, the number of the available contention slots becomes $N = 37$ slots/frame with the same speech delay $D_{\max} = 32$ ms. This example is similar to that of a DECT-like microcell environment.

In this study, we assume that there are N_v active calls in the system. Without considering the reservation states, the active talkspurts are confined to a simple binomial distribution if the number of talkspurts is less than the total N slots per frame:

$$B(N_v, \alpha_v) = \binom{N_v}{j} \alpha_v{}^{j} (1-\alpha_v)^{N_v - j} \tag{4.32}$$

Therefore the mean number of unused slots leftover by the talkspurts is:

$$f(N_v) = N - \alpha_v N_v (1 - P_{drop}) \tag{4.33}$$

The normalised mean number of the available slots per frame for data packets is:

$$\frac{f(N_v)}{N} = 1 - \frac{\alpha_v N_v (1 - P_{drop})}{N} \tag{4.34}$$

Obviously, the active talkspurts that are denied access to channels become lost if the talkspurts j are bigger than N slots per frame:

$$P_{drop}(N_v) = \frac{\sum_{j=N+1}^{N_v} (j - N) B(N_v, \alpha_v)}{\alpha_v N_v} \tag{4.35}$$

Therefore, it is easily to evaluate the normalised available slots per frame, which are left over by voice packets and can be used for the delivery of arrival data packets:

$$\frac{f(N_v)}{N} = 1 - \frac{\alpha_v N_v}{N} + \frac{\sum_{j=N+1}^{N_v} (j-N) B(N_v, \alpha_v)}{N} \tag{4.36}$$

Finally, we have the result as shown in [Figure 4.17].

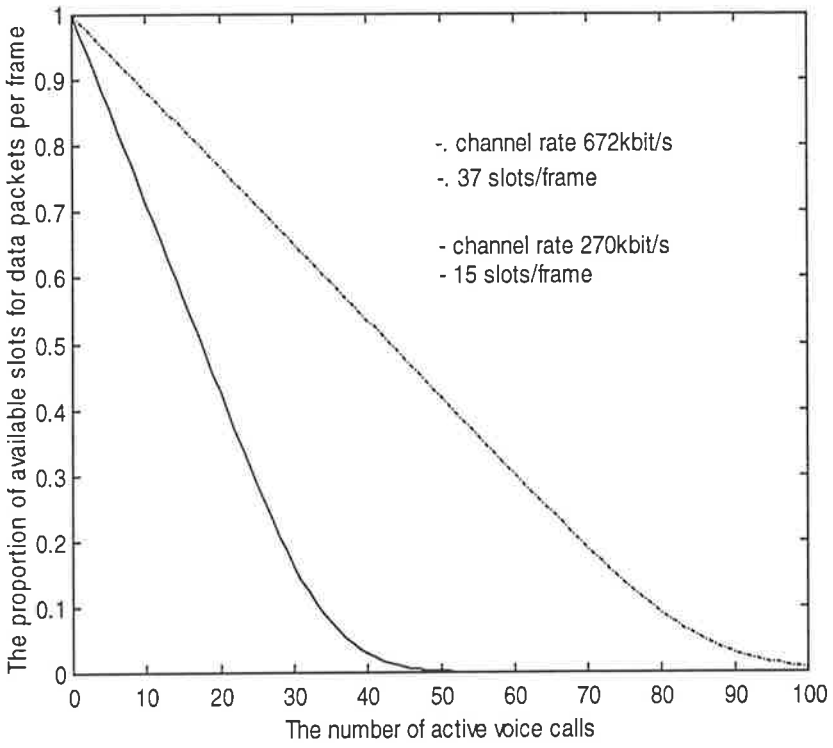


Figure 4. 17: The proportion of available slots for data packets per frame

From the figure, we observe that, although the available slots for data packets decrease according to the increasing active voice call users, the average empty time slots left over by voice packets is still as high as 43% while there are 20 active users in the system. This suggests that the empty capacity left over by the voice users can be further exploited for the delivery of data packets. The reason for the decrease in the proportion of available slots is that the increasing number of users causes more vigorous competition for idle slots. In particular, compared with a low channel rate, a high channel transmission rate tends to cause more leftover slots for data packets.

As in TDMA systems, call control can be used to improve PRMA system performance. Although some control policies have been extensively studied in the wired line ISDN networks [*Willa84*] and [*Kraim86*], the control policies in packetized cellular networks need to be further addressed [*Calleg95*]. In this study, two channel access control policies are used. The first one is to ensure that the maximum number of voice terminals gains access to channels while meeting the required voice packet dropping rate, e.g., 1%. The second one is to force data terminals to share all empty slots leftover by the voice packets in a round robin manner.

As mentioned before, the central control problem in integrated services is to the need to share limited resources efficiently and dynamically among different services. When considering traffic control, channel access strategies are the focus in this study. The fixed boundary, movable boundary and no boundary control schemes are adopted as the candidates for slot allocations.

Let $P(i, j)$ represent j active voice calls and i data calls in progress in the system. The state space can be shown by a 2-D QBD process as described in Section §4.2.

Model Assumptions:

1) Voice arrivals are assumed to follow a Poisson process with a mean arrival λ_v. The successful voice packets can be served according to the reservation protocol as described above until the conversation terminates. Voice packets are allowed to have priority over data packets with a service rate μ_v.

2) The arrivals of data packets λ_d are approximately assumed to follow a Poisson process as well. Data packets are served with a rate μ_d according to the number of idle slots left over by voice packets.

3) When channels are full, data packets are able to wait. The queuing length of buffers is assumed to be infinite, where the queuing data packets follow the FIFO discipline.

4) The transmission channel is assumed to be error free. RF signals are assumed to be at a satisfactory level. Therefore, the channel access conflict becomes the only

source of blocking and delay. For simplicity, we assume that each data call is justified on the basis of one slot per frame. Similar arguments can be found in [*Wen95b*] and [*Wong93b*]. Meanwhile, both capture effect and transmission delay are not included in this study.

Moreover, because the $(m-1)th$ frame and $(m+1)th$ are assumed to be homogenous in the steady state, only the mth frame is considered. It is worth to noting that this assumption becomes untenable due to the correlative properties of each frame under the condition of dynamic analysis.

In the following, the channel access strategies in PRMA systems are examined. Although different channel allocation schemes are investigated in [*Schwartz87*], our study is different because it is based upon the fully packetized PRMA system. The main difference is that resource allocation for data packets must take both a talkspurt level and a call level into account.

4.a) Fixed Boundary Schemes (FBS)

A GSM-like environment as described early is assumed in the following study. The voice call arrival rate is λ_v and the service rate μ_v per frame. Channel access in fixed boundary schemes (FBS) means that a fixed number of K slots per frame are permanently assigned to the voice packets and the rest slots for data packets. Under such an arrangement, voice packets can occupy up to a maximum of K packets and data packets are confined to a maximum of $(N-K)$ packets only.

Since voice performance is not affected by data packets, voice blocking probability is easy to derive from the conventional Erlang formula. If the deterministic service time is used for data packets, a lower bound for data throughput is expected. On the other side, if the general service time is assumed, an upper bound for data packets is obtained [*Zhang90*]. In this study, an exponential service distribution model for data packets is assumed.

Hence the data offered load simply becomes:

$$\rho_d = \frac{\lambda_d}{(N-K)\mu_d} \tag{4.37}$$

The mean number of data packets in the system can be represented by:

$$E(N_d) = \sum_{i=0}^{\infty} iP_d(i) = \frac{\rho_d}{1-\rho_d} = \frac{\lambda_d}{(N-K)\mu_d - \lambda_d} \tag{4.38}$$

Therefore the mean number of data packets, which are independent of the number of voice packets, can be shown as in [Figure 4.18].

From the figure, we observe that the more slots that are preserved for voice packets, the less offered data loads that can be accommodated. This is easy to understand because the fewer slots are left over by voice packets. The problem arising in the FBS is that a number of slots, which are dedicated to voice calls, remain unused if the voice offered loads are not high. This can cause long delay for data packets and hence a lack of channel efficiency. A solution for this situation is to allow data packets to borrow the spare slots left over by voice calls temporarily. Therefore, another flexible scheme is described as follows.

4.b) Movable Boundary Scheme (MBS)

Channel access in the movable boundary scheme (MBS) is to assign K slots to voice calls and $(N-K)$ slots for data packets [*Schwartz87*]. Under the condition of a low voice load, data can exploit the leftover packets by voice calls. As a matter of fact, compared with the FBS, the MBS can provide apparent improvement for data throughput and waiting time. Thus this scheme can lead to the increase of channel utilisation. Because voice quality is constrained by packet dropping rate, data packets are only allowed to exploit the empty slots leftover by voice packets due to voice momentary variations. While the packet dropping rate is below the predefined 1%, the boundary is allowed to move to N_k and thus data packets can have the $(N-N_k)$ slots. Therefore, in order to meet the requirement of having maximum throughput for voice packets, the number N_k can be taken as the equivalent slots, which are occupied by k voice calls.

For the MBS, the maximum number x_K of voice calls that can be admitted into the system will correspond to the use of K slots. The relationship of the number of voice calls x_j corresponding to the number of occupied time slot N_x can be shown as in [*Wen95b*]:

$$x_j = \left\lfloor \frac{M_{0.01}}{N} N_x (1 - \frac{N - N_x}{3N}) \right\rfloor \tag{4.39}$$

where $M_{0.01}$ is the maximum number of voice calls that can maintain the required packet dropping rate 1%. N_x denotes the equivalent number of slots occupied by x_j voice calls, while the packet dropping rate is no more than 1%. The parameter $M_{0.01} = 22$ is used in this study [*Goodm90*].

Thus the blocking probability for voice calls due to all the pre-defined K slots being occupied by x_K voice calls becomes:

$$P_v(x)\Big|_{x=x_K} = \frac{\rho_v^{x_K} / x_K!}{\sum_{j=0}^{x_K} \frac{\rho_v^j}{j!}} \tag{4.40}$$

where $\rho_v = \lambda_v / \mu_v$.

As a result, the blocking probability at a voice call level can be obtained [Figure 4.19]. We easily find that the optimal boundary number in the MBS is critical because it can determine the voice blocking probability and then the voice packet dropping rate. In particular, the higher K can lead to the lower blocking probability for voice calls.

Conditional on the voice packets N_x, the equilibrium probability of data packets becomes a simple $M/M/1$ expression:

$$P_{D|N_x} = P_{0|N_x} \prod_{i=0}^{l-1} \left(\frac{\lambda_d}{(N - N_x)\mu_d} \right) \tag{4.41}$$

$$P_{D|N_x}(i) = (1 - \rho_d)\rho_d^{\,i} \quad where \quad \rho_d = \frac{\lambda_d}{(N - N_x)\mu_d} \tag{4.42}$$

Therefore the unconditional data packets steady-state probability can be shown by:

$$P_D(i) = \sum_{N_x=0}^{x_k} P_{D|N_x}(i) P_v(x\ ) \tag{4.43}$$

Hence, the conditional mean number of data packets in the system can be easily written as:

$$E(N_{d|x}) = \sum_{i=0}^{\infty} iP_{D|N_x}(i) = \frac{\lambda_d}{(N - N_x)\mu_d - \lambda_d} \tag{4.44}$$

Therefore, the unconditional mean data packets in the system is the summation over all voice packets:

$$E[N_d] = E[E[N_{d|x}]P_v(x)] = \sum_{x=0}^{x_k} \frac{\lambda_d}{(N - N_x)\mu_d - \lambda_d} P_v(x) \tag{4.45}$$

In the end, we have the result as shown in [Figure 4.18], where the boundary in the MBS is assumed to be $K = 10$.

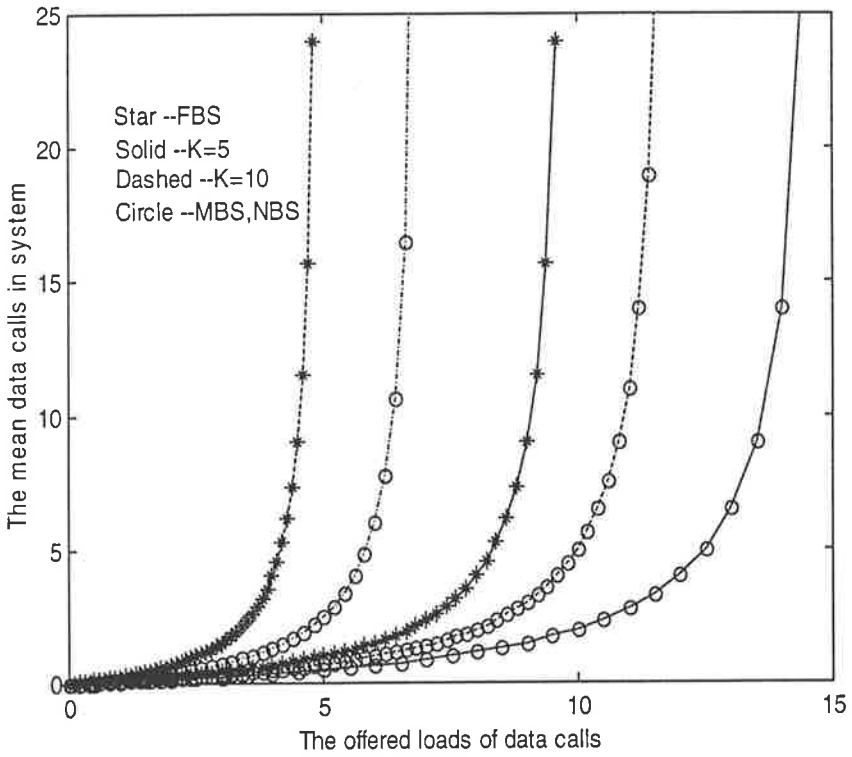


Figure 4. 18: Conditional average data calls in MBS

In contrast to the results with using the FBS, the results in the MBS show that higher number of admitted voice calls will lead to a lower average number of conditional data calls and hence lower data carried loads as well. Therefore, the mean number of data calls

is significantly affected by the number of voice calls. By using Little's law, the average delay of data packets can be easily derived from the above formula as well.

Finally, we observe that data packets in the MBS have to experience long delay if data call offered loads (ρ_d) become high. It is evident that data call performance becomes much worse than its average when the voice packets attain maximum numbers [*Willi84*]. As a result, the excessively long delayed data packets will be dropped if the waiting period has timed out. Apparently, the backlog phenomenon is undesirable to data users who require timing delivery. Therefore, using the MBS, the choice of optimal boundary is crucial. An improper boundary may cause high voice packet dropping rate if the voice arrivals exceed a certain predicted load. The solution to this problem can be attained by relaxing the boundary constraint for voice packets, which leads to a more flexible scheme introduced as follows.

4.c) Non-Boundary Scheme (NBS)

Channel access in the non-boundary scheme (NBS) infers that there is no explicit predefined boundary for voice calls. In other words, an arrival voice packet that finds channels full is allowed to preempt data packets if the data packets are in service. However, the speech packet dropping rate must be maintained at below 1% and the data packets can use up to a maximum of N slots. In the NBS, voice calls x_k are allowed to use up to N_x data packets, whereas data calls can share the leftover $(N - N_x)$ slots. The voice call blocking probability can be shown as in [Figure 4.19].

Generally, voice call blocking probability must be limited to a certain threshold, e.g., 2%. Not surprisingly, we observe that voice call blocking probability in the NBS performs better than in the MBS. The reason for this is that the NBS can always yield more channels for voice calls than the MBS. Therefore, the opportunity for the denial of voice calls in the NBS is found to be less than that in the MBS. In addition, the larger the boundary that is chosen in the MBS, the better the blocking performance for voice calls.

Besides blocking probability, another important parameter is the mean number of data packets in the system. The Markov process of data packets can be expressed by the $M/M/1$ queue because the conditional number of data packets depends upon the number of voice packets.

The conditional steady-state probability of data calls in the system can be expressed by:

$$P_{D|N_x}(i) = (1-\rho_d)\rho_d^{\,i} \quad where \quad \rho_d = \frac{\lambda_d}{(N-N_x)\mu_d} \tag{4.46}$$

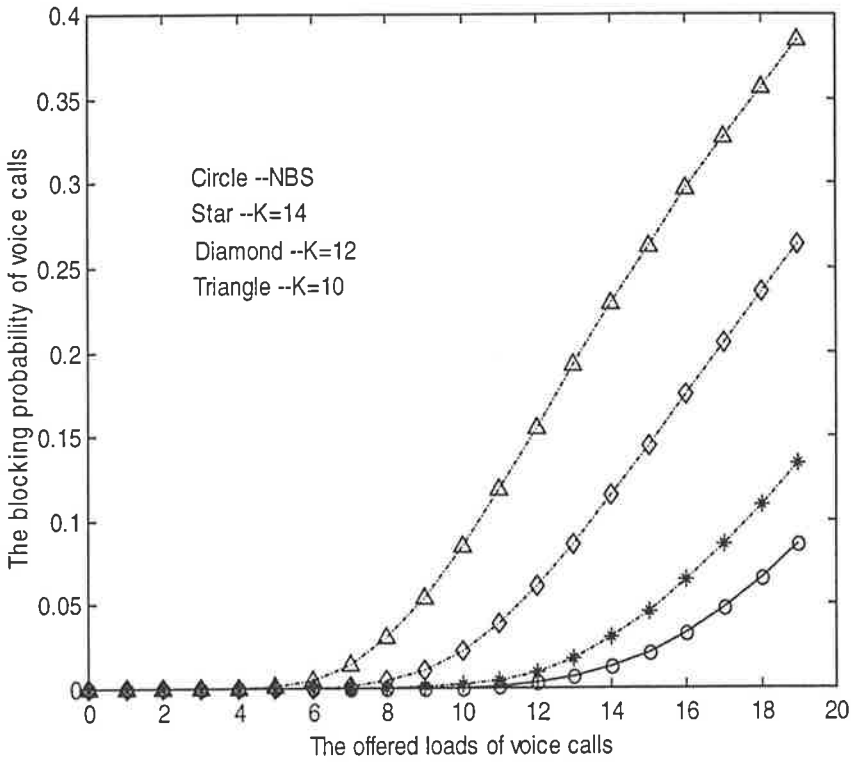


Figure 4. 19: Voice blocking probability in NBS and MBS

Consequently, the unconditional data packet steady-state probability in the system can be given by:

$$P_D(i) = \sum_{S_x=0}^{x_N} P_{D|N_x}(i) P_v(x) \tag{4.47}$$

Then the conditional mean number of data packets in the system can be easily written as:

$$E[N_{d|x}] = \sum_{i=0}^{\infty} i P_{D|N_x}(i) = \frac{\rho_d}{1-\rho_d} = \frac{\lambda_d/\mu_d}{(N-N_x)-\lambda_d/\mu_d} \tag{4.48}$$

The conditional mean number of data packets in the system is shown in [Figure 4.18]. Interestingly, we observe that the conditional mean number of data packets in NBS is the same as the case of MBS. The reason for this is that, given a certain data offered load ρ_d,

the data can exploit the available unused slots to the maximum $(N - N_x)$ under both schemes. Therefore there is no difference between them.

The unconditional mean number of data packets is the summation over all voice packets:

$$E[N_d] = E[E[N_{d|x}]P_v(x)] = \sum_{x=0}^{x_N} \frac{\lambda_d}{(N - N_x)\mu_d - \lambda_d} P_v(x) \tag{4.49}$$

We can have the unconditional mean data calls in the system as shown in [Figure 4.20].

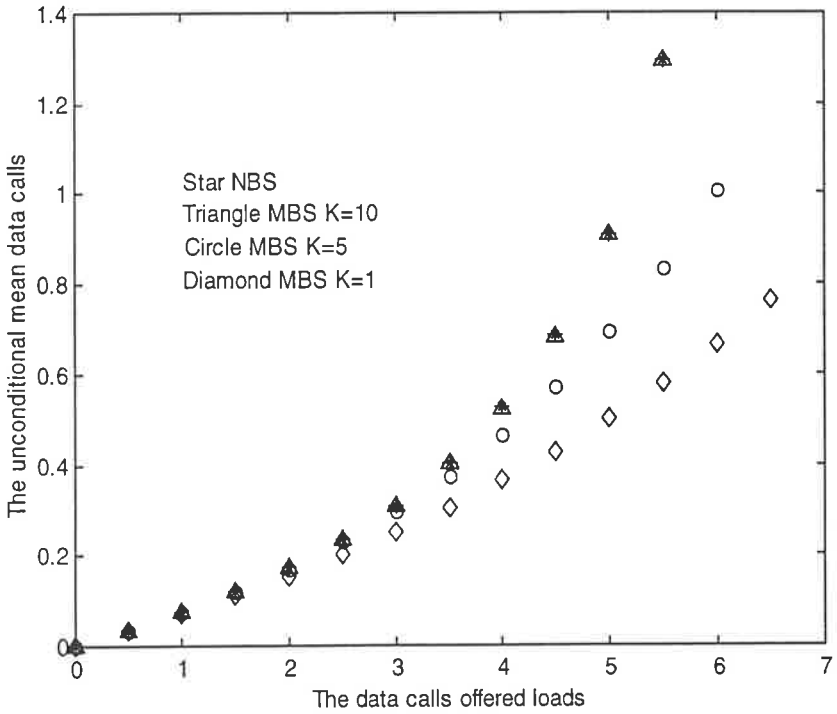


Figure 4. 20: Unconditional mean data calls in NBS

Surprisingly, from this figure, we observe the unconditional mean data calls in the NBS system are very close to that in the MBS when the boundary is large enough, e.g., $K = 10$ in this study. In order to verify this result, we have the probability of voice calls in these two different schemes shown as in [Table 4.5].

We find that there is no significant difference between these two schemes for a large K. In other words, the mean data calls in these two schemes converge while the boundary is large enough. However, if the boundary K in the MBS is reduced, we also find that the

unconditional mean data in the system decreases because less data packets can be accommodated.

Table 4. 5: The probability of voice calls in different schemes

No. voice calls	Movable	No-boundary
0	1.50E-03	1.50E-03
1	9.80E-03	9.80E-03
2	3.20E-02	3.18E-02
3	6.93E-02	6.88E-02
4	1.13E-01	1.12E-01
5	1.46E-01	1.45E-01
6	1.59E-01	1.58E-01
7	1.47E-01	1.46E-01
8	1.20E-01	1.19E-01
9	8.64E-02	8.58E-02
10	5.62E-02	5.58E-02
11	3.32E-02	3.30E-02
12	1.80E-02	1.13E-02
13	9.00E-03	5.20E-03

Consequently, an important observation from this study is that the difference between the NBS and the MBS becomes negligible while the boundary number in the MBS is large enough. While the boundary is $K = 15$, the NBS and the MBS become exactly the same in this study. In other words, the result from the NBS is equivalent to the upper bound of the MBS, and with a lower bound while $K = 1$.

By using Little's law, the delay of data packets can be easily obtained:

$$E[N_d] = T_d \lambda_d \tag{4.50}$$

Then the average time of the data calls spent in the system becomes:

$$T_d = E[N_d]/\lambda_d = \frac{1}{\lambda_d}\left(\sum_{x=0}^{x_N}\frac{\lambda_d}{(N-N_x)\mu_d-\lambda_d}P_v(x)\right) \tag{4.51}$$

Obviously, the high loads of data calls will cause long delay. An interesting result is that the average time of data calls in the system is not equal to zero when the arrival load is equal to zero. This is because there still are some data calls being served in servers even when the arrival load is zero. In fact, this is consistent with the conventional $M/M/1$ queue result. Therefore, we can have:

$$T_{dQ|x} = T_d - \frac{1}{(N-N_x)\mu_d} \tag{4.52}$$

where $T_{dQ|x}$ denotes the average waiting time in buffers only when there are x voice users.

4.3.5 Queuing Performance in MPBS

In order to relieve the over-crowded data packets in buffers and improve the unstable queuing condition, a flexible priority admission policy called multiple priority-based scheme (MPBS) proposed in the last section can be used to hamper the accumulated behaviour of data packets. The basis of this policy is to maintain the short-range channel utilisation stability according to voice traffic variations. A detailed description can be found in Section §4.2. Because this proposal aims to classify the ranking of data users and then restrain the low priority class, it can increase fairness in resource allocations.

The process is assumed to have the state space $\Omega = \{(i,j): 0 \le j \le x_N, i \ge 0\}$, where the index j denotes the state of the voice calls and i represents the number of data calls in the system. Because voice calls are allowed to remove the preoccupying data calls, voice call blocking probability is easy to derive from the standard formula:

$$P_{Bv} = \frac{\rho^{x_N}/x_N!}{\sum_{j=0}^{x_N}\frac{\rho_v^{\,j}}{j!}} \tag{4.53}$$

Note that the blocking probability of voice calls in the MPBS is the same as in the previous NBS. Therefore, voice performance becomes intact with the enforcement of this priority control scheme. For the equilibrium probability of data packets, it is somewhat more complex. Here we can use the results developed in Sections §4.1 and §4.2.

The k classes prioritised data packet arrivals in the uniform admission regions are partitioned as:

$$\lambda_d = \begin{cases} p_k(j)\lambda_d & 0 \le j < \left\lfloor \frac{x_N}{k} \right\rfloor \\ p_{k-1}(j)\lambda_d & \left\lfloor \frac{x_N}{k} \right\rfloor \le j < \left\lfloor \frac{2x_N}{k} \right\rfloor \\ \cdots \quad \cdots & \\ p_1(j)\lambda_d & (k-1)\left\lfloor \frac{x_N}{k} \right\rfloor \le j \end{cases} \tag{4.54}$$

If we choose the priority from high to low, we can have this inequality, $p_1(j) \le p_2(j) \le \cdots \le p_k(j)$. In addition, if we denote the level as $l(i)$ and the phase of a state (j,i) as j, the state space becomes $\Omega = \{(j,i) : i \ge 0, 1 \le j \le x_N\}$. Then this process can be represented by the skip-free property. That is, upon starting from level $l(i)$, the Markov chain can only move to the $l(i-1)$ or $l(i+1)$ in a one step jump. Apparently, we can obtain: $\Omega = \bigcup_{i=0}^{\infty} l(i)$. The details of the analysis can be found in [*Appendix1*].

Due to the block tridiagonal structure of the infinitesimal generator, this is actually a QBD process with a simple boundary condition. QBD processes are the generalisation of the simple birth-death process. The early studies for this process date back to [*Evans67*] and [*Wallace69*]. Later on, this work is continued in [*Neuts81*]. As mentioned in Section §4.1, the salient problem is that the computation of the matrix polynomial requires excessively lengthy time and large memory for a practical complex problem, especially under heavy traffic loads. Therefore, this problem motivated much interest in these areas [*Dailge91*] and [*Latou93*]. The first approach is called the transform method [*Dailge91*]. It claims that this method can efficiently determine the rate matrix under all traffic load conditions. The second approach is conducted by Latouche [*Latou93*] and is adopted for our study. The detail of the algorithm can be found in both Section §4.1 and Section §4.2.

Once R is known, we can obtain the boundary probability $\pi_{\infty}^{(0)}$ as below:

$$\pi_{\infty}^{(0)} B_0 + \pi_{\infty}^{(1)} B_1 = 0$$

$$\pi_{\infty}^{(0)} \sum_{i=0}^{\infty} R^i e = 1 \tag{4.55}$$

or equivalently,

$$\pi_{\infty}^{(0)} B_0 + \pi_{\infty}^{(1)} B_1 = 0$$

$$\pi_{\infty}^{(0)} (I - R)^{-1} e = 1 \tag{4.56}$$

By using the orthogonal-triangular decomposition, we can solve the boundary probability $\pi_{\infty}^{(0)}$ with the normalising equation. Once the boundary probability $\pi_{\infty}^{(0)}$ and the rate matrix R are known, the other steady probabilities can be derived.

In this study, we adopt the coefficients of the arrival rates of 1, 0.9, 0.8, 0.7, and 0.6 as the five-level priority scheme. This can be shown as in [Figure 4.21].

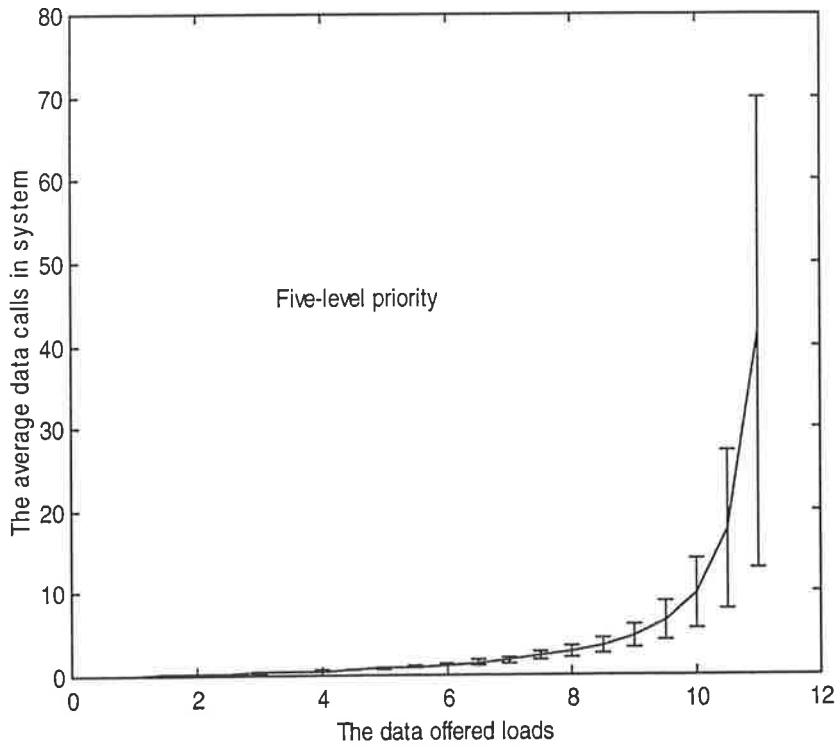


Figure 4. 21: The improvement of priority scheme

Under the five-level MPBS, while data offered loads change from 9 *Erlangs* to 10 *Erlangs* and then 11 *Erlangs*, the change of the mean number of data packets drastically increases from about 1.4, to 4.2 and 28.5 respectively. This shows that the optimal use of MPBS can improve the congestion of data packets in the queue, especially with the increase of offered loads. The mean number of data packets can be easily transformed into the expression of mean waiting time by using Little's law.

In addition, the study of data packet waiting time distribution in a queue is very important. This is because the delay of data packets needs to be estimated so that the queuing congestion can be avoided. Generally, there are two approaches to evaluate the waiting time distribution. The first uses an approximation method [*Bhat76*]. Although the approximation method can easily evaluate the data performance, the feasibility of this method tends to deteriorate when the ratio of the voice duration to the data duration $\alpha_w = \mu_d / \mu_v$ becomes larger. The second approach is regarded as the 'exact' solution [*Serres88*]. This method is used in our study. Note that the waiting time distribution is not investigated in either [*Ren98*] or [*Wen95a*].

The details of the waiting time distribution can be found in [*Appendix 1*]. Then the higher moment of waiting time distribution can be obtained. The *kth* moment of the long term waiting time is shown as:

$$W^{(k)} = k\theta^{-k} \sum_{n=0}^{\infty} d_n \frac{(n+k-1)!}{n!} \quad while \quad k \geq 1 \tag{4.57}$$

As a benchmark parameter for data packets, $D_{d\max}$ can be used to denote the maximum allowable queuing time for data packets. Thus

$$\Pr(W > D_{d\max}) = \sum_{n=0}^{\infty} d_n e^{-\theta D_{d\max}} \frac{(\theta D_{d\max})^n}{n!} \tag{4.58}$$

where $D_{d\max}$ is non-negative.

Finally, the logarithmic unconditional waiting time distribution versus the virtual waiting time can be shown as in [Figure 4.22], where α_w is the ratio of the voice duration to the data duration, e.g., $\alpha_w = 1$. From this result, we can observe that, although blocking probability is insensitive to α_w, the probabilities that a user has to wait longer than 200 ms

are 14%, 33%, 57% and 72% for the data arrival loads 2 Erlangs, 4 Erlangs, 6 Erlangs and 7 Erlangs, represented by circle, star, diamond and triangle lines in the figure respectively. Therefore, we conclude that, given a certain voice load, high data arrivals create a higher possibility of waiting.

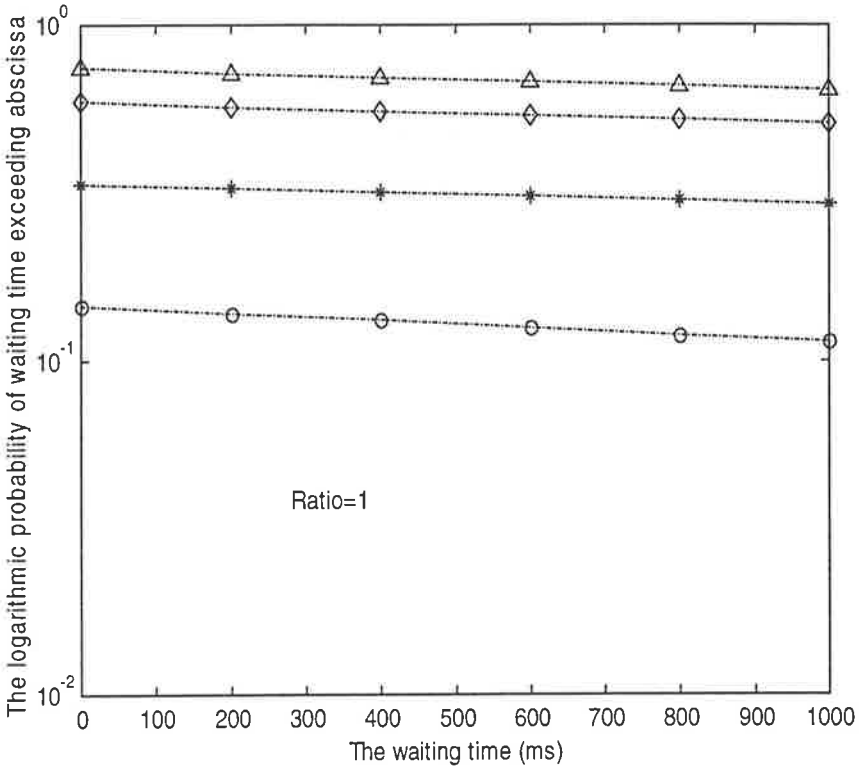


Figure 4. 22: The stationary waiting time distribution (1)

If $\alpha_w = 10$, we can have the waiting time distribution as shown in [Figure 4.23].

Compared with the figure shown before, we observe that the waiting time of data packets becomes shorter when the arrival rates remain unchanged but the service rates become faster. The longer allowable waiting time will lead to data packets having a reduced chance of dropping from the queue.

The waiting time distribution for a special case of $\alpha_w = 100$ is shown in [Figure 4.24].

From the figure [Figure24], we observe that, in the extreme case of $\alpha_w = 100$, data packets only have a slight chance of waiting when the duration ratio and the waiting time increase.

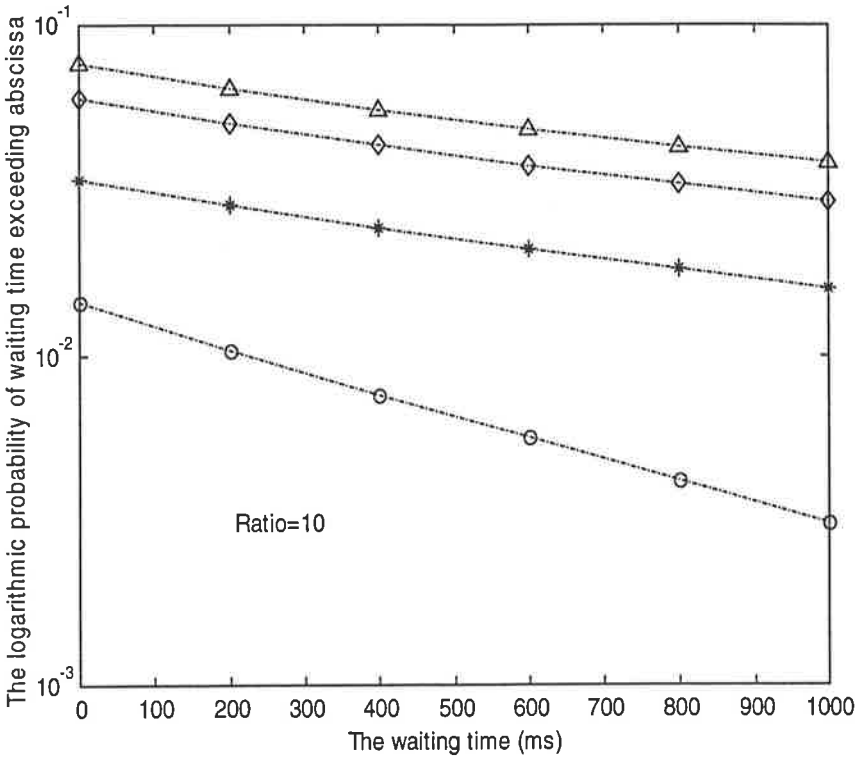


Figure 4. 23: The stationary waiting time distribution (2)

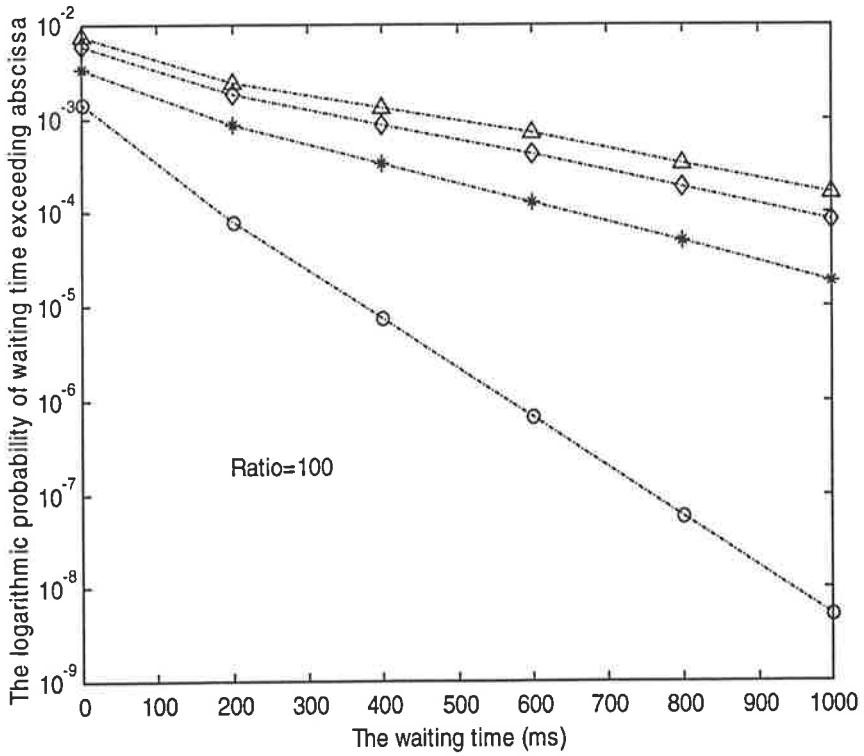


Figure 4. 24: The stationary waiting time distribution (3)

4.3.6 Summary

In order to provide a wide range of integrated services, the efficiency of resource allocation becomes salient. In this study, we elaborate the blocking-delay performance property for mixed traffic in a single-carrier multiple-client PRMA environment. We have investigated the performance of different channel access strategies, i.e., the FBS, the MBS as well as the NBS. Compared with the different channel access strategies, we conclude that NBS can provide the best performance of these channel access schemes. In addition, an efficient MAM is developed to evaluate mixed traffic performance. A prioritised best-fit MPBS is proposed so as to efficiently control the congestion of data packets in overload regions and effectively reduce the excessive packet delay. We show that this flexible scheme can reduce the mean number of data calls in the system without the loss of the high-priority class data packets. Because channel condition needs to be instantly measured, this scheme can lead to efficient utilisation of the bandwidth-limited channels without degrading voice services. Moreover, a benchmark parameter for data packets is numerically analysed. It shows that the waiting time distribution is largely dependent on the ratio of call duration to packet message length and input traffic loads.

4.4 Conclusion

High-speed data is seen as the key to the successful delivery of wireless multimedia applications. Enhancement GSM systems can support both circuit-mode high data rate services and packet-mode high packet rate services. The consideration of such enhancements is based upon two ideas: one is to increase capacity for future data applications, the other is to exploit existing system resources. Beside the existing GSM system, these concepts of enhancements are generic and applicable to other TDMA systems as well.

In this chapter, a systematic MAM is proposed to be applied in analysing high data rate integrated systems and integrated queuing systems with reservation protocols. In Section §4.1, an efficient MAM is proposed to analyse HSD in circuit-mode integrated services. Moreover, a hybrid reservation scheme is proposed to optimise handoff call performance. Meanwhile, the impact of terminal mobility on call performance is then discussed. In Section §4.2, the integration of HSD into voice services in the packet-mode switching mechanism is analysed by an efficient MAM. Subsequently, a new multiple priority-based control scheme and a voice coding scheme are proposed to ease the congestion of data packets. The impact of terminal mobility on packet performance is also investigated. In Section §4.3, different channel access strategies in reservation protocols are compared. A benchmark for data packet is established by evaluating the maximum allowable data packet queuing time. Furthermore, a prioritised queuing scheme is proposed to control packet congestion in overloaded regions without degrading voice packet services.

The main conclusions of this section are summarised as follows. Firstly, MAMs are robust enough to analyse the traffic queuing performance of HSD integrated services. Secondly, the optimal resource allocation in integrated services mainly depends on the mixed traffic characteristics. The effect of terminal mobility and data packet message length on call performance under high traffic loads is highly significant and thus important to examine. Finally, the multiple priority scheme is a promising technique for easing packet congestion without the loss of data preferences.

Chapter 5

Quality-Based Dynamic CAC in IWSs

The previous chapter is concerned with the performance analysis of high data rate in time-channel multimedia systems. In this chapter we concentrate on the development of call control schemes and propose some new control methods or design guideline to improve the performance of integrated services. As mentioned earlier, apart from the existing voice service, future wireless systems must support multimedia services. In order to provide a wide range of integrated services, resources must be allocated in an efficient and dynamic manner. Under the constraint of limited resources, call admission control (CAC) can be used as a design tool to satisfy different levels of QoS while maximising spectral efficiency. This chapter is devoted to the investigation of such schemes in two kinds of integrated services. The first one is in the interference limited DS-CDMA systems and the second one is in the overlaid systems. In Section §5.1, quality-based call control schemes in the integrated DS-CDMA systems are studied. Subsequently, the overflow traffic handling scheme is investigated in Section §5.2. Summaries appear at the end of each section.

5.1 Quality-Based CAC in the Power-Controlled DS-CDMA IWSs

Since capacity in DS-CDMA systems is interference-limited, there exist tradeoff relationships among system capacity, interference, coverage and communication quality. One of the keys in DS-CDMA system design is to guarantee the required QoS at an average level while providing maximum system capacity. Therefore, CAC plays an important role because it can directly control the maximum number of users and hence affect the interference. In order to achieve maximum capacity, the interference fluctuation needs to be smoothed. In a voice system, the study of intercell interference modelling is of importance because it can determine allowable maximum number of users. In this study, the intercell interference is modelled as a gamma distribution rather than a conventional Gaussian distribution. An optimal threshold design method for a single voice service is derived so as to satisfy the requirements for both loss and blocking probability. For the integrated services, a Non-Collision CDMA (NC-CDMA) mechanism is developed to

integrate data packets into the conventional voice system. Subsequently, the delay constraint for data packets is numerically analysed. In the end, a new design guideline for satisfying both loss probability and delay limit is derived. The result shows that the gamma distribution can yield a more accurate bound than using Gaussian distribution and the optimal design methods can be effectively applied to the interference-limited integrated systems.

This section is organised as follows. Firstly, the development of CDMA systems is introduced in Subsection §5.1.1. The channel efficiency of spread ALOHA is then discussed in Subsection §5.1.2. Subsequently, the CAC for the single voice service is investigated in Subsection §5.1.3. Accordingly, the mixed traffic performance is studied in Subsection §5.1.4. Finally, a summary is presented at the end.

5.1.1 Development Overviews

Spread spectrum is a technique that allows multiple users to efficiently use the same bandwidth. Historically, the use of spread spectrum techniques for anti-jamming in military communications has been long established. The use of wider bandwidth allows spread spectrum to effectively mitigate interference [*Pick91*] and [*Kohno95*]. In recent years, the potential use of spread spectrum in wireless commercial environments has aroused much more interest, because the emerging new generation systems must provide a more flexible air interface than the present ones.

For more than a decade, research has been intensifying to find enabling techniques to introduce multimedia capabilities into cellular communications. In Europe, the third generation system is known as UMTS originally and has been developed by the Europe Telecommunications Standards Institute (ETSI). Work towards the third generation system has been put into operation by several major programs, such as RACE and Advanced Communications Technologies and Services (ACTS) as described in [*Nikula98*]. Among the spread-spectrum communications, CDMA systems are known as the most widely deployed commercial wireless systems, e.g., the IS-95 systems.

Note that the actual commercial development largely depends on the two key factors, i.e., technology push and market pull, although the industry regulating body also plays a significant role. It is worthwhile to mention here that, in January 1998, wideband CDMA for the UMTS terrestrial radio access (UTRA) system was selected to be used in frequency-division-duplex (FDD) operation [*Erik98*]. In Japan, the Association of Radio

Industries and Business (ARIB) had made an early decision to use W-CDMA technology. A merger between European and Japanese technology based on W-CDMA using a GSM core network is being promoted as a possible universal standard, commonly known as International Mobile Telecommunications in the year 2000 (IMT-2000) [*Erik98*]. Recently, QUALCOMM and Ericsson have also agreed to support a single next-generation wireless standard based on CDMA2000 system. At the time of writing this thesis, the Third-Generation Partnership Project (3GPP) is working towards the harmonisation of these two proposals.

Apart from the many advantages of CDMA systems, which have been mentioned in the first chapter, DS-CDMA is well suited for transmitting VBR services [*Baier93*] and [*Baier94*]. Because the use of VBR transmission can reduce the interference and hence smooth the interference fluctuation, the transmitting power can be reduced. The reason for this is that the reduced bit rate can transform into the increased processing gain if the chip rate needs to be constantly maintained.

However, CDMA systems have two noticeable problems, i.e., near-far effect and self-jamming. Near-far effect means the user close to the receiver will suffer less attenuation than the one far away, while self-jamming refers to interference arising from the non-orthogonal users. As a consequence, the price paid for these is the required precise and fast power control algorithms. The actual QoS of a CDMA system becomes a function of the number of active users and the tightness of the power control. It is known that open loop power control and closed loop power control are implemented in the present IS-95 reverse link. The capacity can be maximised if the transmitted power from mobile terminals can be perfectly controlled within the required SIR. However, practical limitations prevent this perfect power control scheme from being realised due to the terminal mobility and propagation conditions. Generally, the impact of traffic variation on performance needs to take the power control into account. However, in order to simplify the analysis, a perfect power control and a CBR transmission are assumed in this study.

As mentioned in Chapter 3, future wireless services must focus on the provision of multimedia services. CDMA is suitable for the delivery of multimedia services because it exploit the gain of statistical multiplexing. In addition, different levels of communication QoS must be satisfied while achieving maximum spectral efficiency. As a result, packet-based transmission is seen as a favourable technique for the delivery of bursty data. As in the case of integrated services in TDMA, resource allocation for packet data traffic in the

integrated services of CDMA systems is required to follow efficient and dynamic principles.

CDMA capacity is limited by the total interference, which includes intercell interference and intracell interference [*Viterbi93*]. The key to achieve the provision of maximum capacity is to smooth interference fluctuation. Therefore, in order to allocate resource optimally and exploit capacity maximally, the study of CAC in traffic integration becomes important. The investigation of voice and data simulation appeared in [*Wilson93*]. The results show that data packets can be integrated in the near-perfect statistical multiplexing for mixed traffic and the propagation factor can reduce the spectral efficiency. In particular, compared with D-TDMA, the integrated services in CDMA have higher bandwidth efficiencies but longer data delay. Because the excessive data delay is undesirable for data users, the impact of CAC on data packet delay in system design needs to be further studied.

5.1.2 Channel Efficiency in SAMA

Generally, because spread spectrum multiple access (SSMA) systems are power-limited, the metric of throughput in the conventional ALOHA TDMA systems is found to be inappropriate for the accurate measurement of multiple-access channels in the SSMA systems. The multiple access channel efficiency, which takes average power and bandwidth resource into consideration, is especially designed for this purpose. The multiple access channel efficiency of ALOHA random access called spread ALOHA multiple access (SAMA), has been studied in [*Abram94*]. Due to proposals to use slotted ALOHA in future W-CDMA systems [*Erik98*], in this study, we show the channel efficiency of slotted ALOHA as follows.

Let G denote the channel traffic, which is a fraction of the maximum possible data rate, and S denote the channel throughput. It is well known that the relationship of the channel traffic and throughput for the case of slotted ALOHA can be shown by: $S = Ge^{-G}$.

If the capacity C_{con} (b/s) of the conventional additive white Gaussian noise channel can be given by:

$$C_{con} = W \log(1 + SNR) \tag{5.1}$$

where W is the channel bandwidth in hertz, and SNR is the average signal-to-noise ratio (SNR). The capacity C_{ma} of the multiple access channel can be shown as [*Abram94*]:

$$C_{ma} = SW\log(1+SNR/G) \tag{5.2}$$

Therefore the channel efficiency can be defined as the ratio of the multiple access capacity to the conventional capacity:

$$r = \frac{C_{ma}}{C_{con}} = \frac{Ge^{-G}\log(1+SNR/G)}{\log(1+SNR)} \tag{5.3}$$

Finally, we have the channel efficiency as shown in [Figure 5.1].

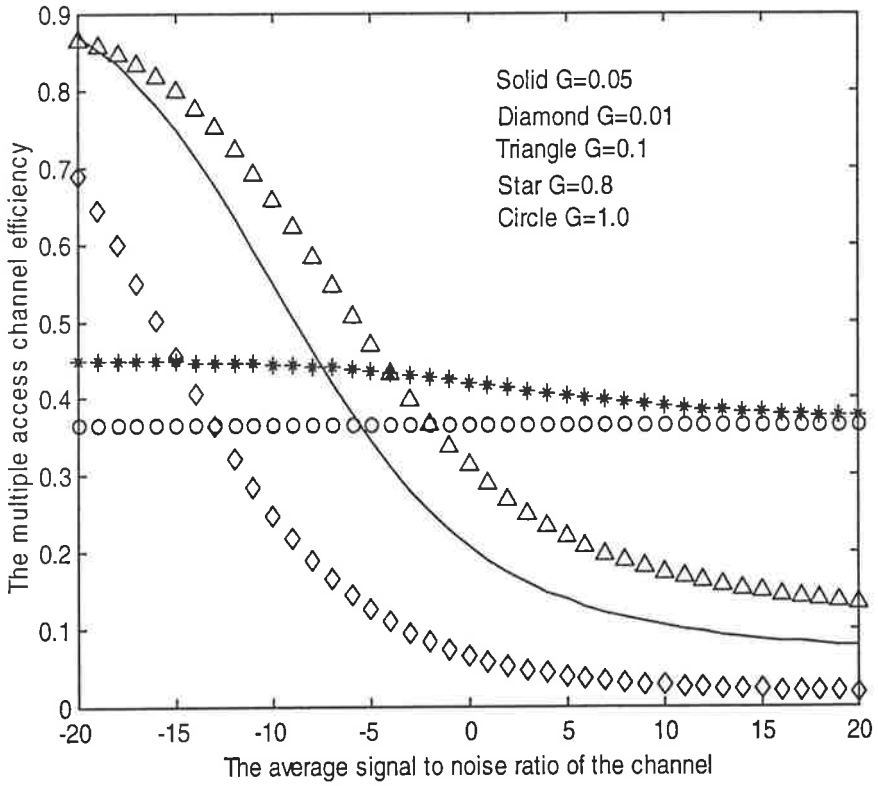


Figure 5. 1: The multiple access channel efficiency

From the figure, we observe that channel efficiency will be reduced according to the increase of SNR. Importantly, the channel efficiency has a non-linear relationship with the channel traffic. This is because the collision of packets will increase when the traffic load

becomes heavy. In particular, when the channel traffic G is equal to a maximum ($G=1$), the channel efficiency is actually equivalent to the maximum throughput of slotted ALOHA.

For a SAMA system, the studies of CPDC, in which multiple users compete for the same code channel, can be found in [*Guo96*] and [*Das97*]. The reason for using the same code is that the common packet channel can be efficiently used to transport the short and bursty packets and hence improve channel efficiency.

5.1.3 Threshold-Controlled CAC in a Voice System

For voice service with multiple code channels, the aim of using call admission control (CAC) in DS-CDMA systems is to guarantee both the required blocking probability and loss probability for communication quality. In general, there are two kinds of CAC divisions [*Ishika97*]. The first one is called interference-based CAC (ICAC), where a new call is blocked if the total interference level exceeds an allowable threshold. The second one is to maximise the Erlang capacity in each cell, which is named threshold-controlled CAC (TCAC). In the TCAC, a new call is rejected if there are already N_u users in the system. As a comparison, from a signalling point of view, the implementation of ICAC requires extra overheads for BS hardware, whereas TCAC can be simply implemented by using software in the BS.

Since the study of the reverse link is a much more challenging problem, only the reverse link is concentrated here. The mean voice call arrival rate is assumed to be λ_v and the mean service rate is μ_v. A slow speech activity detector is employed with activity factor α_v. I_{other} and I_{req} represents the interference coming from other cells and the allowable maximum total interference respectively.

Hence, the aggregate interference in a reverse link of a multicell system can be shown as [*Ishika97*]:

$$E_b(k-1)/PG + N_0 + I_{other} \le I_{req} \tag{5.4}$$

where k is the number of active users in a cell, E_b is signal energy per bit, N_0 is thermal noise power density and PG denotes the processing gain.

If the intercell interference is transformed into an equivalent number of active users m, which is a nonnegative real number, we can have $m = I_{other} PG / E_b$. Taking the intracell interference into account, then we have this expression: $k + m \le c_{\max}$, where $c_{\max} = \frac{PG(1-\eta_{\inf}^{-1})}{E_b / I_{req}} + 1$ and $\eta_{\inf}$ is defined as $\eta_{\inf} = I_{req} / N_0$. Apparently, there exists a tradeoff relationship between capacity and E_b / N_0. In other words, high capacity is at the expense of E_b / N_0, and vice versa.

The blocking probability in the TCAC scheme can be simply expressed by the conventional Erlang formula:

$$P_B = \frac{(\lambda_v / \mu_v)^{N_u} / N_u!}{\sum_{j=0}^{N_u} (\lambda_v / \mu_v)^j / j!} \tag{5.5}$$

As an example, we plot the blocking probability under the TCAC policy as shown in [Figure 5.2].

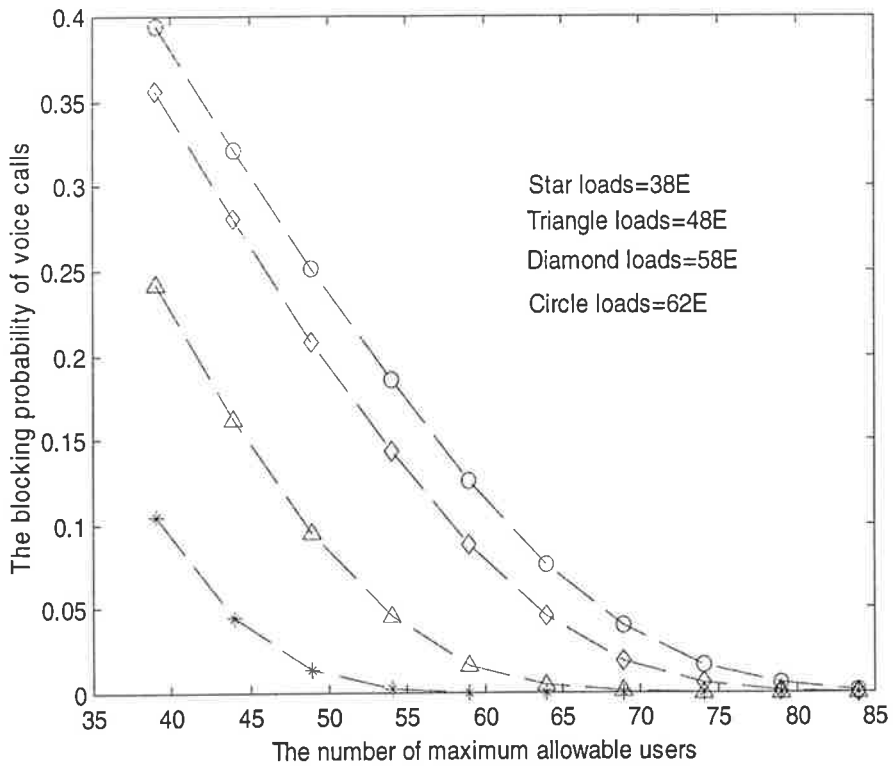


Figure 5. 2: The blocking probability for the TCAC

From this figure, we observe that the number of maximum allowable users must be limited to a predefined threshold if the QoS needs to be guaranteed to a certain amount, i.e., $P_B = 1\%$. In addition, we find that this threshold is dependent upon the offered traffic load ρ_v as well. If the offered load increases, more calls will be possibly rejected. In the end, through the control of the maximum number of active users under a certain offered load, we are able to guarantee the required communication QoS at call level.

Taking voice activity into account, the probability of k active users can be shown by:

$$P_k = \sum_{r=0}^{N_u} Bi_v(k,r,\alpha_v) \frac{(\lambda_v/\mu_v)^r / r!}{\sum_{j=0}^{N_u} (\lambda_v/\mu_v)^j / j!} \tag{5.6}$$

where the binomial distribution is represented by:

$$Bi_v(k,r,\alpha_v) = \binom{r}{k} \alpha_v{}^k (1-\alpha_v)^{r-k} \quad (r \geq k) \tag{5.7}$$

Here we can use a step function $u(x)$ to denote the threshold-based capacity:

$$u(k+m-c_{\max}) = \begin{cases} 1 & k+m>c_{\max} \\ 0 & k+m<c_{\max} \end{cases} \tag{5.8}$$

Therefore, the loss probability can be written as:

$$P_{loss} = \frac{T_c \cdot E(ku)}{T_c \cdot E(k)} = \frac{\sum_{k=0}^{N_u} kP_k \int_{m^+}^{\infty} g(m)dm}{\sum_{k=0}^{N_u} kP_k} \tag{5.9}$$

where T_c is the observed time and $g(m)$ is a gamma distribution, which is represented by two parameters v and α_g.

Namely, we have:

$$g(m) = \frac{\alpha_g{}^v m^{v-1}}{\Gamma(v)} e^{-\alpha_g m} \tag{5.10}$$

where $\Gamma(v) = \int_0^{\infty} \alpha_g{}^v m^{v-1} e^{-\alpha_g m} dm$ is a gamma function. The mean of the gamma distribution can be shown by $E(m) = v/\alpha_g$ and the variance by $Var(m) = v/\alpha_g{}^2$.

Although the Gaussian distribution is adopted in some previous studies [*Gilhou91*], [*Viterbi93*] and [*Viterbi94*], here we use the assumption of gamma distribution for an equivalent number of users resulting from intercell interference. By using the Gaussian interference assumption, a modified Chernoff upper bound or a central limit theorem can be used to approximately derive the loss property [*Viterbi93*]. It is well known that Gaussian distribution has a bell-shaped curve, which is symmetrically centred on the mean. In contrast to the Gaussian distribution, Gamma distribution exhibits slower decay than Gaussian distribution on the upper side tail. Therefore, the use of the gamma distribution can avoid the overestimation of system capacity caused by the Gaussian approximation. This explains the reason that the gamma distribution is used in this study rather than the conventional Gaussian distribution.

Then we can have:

$$\int_{m^+}^{\infty} g(m)dm = \int_{m^+}^{\infty} \frac{\alpha_g{}^{v} m^{v-1}}{\Gamma(v)} e^{-\alpha_g m} dm \tag{5.11}$$

Namely,

$$\int_{m^+}^{\infty} g(m)dm = \frac{\int_{\alpha_g(c_{\max}-k)}^{\infty} (\alpha_g m)^{v-1} e^{-\alpha_g m} d(\alpha_g m)}{\Gamma(v)} = \frac{\int_{\alpha_g(c_{\max}-k)}^{\infty} (\alpha_g m)^{v-1} e^{-\alpha_g m} d(\alpha_g m)}{\int_0^{\infty} \alpha_g{}^{v} m^{v-1} e^{-\alpha_g m} dm} \tag{5.12}$$

It is known that a gamma function $\Gamma(v)$ can be split into two incomplete functions as:

$$\Gamma(v) = \Gamma_P(v,x) + \Gamma_Q(v,x) \tag{5.13}$$

where $\Gamma_p(v,x) = \int_0^x e^{-t} t^{v-1} dt$ and $\Gamma_Q(v,x) = \int_x^{\infty} e^{-t} t^{v-1} dt$.

Then we can have this expression:

$$\int_{m^+}^{\infty} g(m)dm = \frac{\Gamma_Q(v, \alpha_g(c_{\max}-k))}{\Gamma(v)} \tag{5.14}$$

Finally, we obtain the loss probability for a voice system:

$$P_{loss} = \frac{\sum_{k=0}^{N_u} \{kP_k \cdot \Gamma_Q[v, \alpha_g(c_{\max} - k)]\}}{\sum_{k=0}^{N_u} kP_k \cdot \Gamma(v)} \tag{5.15}$$

In the end, we can plot the logarithmic loss probability against the allowable number of users as shown in [Figure 5.3].

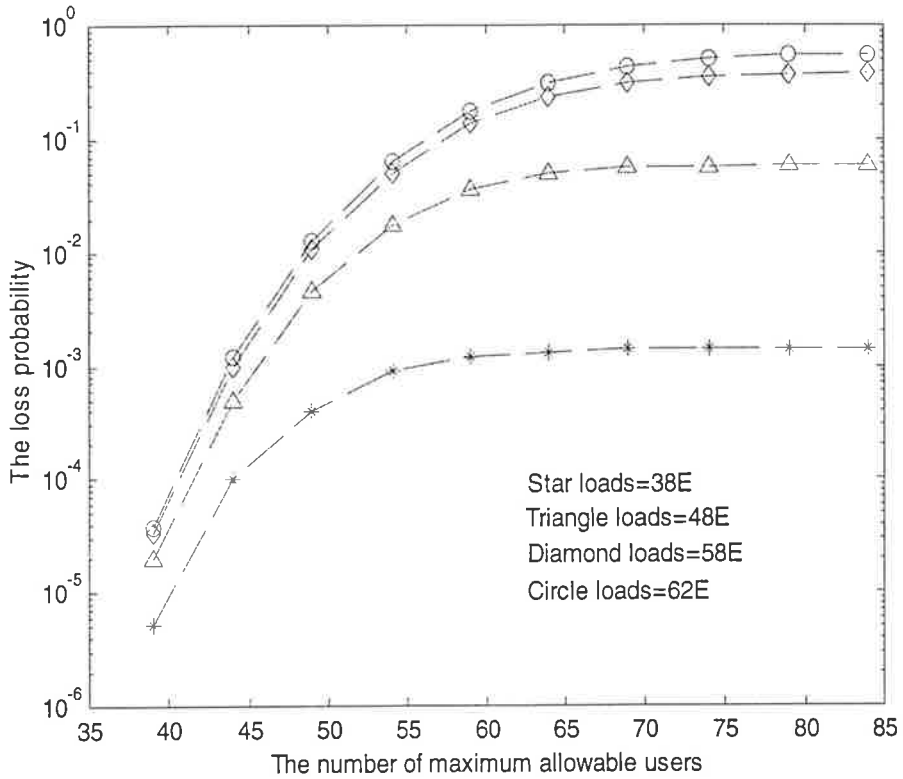


Figure 5. 3: The loss probability for the TCAC

To summarise the optimal threshold design method for the single voice service, we describe it as below:

Case 1: It assumes that there exists a common intersection between the constraints of loss probability and blocking probability. If the maximum number of allowable voice users is $a_{P_{loss}}$, which satisfies $P_{loss} \le P_{loss}^{(req)}$, and the minimum number of users $a_{P_{Block}}$, which satisfies the blocking probability limit $P_B \le P_B^{(req)}$, then we can have the maximum threshold of the CAC scheme as $a_T = a_{P_{loss}}$ if $a_{P_{loss}} \ge a_{P_{Block}}$.

In order to guarantee both loss probability and blocking probability, this can be shown by the dotted line as in [Figure 5.4]. In this study, we find that the threshold for the $P_B = 10^{-2}$ bocking probability is $a_{P_{Block}} = 50$ and the threshold for the loss probability $P_{loss} = 10^{-3}$ is $a_{P_{loss}} = 55$. Therefore, the overall threshold of CAC in the system should be $a_T = 55$. Similarly, as another example, there still exists a solution $a_T = a_{P_{loss}}$ as shown by the dashed line in [Figure 5.4].

Case 2: If the constraints of the loss probability and blocking probability are mutually exclusive, then there is no solution to satisfy both the constraints. That is, if $a_{P_{loss}} < a_{P_{Block}}$, the maximum threshold input $a_T = \{\Phi\}$, where Φ denotes an empty set.

This can be illustrated by the solid line shown as in [Figure 5.4]. In this case, either one of the two QoS parameters must be allowed to be degraded so as to have a solution for the controlled threshold under such a CAC scheme.

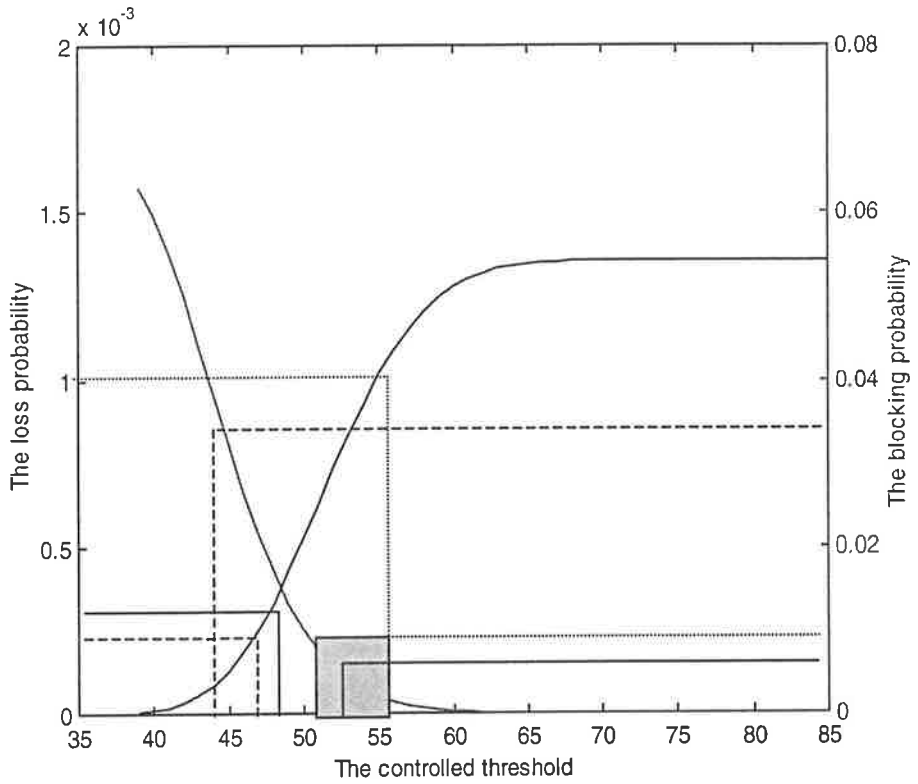


Figure 5. 4: The threshold design in DS-CDMA systems

5.1.4 Dynamic CAC Schemes in Integrated Services

As far as multimedia applications are concerned, there is a need for better understanding of the tradeoff relationship between the different levels of required QoS and the capacity for each type of service. To achieve the required QoS, this can be realised through the use of CAC policies. As mentioned previously, because DS-CDMA can statistically multiplex heterogenous traffic without complicated radio resource management, it is found to be suitable for multimedia services and no special protocol structures are required for the special arrangement of integrated CBR and VBR services. The investigation of single voice services has been presented in the previous studies [*Cooper78*], [*Pick91*] and [*Gilhou91*]. Until recently, the integrated services in the CDMA systems have drawn much more attention and studies have appeared in [*Wilson93*], [*Abram94*], [*Guo96*] and [*Naga98*], etc. As mentioned in Chapter 2, the key problem in integrated traffic access control is to smooth interference fluctuation. In this study, we propose and develop a non-reservation method to integrate data packets into voice service and then evaluate traffic performance of integrated services in terms of system capacity.

5.1.4.1) Voice Traffic Dynamic Control Scheme

If the number of active users exceeds the maximum allowable users, the actual signal-to-interference ratio (SIR) becomes smaller than the required SIR and communication quality will be degraded. Therefore, the voice control scheme is to guarantee that the active users are below the maximum level. Without considering intercell interference, the number of maximum allowable users (MAU) C_m, which are active in a single cell, can be derived from the last subsection:

$$E_b(k-1)/PG + N_0 \le I_{req} \tag{5.16}$$

In order to maintain the required communication quality, the number of active voice users k in a cell must satisfy this:

$$k \le C_m = \left\lfloor \frac{PG(1-\eta_{\inf}^{-1})}{E_b / I_{req}} + 1 \right\rfloor \tag{5.17}$$

where the parameters are described in the last Subsection §5.1.3 and $\eta_{\inf} = I_{req} / N_0$ is used to control the maximum allowable transmit power of the MS.

If the voice activity factor is denoted by α_v, the probability that m active voice users resulting from the M_u voice terminals can be represented by a binomial distribution:

$$Bi_v(m, M_u, \alpha_v) = \binom{M_u}{m} \alpha_v^{\,m} (1-\alpha_v)^{M_u - m} \quad (M_u \geq m) \tag{5.18}$$

Due to the constraint of interference, the probability that more than C_m active users among the rest of the $(M_u - 1)$ users cause call dropping can be expressed as:

$$P_{loss} = \sum_{m=C_m}^{M_u - 1} \binom{M_u - 1}{m} \alpha_v^{\,m} (1-\alpha_v)^{M_u - 1 - m} \quad (M_u - 1 \geq m) \tag{5.19}$$

In order to maintain the loss probability at a required level, we can use the following algorithm to determine the maximum allowable number of users M_u in the system:

Algorithm 5.1:

1) Initiate the voice activity factor α_v, the available channel C_m, and also the pre-assumed number of users $M_u^{(0)}$.
2) Compute the loss probability $P_{loss}^{(0)}$.
3) Increase the number users from $M_u^{(0)}$ to $M_u^{(1)}$ and then compute $P_{loss}^{(1)}$.
4) Iterate successively until $\left\| P_{loss}^{(k+1)} - P_{loss}^{(k)} \right\| \leq \varepsilon$.
5) Obtain $M_{u,\max} = M_u^{(k+1)}$.

If we assume that the voice activity factor is $\alpha_v = 0.4$ and the maximum available channels are $C_m = 10$, we can have the MAU equal to $M_{u,\max} = 12$ in order to satisfy the loss probability of no more than 10^{-3}.

Subsequently, it is easy to derive the leftover capacity M_v for data packets by voice users:

$$M_v = C_m - \alpha_v M_u (1 - P_{loss}) = C_m - \alpha_v M_{u,\max} \tag{5.20}$$

Then the normalised leftover capacity for data packets can be expressed by:

$$\frac{M_v}{C_m} = 1 - \alpha_v M_u (1 - P_{loss}) / C_m = 1 - \alpha_v M_{u,\max} / C_m \tag{5.21}$$

In the end, we can have the loss probability versus the number of users while the communication quality satisfies the required SIR as shown in [Figure 5.5].

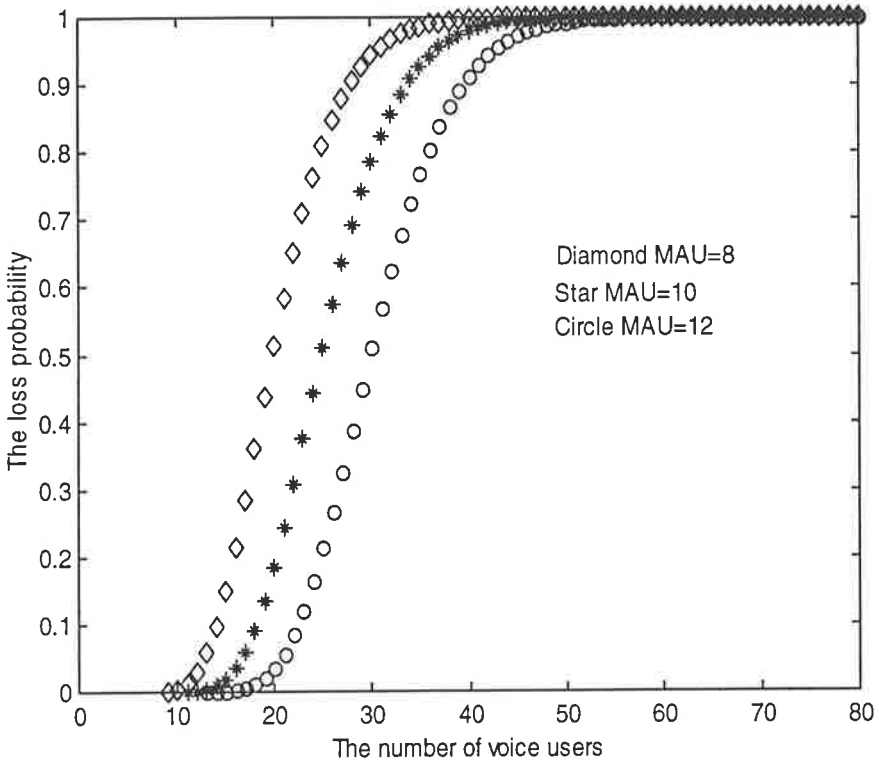


Figure 5. 5: The effect of the number voice users on loss probability

From this figure, we observe that the total loss probability increase according to the increased voice users M_u. In addition, for a fixed number of voice users M_u, the smaller the MAU is, the bigger the loss probability becomes.

5.1.4.2) Data Traffic Dynamic Control Scheme

In order to utilise network resources in an efficient way, resource allocation must be achieved as flexibly as possible in response to the instantaneous traffic demands. More importantly, as mentioned before, maximum throughput, minimum delay, low blocking and loss probability must serve as the guidelines in network design.

Since data packets are delay insensitive, we can schedule the transmission of data packets in both the voice idle periods and the periods of vacant channels. This can result in improved channel efficiency but incur long data packet delay. Sampath [*Sam97*] proposes to use permission probability for data packets to dynamically control packet access in the integrated services. In fact, such a proposal is similar to the one used in PRMA systems [*Goodm89*]. Tan [*Tan96*] analysed the RCMA protocol for integrated voice and data services in. In particular, Nagatsuka [*Naga98*] studied the integrated services in a contentionless mode.

In this study, we develop a NC-CDMA channel access method for integrating data services into voice services. Our proposal is different from the one used in [*Naga98*]. Nagatsuka uses a reservation mechanism for data packets. As a result, control time difference (CTD), which represents the time when permission is detected at a base station until the time when a data packet in a terminal is transmitted, becomes necessary. Moreover, CTD has a potential to reduce system capacity due to the use of extra overhead [*Naga98*]. Conversely, the use of CPSM data packets can increase traffic throughput.

The control scheme for data packets in the integrated services are described as follows:

1) A data user who wants to transfer packets needs to lodge a request to a BS first and the receiving power at the BS is perfectly controlled. Data users must hold permission from the BS for transmission.

2) The structure of data packets consists of packet header, packet body and packet tailer. The packet header is for indicating message length, packet body for information and packet tailer for cyclic redundancy check (CRC) in order to improve the BER performance. In addition, the frame header is used for sending the request to the BS.

3) By giving priority to voice users over data users, data users are assigned a unique code only when voice users do not occupy or request them.

4) A base station always monitors the number of active voice users. Data terminals are then allowed to transmit packets immediately as long as there exist vacant channels and satisfy the required SIR. The redundant packets are allowed to wait in

the queue. The waiting data packets in the queue are served on a basis of the FIFO discipline. The advantage of the queuing buffers is to increase the carried loads at the expense of the loss probability. This can be shown as in [Figure 5.6].

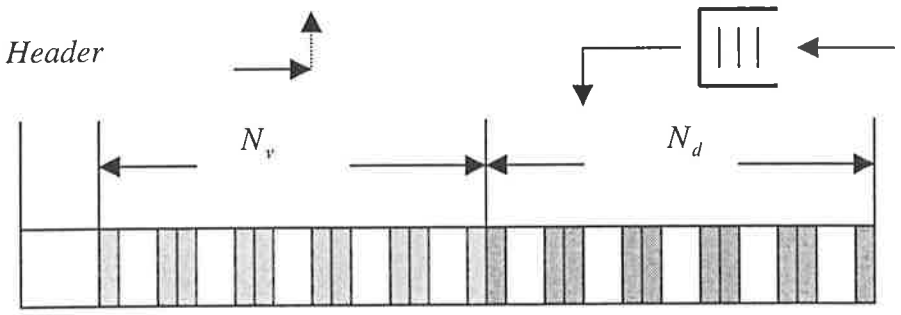


Figure 5. 6: The uplink structure of voice and data integrated services

If the probability that k data users are active is donated by P_k, this can be written as:

$$P_k = \sum_{r=k}^{N_u} Bi_d(k,r,\alpha_d) \frac{{\rho_d}^r / r!}{\sum_{j=0}^{N_u} {\rho_d}^j / j!} \tag{5.22}$$

where ρ_d is the offered load of data users, α_d is the data activity factor.

Based on the required SIR and a certain number of voice users M_u, the approximation of the loss probability in the integrated services can be represented by:

$$P_{loss} = \frac{\sum_{k=C_M-M_u+1}^{N_u} kP_k \sum_{x=0}^{C_M-k+1} Bi_v(x,M_u,\alpha_v)}{\sum_{k=0}^{N_u} kP_k \sum_{l=0}^{M_u} Bi_v(l,M_u,\alpha_v)} \tag{5.23}$$

Therefore, we can have the result for the loss probability versus the data offered load as shown in [Figure 5.7].

From the figure, we observe that the total loss probability of the integrated voice and data packet services will reduce according to the decrease of data packet activity factor α_d. For instance, the data packet activity factor is assumed to be $\alpha_d = 1$ in the World Wide Web (WWW) service [*Naga98*] and the required loss probability is equal to 1%. The loss probability achieves this requirement when the data user offered load is

$\rho_d = 0.43$ Erlangs. However, when the data activity factor is $\alpha_d = 0.8$, the loss probability which corresponds to the same data offered load $\rho_d = 0.43$ Erlangs is about 0.5%.

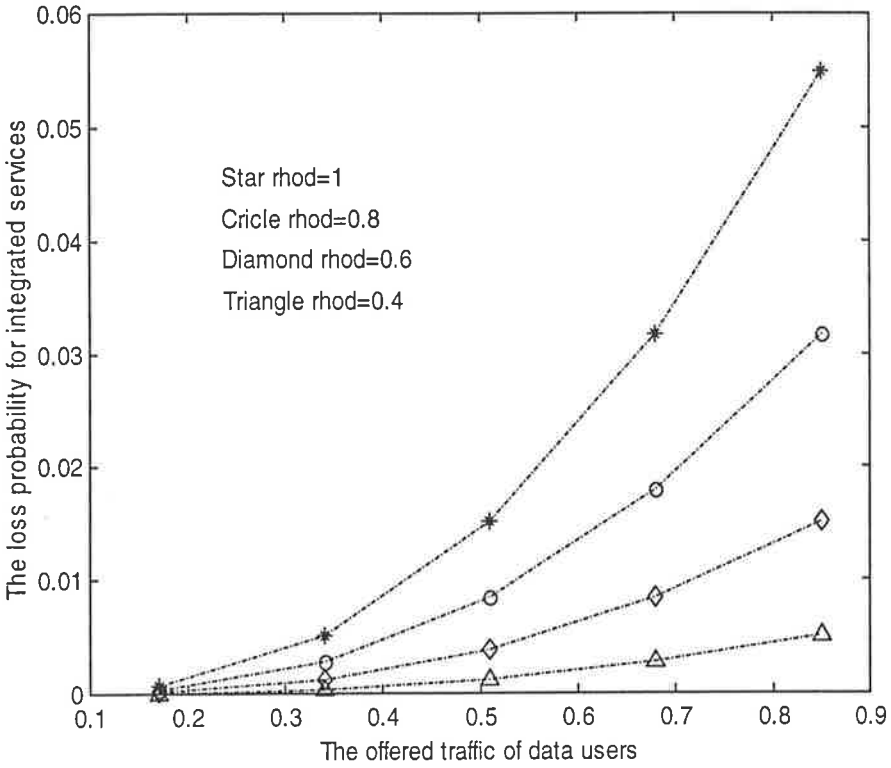


Figure 5. 7: The loss probability of integrated services

Next we proceed to analyse data packet delay. First, we assume that data packets can reach equilibrium states faster than that of voice calls. Under this assumption, we regard that voice calls always occupy C_v channels and data packets can exploit the leftover capacity $(C_m - C_v)$. Therefore, a M/D/c queuing delay can be applied to this specific problem.

The queuing delay time for data packets can be simply shown as:

$$W_{Delay} = \frac{\rho_d^2}{2\lambda_d(1-\rho_d)} \tag{5.24}$$

where the data packet offered load is $\rho_d = \frac{\lambda_d}{\mu_d(C_m - C_v)}$.

That is,

$$W_{Delay} = \frac{\rho_d}{2\mu_d (C_m - C_v)(1-\rho_d)} \tag{5.25}$$

Therefore, we have the result for the loss probability against the data delay as shown in [Figure 5.8].

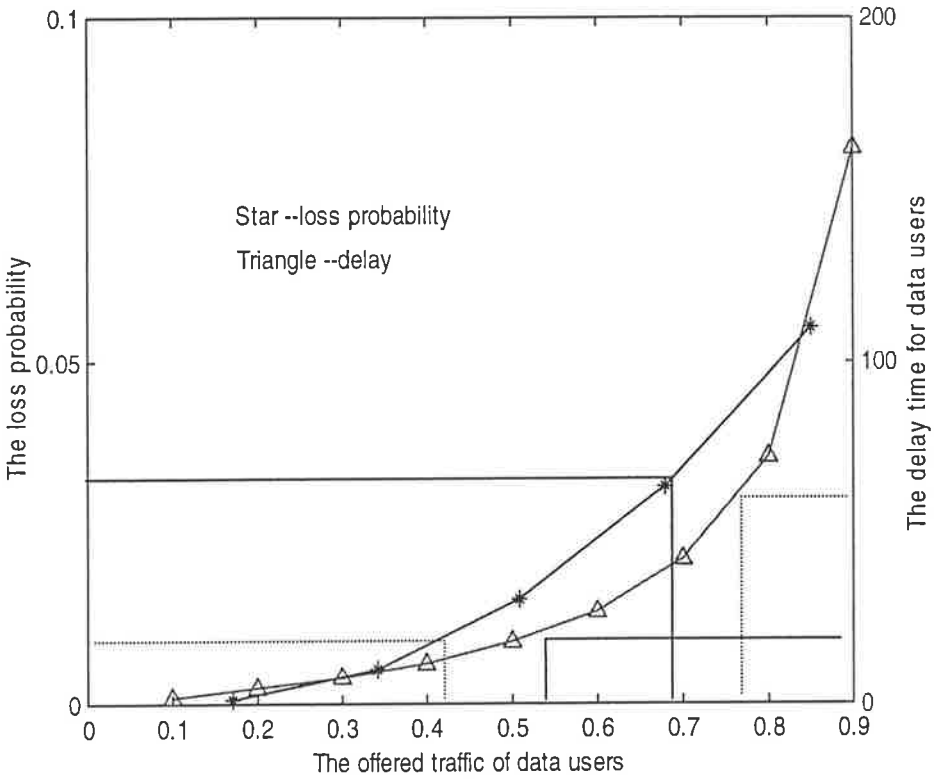


Figure 5. 8: The loss-delay in integrated services

As a result, we can form a design guideline for the integrated services in DS-CDMA systems as follows.

If the maximum input load is $a_{P_{loss}}$, which satisfies $P_{loss} \le A$, and $a_{W_{Delay}}$ for the delay limit B, the maximum allowable data packets input is $a_T = \min\{a_{P_{loss}}, a_{W_{Delay}}\}$.

This can be simply proved as follows.

We assume there are two functions, $y_1 = f_1(x)$ for loss probability and $y_2 = f_2(x)$ for packet delay. Both are the monotonically increasing functions.

Case 1: if $A \geq B$, this means $f_1(x_1) \geq f_2(x_2)$. Since $y_1 = f_1(x)$ is monotonically increasing, we can have $f_1(x_2) \leq f_1(x_1)$ while $x_2 \leq x_1$. Noting $f_1(x_1) = A$, therefore we can have $f_1(x_2) \leq A$. Because we already have $f_2(x_2) = B$, the minimum value x_2 satisfies the requirements of both loss and delay.

Case 2: if $A < B$, this means $f_1(x_1) < f_2(x_2)$. Since $y_2 = f_2(x)$ is monotonically increasing, we can have $f_2(x_1) < f_2(x_2)$ while $x_2 > x_1$. Noting $f_2(x_2) = B$, therefore we can have $f_2(x_1) < B$. In addition, we already have $f_1(x_1) = A$. So the minimum value x_1 becomes the solution.

In the end, this completes the proof. For instance, to satisfy the delay constraint of 60 seconds, we find that the input data packet load corresponding to this delay is equal to 0.77 Erlangs [Figure 5.8]. However, taking the SIR into account, the input data load corresponding to no more than 1% loss probability must be less than 0.43 Erlangs. Therefore, in accordance with the proposed design guideline, we only can adopt the maximum input load equal to 0.43 Erlangs.

5.1.5 Summary

Since capacity in the DS-CDMA systems is interference-limited, there is a tradeoff relationship between system capacity, interference, and communication quality. Therefore, the CAC method needs to be used to guarantee the QoS in the DS-CDMA systems. In a voice system, the intercell interference is modelled as a gamma distribution rather than a conventional Gaussian distribution. For the voice system, the main contribution is that an optimal threshold design method for the single voice service is developed to satisfy both the required loss probability and blocking probability. For integrated services, the main contributions are: firstly, a NC-CDMA mechanism is proposed to integrate data packets into conventional voice systems; secondly, the delay constraint for data packets is then analysed theoretically; finally, a new design guideline for satisfying both loss probability and delay limit is proposed.

The result shows that the gamma distribution can yield a better bound in terms of interference than that of the conventional Gaussian distribution. Optimal threshold design method can be used as a design tool in a voice system to control the access of maximum

allowable users. For integrated services, the proposed voice and data CAC schemes can restrain the interference efficiently. A design guideline must take into account both the required loss probability and delay.

5.2 Overflow Traffic Handling Schemes in HCSs

In order to increase network capacity and radio coverage, multi-tier systems can be implemented in 'black spot' areas, where radio is difficult to cover, or 'hot spot' areas, where traffic congestion often occurs and also microcells have no overlapping coverage. Recently, particular attention has been paid to the implementation of multilayer hierarchical cell structure (HCS) systems because they can provide users with global access services with diverse mobility platforms. In order to allocate resources among layers efficiently, the performance of overflow calls needs a precise and quantitative analysis. Similar to the CAC schemes in last section, the call handling schemes can be adopted to enhance overflow traffic performance in such a mixed cell environment.

A mathematically generic model, which not only includes the performance analysis of new voice calls and handoff calls, but also can take the correlative arrival behaviour into account, is developed through a theoretical method. An exact solution for the multiple overflow approximation problems is derived by using a two-state Markov Modulated Poisson Process (MMPP), which is subsequently approximated by an interrupted Poisson process (IPP) model for the non-random overflow process. The important parameters, such as call dropping probability and waiting time distribution, are subsequently derived according to the speed classification. As a result, it concludes that the overflow traffic has to be modelled by the higher moment matching techniques if the accuracy of performance is critically required. Terminal mobility has a noticeable effect on the overflow call performance. Moreover, the new disposition policy can offer better performance than the previous schemes, especially under a heavy load condition.

This section is organised as follows. The introduction begins in Section §5.2.1. Secondly, the operation of HCSs is depicted in Section §5.2.2. The assumptions of the model are given in Section §5.2.3. Accordingly, the moment matching techniques and mathematical preliminaries appear in Section §5.2.4. Section §5.2.5 details the solutions for the overflow traffic. Then the disposition policy and network optimisation are discussed. Finally, a summary is presented.

5.2.1 Introduction

For future PWC, one of the important features is to achieve global access. The use of the satellite component for next generation networks called Satellite UMTS (S-UMTS) has

been defined in the early development of the UMTS [*Gunts98*], whereas the terrestrial component of UMTS is known as T-UMTS. The terrestrial systems consist of macrocells, microcells and even picocells. Without the satellite services, the terrestrial services are restricted to a regional coverage. In the ACTS program, different kinds of projects for integrating the S-UMTS into the T-UMTS have been undertaken, such as *INSURED*, *SINUS* and *TMMAS* as described in [*Gunts98*]. These projects are intended to provide next generation mobile satellite systems with a high data rate access of 2 Mb/s.

The deployment of satellite systems can not only provide international roaming, but also expand the present limited regional communications into seamless communications with the aid of multi-mode terminals. As a matter of fact, the overlaid systems form the multi-tier HCSs. For example, Low Earth Orbit (LEO) overlays the present cellular networks. Examples of the satellite projects include Iridium, Globalstar, and Odyssey, etc [*Kim98*]. Satellite communications can also provide services for maritime, aeronautical and rural environments, especially for sparsely populated areas, in an economical manner. Meanwhile, a variety of services must support both terrestrial and space inter-environment mobility. Therefore, the employment of small cells not only can increase traffic capacity, but also can accommodate different mobility patterns and provide backup channels for the failure calls.

Because of the constraints of propagation and terminal power consumption, it is evident that there exists a trend to smaller cell sizes. However, with the shrinking of cell size, traffic patterns with fast mobility may not be well handled by a range of small cells due to the handoff processing speed or the insufficient radio coverage. Meanwhile, as the cell size decreases, the handoff rates will increase and the signalling control messages for the frequent handoffs will exacerbate the burdens of mobile switching nodes. On the other side, from a user point of view, it requires that handoff dropping probability between layers must be designed to a minimum. Therefore, in order to reduce the handoff times and handoff dropping probability, it is necessary to develop a traffic model, which can closely reflect the correlative behaviour between the low layer and high layer while taking terminal mobility into account. The ultimate objective for the study is to ensure that the precious resource is allocated efficiently.

HCSs have many advantages over a single tier structure. The advantages of HCSs and the overflow traffic handling procedures are subsequently described as follows.

Firstly, as aforementioned, HCSs can integrate terrestrial radio access with space access. Apart from international roaming technique, HCSs are seen as a means to achieve global access. Truly seamless global communications can not be realised until the facilitation of HCSs. In particular, HCSs are able to accommodate different traffic densities and mobility patterns so as to reduce the handoff dropping probability. Secondly, HCSs can act as a traffic reliever for hot-spot cells or unevenly distributed traffic. By using an overlaid structure, the upper cell can act as an umbrella cell and provide secondary resources for hot-spot cells. More importantly, apart from the other techniques, i.e., cell splitting or channel borrowing, cell overlaid systems can be used as an effective method to alleviate traffic congestion. Next, HCSs can increase spectrum efficiency and system capacity. By using small cells, this can reduce the elevation of the antenna and power transmission. Finally, HCSs can improve radio link quality. It is evident that call quality may be degraded due to poor propagation conditions or terminal movement. The use of HCSs can re-direct the call to the upper layer with better radio coverage.

In summary, the HCSs can provide an alternative approach to re-allocate network resources. This enables us to reduce the failure probability of multiple handoffs. During the execution of overflow calls, handoff initial access is determined by a set of parameters, such as vehicle movement speed, the actual traffic load and link budget.

Generally, the overflow traffic handling procedures consist of the following aspects.

- **Cell (Re)Selection:** An initial cell selection can be based upon speed sensitive or speed insensitive schemes. In the case of speed sensitivity, the HCSs direct the mobile subscribers to the appropriate layers according to their speeds. Terminals with high mobility can be initially assigned to macrocells and low mobility terminals to microcells [*Hu95*]. Following speed-sensitive criteria, arriving calls are assigned to cells according to the call present or past CST. Conversely, in the case of speed insensitivity, the initial cell determination only depends upon the default layers irrespective of the CST [*Lagran95*]. As a result, the terminals with high mobility can avoid the increase in handoff failure rate associated with microcells. Similarly, the terminals with low mobility can be separated from the macrocells.

- **Overflow Handling Strategies:** From an overflow traffic control point of view, there are two categories, which are classified into overflow and underflow controls. Underflow control means that traffic in higher layers can make a retry to lower layers once channels in higher layers become unsuitable to handle the calls, while overflow is from the lower layers to the upper layers. From a call control point of view, there exist different ways. Only handoff calls were allowed to overflow onto the higher layer [*Langran95*]. Both new calls and handoff calls are allowed to overflow onto the higher layer [*Hu95*]. In order to give priority to handoff calls, a reservation scheme, which is exclusively for handoff calls, is taken into account in [*Hu95*].

This study is motivated by the problems of optimal resource allocation among layers and congestion control in overlaid systems. Therefore, the overflow control policy is focused on. Specifically, this study is to contrast the different analytical methods for the time-varying overflow traffic in the multiple layers and intends to improve the dropping probability by using overflow handling strategies. Whilst the overflow calls are analysed in [*Hu95*] and [*Langran95*], the take-back underflow traffic is investigated in [*Jabb97*]. However, all these studies are restricted to loss systems only. In the search of available studies, it is apparent that a queuing disposition policy, which intuitively can improve overflow traffic performance, has not been paid particular attention. As distinct from the previous studies in [*Hu95*], [*Lagran95*], [*Lagran96*] and [*Jabb97*], we propose a disposition policy for a multiple overflow queuing system, in which overflow traffic can launch a set-up request from a finite queue.

A mathematical model for a micro-macro wireless overlaid system is developed. However, this model is applicable to the generic systems, such as the satellite component overlaying the terrestrial component or even multiple overlaid systems.

5.2.2 The Operation of the System in HCSs

The overflow control policy adopted here is to allow both new calls and handoff calls to be able to overflow into macrocells once there are no idle traffic channels existing in their original cells. This is different from the policy used by Lagrange [*Lagran95*], where only handoff voice calls are allowed to overflow. In order to achieve optimum frequency reuse,

low mobility users camp on microcells initially, whereas high mobility vehicles adhere to macrocells first. Although high and low mobility has been classified [*Lagran96*] and [*Hu95*], the overflow calls will become lost once they find channels full in macrocells. In contrast, our queuing overflow control policy enables resource allocation to become more flexible.

For the sake of simplicity, only the case of fixed channel assignment (FCA) in a two-stage error-free channel HCS system is considered. However, the cases of multi-layers and multiple overflow streams are easily extended from this study. In addition, all cells are considered to be statistically identical here. The case for inhomogeneous cells can be extended from this study by using other methods, e.g., decomposition methods.

The illustration of one macrocell covering four microcells used in our study can be shown in [Figure 5.9]:

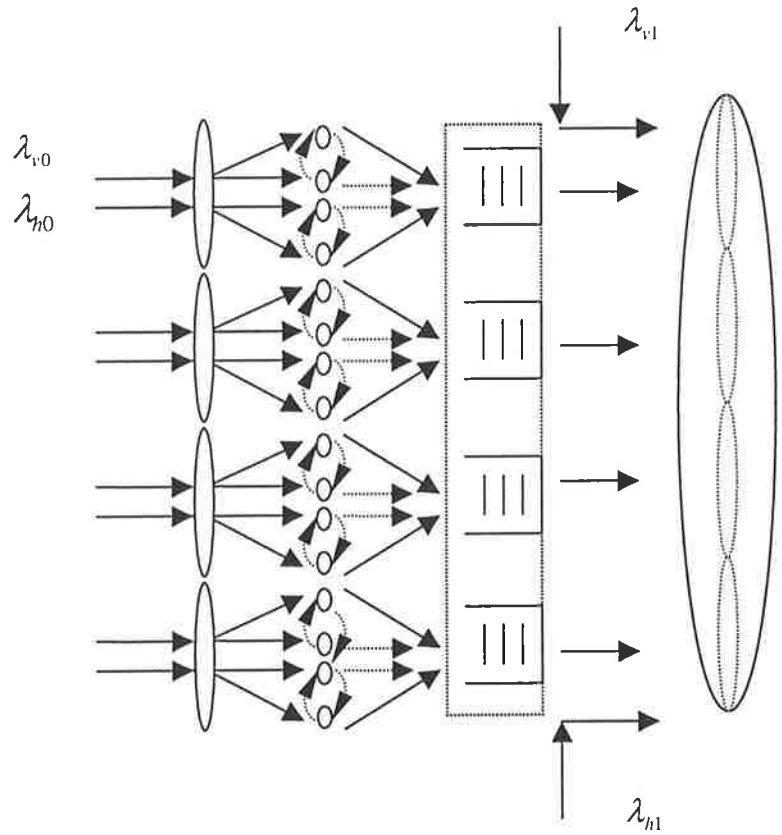


Figure 5. 9: The multilayer hierarchical overflow queuing system

In this figure, one umbrella-cell seamlessly covers four lower cells. There are new calls and handoff calls arriving at each lower cell. The overflow traffic arrives at the random switch and is then handled in the higher layer. The blocked overflow traffic can be stored in the finite buffers.

The operation of the system can be described as follows:

1) The lower layer structure is assumed to be fully embedded under the contiguous wide area coverage of the upper layer structure. If an arriving new call finds channels C_1 full in a microcell, it is allowed to overflow into macrocells. If the channels C_2 in a macrocell are temporarily unavailable, the overflow calls are allowed to wait in a finite queue and then have a potential request for set-up as shown in [Figure 5.9].

2) A handoff request is initiated when a terminal moves beyond the range of the default communication area. A handoff terminal with slow speed first searches the idle channels in the same hierarchical level. If all channels in the microcell are busy, the handoff call is allowed to re-direct to a higher level.

3) If a new arrival in a macrocell finds all channels C_2 full, this call will be blocked and is not allowed to underflow into microcells. In other words, a reversible policy is not included but can be extended from this study.

5.2.3 The Model Descriptions

In order to derive a tractable solution, we have to make some assumptions as follows:

1) New call arrivals are confined to a Poisson process. The new call arrives in microcells with mean rate λ_{v0} and in macrocells with mean rate λ_{v1}. The total number of channels C_1 in a microcell can be shared between originating calls and handoff calls. There are a total of m_0 macrocells in the network. Every macrocell can seamlessly cover n_0 microcells. The maximum capacity in macrocells is assumed to be C_2 and C_1 in microcells.

2) Handoff calls are approximately assumed to follow a Poisson process as well. The average handoff arrival rate is λ_{h0} in microcells and λ_{h1} in macrocells. Therefore the total mean arrival rate in microcells is $\lambda_0 = \lambda_{v0} + \lambda_{h0}$. In a similar way, we can

simply use $\lambda_1 = \lambda_{v1} + \lambda_{h1}$ to represent the aggregate mean arrival rate in macrocells. An overflow call that finds all channels busy in a corresponding macrocell is allowed to wait in the queuing room with a capacity of K.

3) The duration of CUT $1/\mu_v$ denotes the time that a call is in progress without any forced termination. Taking terminal movement into consideration, the rate of CHT for voice calls can be represented by the summation of the mean rate μ_v of CUT and the cell cross-over rate β_n in microcells: $\mu_1 = \mu_v + \beta_n$. Similarly, in macrocells, the mean rate of CHT can be represented by: $\mu_2 = \mu_v + \beta_M$, where the cell cross-over rate in macrocells is β_M [*Lin94*].

The simulation for the overflow traffic can be shown in [Figure 5.10].

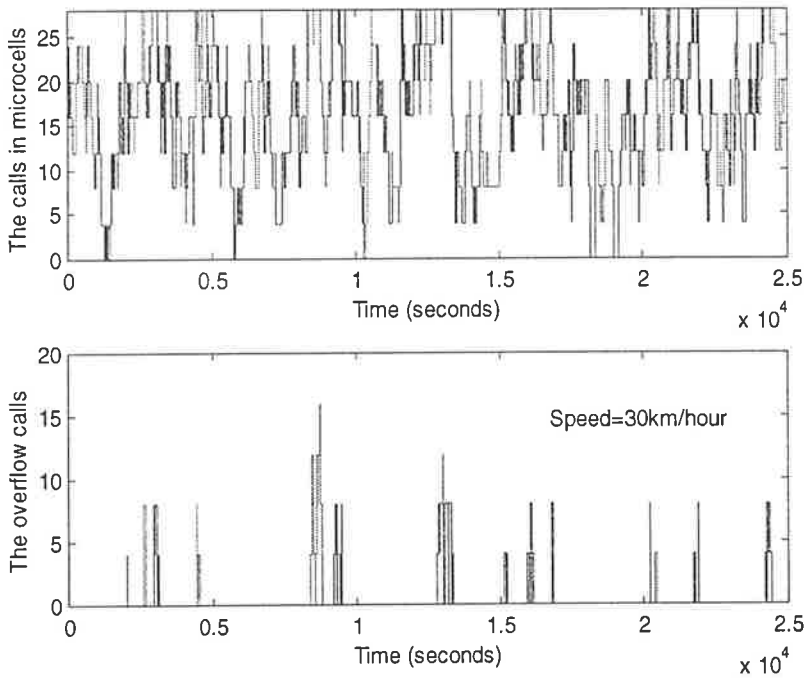


Figure 5. 10: The simulation for overflow traffic

In this simulation, traffic arrival rates at four microcells are treated as homogeneous and equal to 0.0417 calls/s in each cell. The average call duration is 2 minutes. The terminal speed is assumed to be 30 kilometres per hour.

From this study, we observe that the overflow traffic is different from the original traffic and the bursty characteristic must be properly dealt with.

5.2.4 The MMT and Mathematical Preliminaries

Although the theoretical development for overflow traffic has been long established in the wired line networks, the derived results can not be directly applicable to the context of wireless networks. As mentioned earlier, the reason for this is that the analysis in cellular networks needs to take both terminal mobility and handoff calls into account.

To characterise a process, which may have greater variability than a Poisson process, there are two kinds of methods that can be used. One is called index-of-dispersion for count (IDC) measurement. The other one is peakedness measurement. By using the first method, the variability of the process is featured by counting the number of arrivals. The IDC at time t is defined as the ratio of the variance to the mean of the arrival number $N(t)$ during an interval of time length $[0,t]$ [*Gusella91*]:

$$IDC(t) = \frac{Var(N(t))}{E(N(t))} \tag{5.26}$$

If the process is a Poisson process, we can simply have $IDC = 1$ for all time t.

In order to characterise the dependency of the instantaneous arrival rate $\lambda(t)$ at one instant of time to another, a covariance function $r(t)$ is usually used. The time constant is defined as the mean of the integral of the covariance function [*Heffes80*]:

$$\tau_c = \frac{1}{Var}\int_0^{\infty} r(t)dt \tag{5.27}$$

where $r(t)$ is:

$$r(t) = Cov(t_1, t_2) = E((\lambda(t_1) - \overline{\lambda})(\lambda(t_2) - \overline{\lambda})) \tag{5.28}$$

Obviously, the arrival rate variance becomes the covariance function while the time is equal to zero: $Var = r(0)$. Therefore, the covariance function can show the time evolution of variance. Given the arrival rate variability, longer time constants mean that there is more variability in the number of arrivals. In order to match the first three noncentral moments of the MMPP arrival rate, the approach derived by Heffes can be adopted [*Heffes80*].

Alternatively, in this study, another approach derived in [*Appendix 2*] is used. Regardless of the time effect of arrivals, the moments of the process can be used to characterise the distribution of arrival rates at an arbitrary time. By using the peakedness measurement method, a fictitious server is always assumed. Although the peakedness is not a complete characterisation for overflow traffic, it is useful and also is seen to be at the heart of the moment matching techniques (MMTs).

The development of the MMTs dates back to the middle of the 1950s. In fact, the idea comes from alternate routing in fixed telephone networks [*Girard90*]. The blocked multiple traffic streams in the primary group can be re-routed to the secondary group. As a matter of fact, this can reduce the loss possibility for the arriving traffic streams in the lower layer.

Although the exact solution for the overflow problem may be feasible, the high dimensionality of the state space and the complexity of the computation are found to be difficult to apply in practical systems. Therefore, the search for accurate approximation techniques has aroused much interest over the years [*Guerin90*]. There exist several methods that can be used to measure the overflow traffic performance.

Firstly, it is well known that the overflow process is non-random and the higher moments must be used to describe the fluctuation characteristics. An equivalent random method (ERM) is developed by Wilkinson to characterise such non-random traffic in [*Wilki56*]. The basic idea of ERM is to provide an equivalent offered traffic with equivalent mean and variance, which are equal to the actual overflow mean M_e and variance Var.

Generally, the peakedness is defined as the ratio of variance to mean of busy calls offered in the infinite servers, that is, $Z = Var / M_e$. In order to maintain a specific QoS, peaked traffic requires more channels to handle than the normal random traffic does. The required channels depend upon the degree of peakedness $(Z > 1)$. Conversely, for the case of smooth traffic $(Z < 1)$, other extended methods must be used [*Girard90*].

For example, if the equivalent offered loads are A and the equivalent channel numbers are S, an approximation is given by Rapp [*Girard90*]:

$$A = Var + \frac{3Var}{M_e}\left(\frac{Var}{M_e} - 1\right) \tag{5.29}$$

and

$$S = \frac{A(M_e^{\,2} + Var)}{M_e^{\,2} + Var - M_e} - M_e - 1 \tag{5.30}$$

Another equivalent method called Hayward approximation (HA) has been studied in wired line networks [*Fred80*]. In fact, HA is an extension of Erlang formula by taking the traffic peakedness into consideration. A simple approximate method for calculating the blocking probability in the overflow group with capacity N_{of} and arrival load ρ_v can be expressed as:

$$P_{BHA}(N_{of}, \rho_v, Z) \approx P_B(\frac{N_{of}}{Z}, \frac{\rho_v}{Z}) \tag{5.31}$$

where $Z = Var / M_e$ is the peakedness on the infinite group and P_B denotes the Erlang formula.

The theoretical basis for HA is the heavy traffic limit theorem. It states that, for a pure loss system with the assumptions of a renewal arrival process and exponential service times, the blocking experienced by arrivals only depends on the first two moments of the arrival traffic [*Guerin90*]. Simply speaking, the blocking probability by using HA is approximately given by scaling down for peaked traffic, and conversely, enlarging up for smooth traffic. Taking the waiting queue into consideration, by using the Laplace-Stieltjes transform (LST) method, an extension of HA is subsequently obtained in [*Guerin90*].

In particular, using the HA method, the arising problem is that x can be a nonintegral number in $P_B(x, \rho_v)$. In this case, the interpolation method can be used to overcome this difficulty. For example,

Let

$$f(x) = \ln B(x, \rho_v) \tag{5.32}$$

The first forward difference of f at x can be defined as:

$$\Delta f(x) = f(x+1) - f(x) \tag{5.33}$$

The second forward difference of f at x is:

$$\Delta^2 f(x) = \Delta f(x+1) - \Delta f(x) \tag{5.34}$$

Namely,

$$\Delta^2 f(x) = f(x+2) - 2f(x+1) + f(x) \tag{5.35}$$

According to the Newton's forward difference interpolation formula:

$$f(x+h) \approx p_n(x) = \sum_{j=0}^{n} \binom{h/\Delta h}{j} \Delta^j f(x) \tag{5.36}$$

If we assume that $\Delta h = 1$, then we have:

$$p_n(x) = \sum_{j=0}^{n} \binom{h}{j} \Delta^j f(x) = f(x) + h\Delta f(x) + \frac{h(h-1)}{2} \Delta^2 f(x) \tag{5.37}$$

If the nearest integral number of x is denoted as: $n = \lfloor x \rfloor$ and $h = x - n$, the abbreviations, which are used in the Erlang formula, are shown as:

$$B = B(n, \rho_v) \quad , \quad B_1 = B(n+1, \rho_v) \quad B_2 = B(n+2, \rho_v) \tag{5.38}$$

Hence,

$$f(x, \rho_v) = f(n+h, \rho_v) = f(n) + h\Delta f(n) + \frac{h(h-1)}{2} \Delta^2 f(n) \tag{5.39}$$

Substituting $f(n) = \ln B(n, \rho_v)$,

$$\ln B(x, \rho_v) = \ln B(n, \rho_v) + h\Delta \ln B(n, \rho_v) + \frac{h(h-1)}{2} \Delta^2 \ln B(n, \rho_v) \tag{5.40}$$

Finally,

$$B(x, \rho_v) = B^{(1-h)} B_1^{\,h} \left(\frac{B_2 B}{B_1^{\,2}}\right)^{\frac{h(h-1)}{2}} \tag{5.41}$$

Apart from the HA method, another important method is called the IPP [*Kuczu73*] and [*Kuczu79*]. The overflow stream can be approximated by a simple renewal process, which is alternately turned on for an exponentially distributed time and turned off for another exponentially distributed time. The three-moment matching method has proved to be sufficiently accurate for the overflow traffic [*Kuczu73*] and [*Kuczu79*]. In fitting the

parameters of IPP, the key points are to determine the arrival intensity λ and the on-off switch parameters ω and γ. As a special case, Kuczura's results only pertain to the assumption of average service rate equal to one [*Kuczu73*]. Here the details for the derivation of the IPP parameters under the general condition are provided in [*Appendix 2*].

In order to proceed with the analysis, we first need to use the recursive algorithm to determine the handoff arrival rate, which has been derived in Section §3.1. If the arrival calls find channels full, the blocking probability in lower layers can be easily shown as:

$$P_{C_1} = \frac{(\lambda_{v0} + \lambda_{h0})^{C_1} / ((\mu_v + \beta_n)^{C_1} C_1!)}{\sum_{i=0}^{C_1} \frac{(\lambda_{v0} + \lambda_{h0})^i}{(\mu_v + \beta_n)^i i!}} \tag{5.42}$$

Since new arrival calls and handoff calls are treated the same, the handoff failure probability is actually equal to the blocking probability: $P_{hf} = P_{C_1}$. If the CST can be represented by an exponential distribution and v is the handoff rate without considering handoff failure, the handoff proportion of voice calls can be shown as [*Lin94*]:

$$\lambda_{h0} = \frac{v(1 - P_{C_1})}{1 + vP_{C_1}} \lambda_{v0} \tag{5.43}$$

where $v = \beta_n / \mu_v$ and the handoff probability: $P_h = \frac{\beta_n}{\mu_v + \beta_n} = \frac{v}{1+v}$. Therefore we conclude that the bigger handoff rate leads to a bigger handoff probability. The total traffic offered load in a microcell is: $\rho_1 = \frac{\lambda_{v0} + \lambda_{h0}}{\mu_v + \beta_n}$. Apparently, the queue becomes stable when the summation of the working rates of the two types of calls does not exceed the total channel capacity in microcells. Namely,

$$\frac{\lambda_{v0} + \lambda_{h0}}{\mu_v + \beta_n} < C_1 \tag{5.44}$$

Similarly, we can have the expressions for the case of macrocells.

5.2.5 Modelling HCSs and Overflow Traffic Control

Since the properties of the overflow traffic are stochastically time-varying, the correlated arrivals need to be taken into account in overflow traffic modelling. Generally

speaking, such processes can be estimated by doubly stochastic processes. Because our objective is to have a solution for statistical purposes, we are interested in the arrival distribution at an arbitrary time and the degree of arrival correlation.

For an overflow traffic stream, it can be modelled as a MMPP [Figure 5.11].

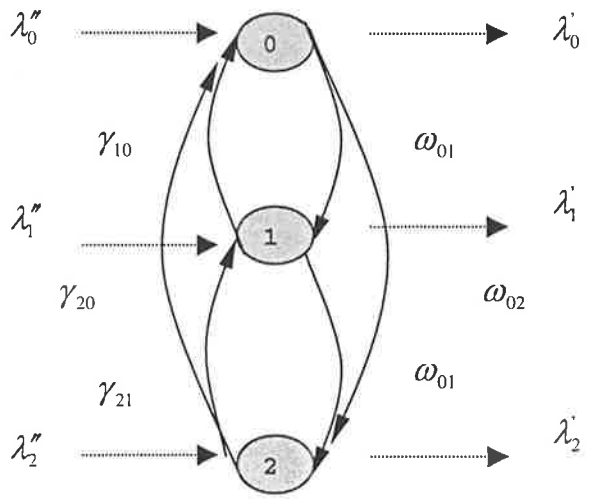


Figure 5. 11: A three-state MMPP for overflow traffic

Strictly speaking, MMPP is a Markov renewal process rather than a pure renewal process. In fact, MMPP is a doubly stochastic Poisson process, whose instantaneous arrival rate is a stationary random point process according to an irreducible m state continuous-time Markov chain. A MMPP can be represented by an infinitesimal generator Q and an arrival matrix by $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$ for the underlying Markov process. Specifically, when the MMPP is in the phase k $(1 \le k \le m)$, the arrivals will follow a Poisson process with a mean rate of λ_k. The reason for the use of MMPP in the overflow process is that the MMPP can capture the correlations between the subsequent arrivals. In particular, a MMPP not only can match the first three moments, but also can match any number of moments. On the contrary, hyperexponential interarrival times and batch Poisson process are found less appropriate [*Gusella91*].

If the overflow arrival streams from the microcells are all assumed to be independent, the superposition of n independent multi-state MMPP can be expressed by a Kronecker summation with the parameters (Q, Λ) [*Meier89*]:

$$Q = Q_1 \oplus Q_2 \oplus \cdots \oplus Q_n = Q_1 \otimes I_2 \otimes \cdots \otimes I_2 + I_2 \otimes Q_2 \otimes \cdots \otimes I_2 + I_2 \otimes I_2 \otimes \cdots \otimes Q_n$$

$$\Lambda = \Lambda_0 I_{2^n} + \Lambda_1 \oplus \Lambda_2 \oplus \cdots \oplus \Lambda_n = \Lambda_0 I_{2^n} + \Lambda_1 \otimes I_2 \otimes \cdots \otimes I_2 + I_2 \otimes \Lambda_2 \otimes \cdots \otimes I_2 + I_2 \otimes I_2 \otimes \cdots \otimes \Lambda_n \tag{5.45}$$

where $\oplus$ denotes the Kronecker summation and $\otimes$ represents the Kronecker product. Λ_0 is an arriving matrix for new calls, I_2 is an identity matrix of order 2 and I_{2^n} is an identity matrix of order 2^n for a two-state MMPP,

We can find that Q has 2^n states for a two-state MMPP. Apparently, if the number of n becomes large, the states will dramatically increase and then performance evaluation will become difficult. This is the reason that a two-state MMPP is approximated by a simple renewal process called Interrupted Poisson Process (IPP) in this study. In fact, the use of IPP or hyperexponential distribution is the special case of MMPP. The advantage of IPP is that it can still capture the peakedness of the arrival process and remain tractable. Instead of MMPP, the use of IPP can result in both the reduction of computer simulation time and the simplicity of analytical study [*Kuczu73*]. This approximate method is to match the first three moments of the overflow process. Note that this approach is exact except for the use of IPP approximation.

As mentioned above, a two-state IPP can be represented by:

$$Q_g = \begin{bmatrix} -\omega & \omega \\ \gamma & -\gamma \end{bmatrix} \qquad \Lambda_g = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix} \tag{5.46}$$

where λ is the arrival rate at an on-off random switch.

The mean arrival rate λ_T towards the overflow group can be calculated by:

$$\lambda_T = (\lambda_{v0} + \lambda_{h0}) P_{C_1} \tag{5.47}$$

Since we assume that the initial offered load to the microcells is ρ_1, the next step is to have the solution for the mean and variance in the infinite servers group. The computation

of the mean M_e and variance Var can be obtained from the global balance equations, in which the state space is shown in [Figure 5.12].

If we only consider the superposition of n IPP, the dimension of the matrix can significantly reduce from 2^n to $(n+1)$. This can significantly reduce the complexity in computation. The superposition of n IPP can be parameterized by (Q_q, Λ_q) in the n microcells.

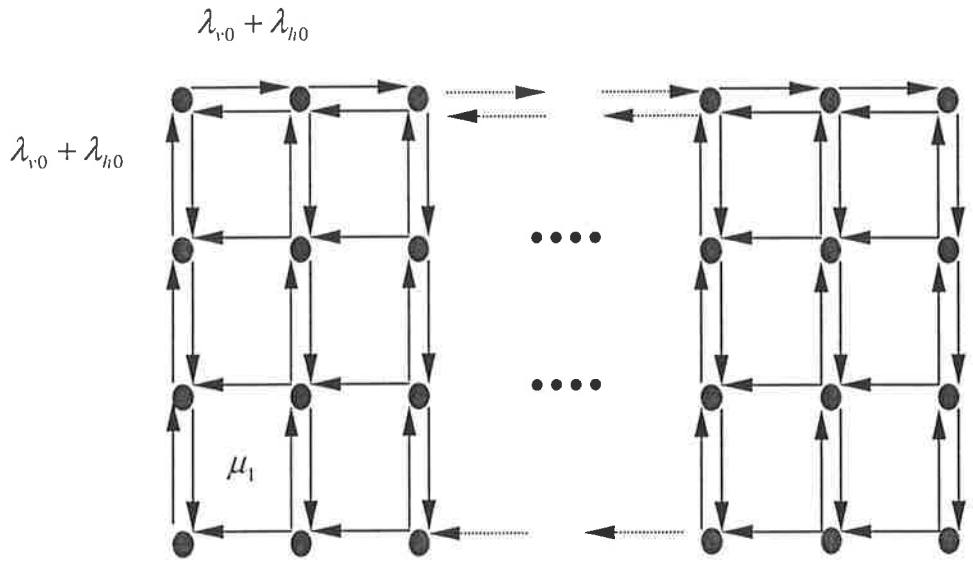


Figure 5. 12: The state space for overflow traffic

Then we have $\Lambda_q = diag(0, \lambda, 2\lambda, \ldots, n\lambda)$. Meanwhile, the infinitesimal generator for the superposition can be written as:

$$\begin{aligned} \{Q_q\}_{i,i} &= -(n-i)\omega - i\gamma \quad , \quad 0 \le i \le n \\ \{Q_q\}_{i,i+1} &= (n-i)\omega \quad , \quad 0 \le i \le n-1 \\ \{Q_q\}_{i,i-1} &= i\gamma \quad , \quad 1 \le i \le n \end{aligned} \tag{5.48}$$

If we use j to represent the number of overflow calls at macrocell servers, which is from the subordinate microcells, and i for the number of calls at the destination, the state

space of a finite two-dimensional Markov process can be represented by $\Omega = \{(i, j), 0 \le j \le n, 0 \le i \le C_2 + K\}$.

The infinitesimal generator Q for the state space can be written as:

$$Q = \begin{bmatrix} Q_q - \Lambda & \Lambda & & & & & & \\ U & Q_q - \Lambda - U & \Lambda & & & & & \\ & 2U & Q_q - \Lambda - 2U & \Lambda & & & & \\ & & \cdots & \cdots & & & & \\ & & & C_2U & Q_q - \Lambda - C_2U & \Lambda & & \\ & & & & U_{q1} & Q_q - \Lambda_{q1} - U_{q1} & \Lambda_{q1} & \\ & & & & & \cdots & \cdots & \\ & & & & & & U_{qK} & Q_q - U_{qK} \end{bmatrix} \tag{5.49}$$

Subsequently, the sub-matrices for arrivals can be obtained. Note that Λ and Λ_{qk} are all with dimensions of $(n+1)\times(n+1)$. Then we can have:

$$\begin{aligned} \Lambda &= diag(\lambda_1, \lambda + \lambda_1, 2\lambda + \lambda_1, \ldots, n\lambda + \lambda_1) \\ \Lambda_{qk} &= diag(0, \lambda, 2\lambda, \ldots, n\lambda) \end{aligned} \tag{5.50}$$

where $\lambda_1 = \lambda_{v1} + \lambda_{h1}$ and $1 \le k < K$.

Similarly, the submatrices U and U_q for services can be obtained as well. They all have the dimensions of $(n+1)\times(n+1)$. Specifically,

$$\begin{aligned} U &= diag(\mu_v + \beta_M, \mu_v + \beta_M, \ldots, \mu_v + \beta_M) \\ U_{qk} &= diag(C_2(\mu_v + \beta_M), C_2(\mu_v + \beta_M), \ldots, C_2(\mu_v + \beta_M)) \end{aligned} \tag{5.51}$$

where $1 \le k \le K$.

The equilibrium of the Markov chain can be found from the calculation of the stationary vector π_∞:

$$\pi_\infty Q = 0 \quad , \quad \pi_\infty e = 1 \tag{5.52}$$

where $\pi_\infty = [\pi_\infty^{(0)}, \pi_\infty^{(1)}, \ldots, \pi_\infty^{(C_2)}, \ldots, \pi_\infty^{(C_2+K)}]$ and $e = [1, 1, 1, \ldots, 1, \ldots, 1]^T$.

Finally, we can use Gauss-Seidel iteration to compute the equilibrium probability successively.

5.2.5.1) Performance Measures

After the equilibrium probability is obtained, the parameters of interest for this process can be easily derived.

1) If the disposition policy is taken into consideration, the equilibrium probability of the number of calls in overflow destination x_i can be shown by:

$$x_i = \pi_\infty^{(i)} e \quad while \quad 0 \le i \le C_2 + K \tag{5.53}$$

2) The probability of the jth arrival finding i calls in the system can be expressed by:

$$y_i(j) = \frac{\pi_\infty^{(i)}(j)\Lambda(j)e}{\lambda_{tot}} \quad 1 \le j \le n, \quad 0 \le i \le C_2 + K \tag{5.54}$$

where $\lambda_{tot} = \sum_{i=0}^{C_2+K} \pi_\infty^{(i)} \Lambda e$.

4) The blocking probability seen by the jth arrival can be obtained by:

$$P_B(j) = \frac{\pi_\infty^{(C_2+K)}(j)\Lambda(j)e}{\lambda_{tot}} \quad while \quad 1 \le j \le n \tag{5.55}$$

While the disposition policy is not taken into account, we can have the blocking probability seen by all the arriving overflow streams (from $j = 0$ to $j = n$) as:

$$P_{Blocking} = \sum_{j=0}^{n} P_B(j) = \sum_{j=0}^{n} \frac{\pi_\infty^{(C_2)}\Lambda(j)e}{\lambda'_{tot}} \tag{5.56}$$

Namely,

$$P_{Blocking} = \frac{\sum_{j=0}^{n} j\pi_\infty^{(C_2)}(j)}{\sum_{i=0}^{C_2}\sum_{j=0}^{n} j\pi_\infty^{(i)}(j)} \tag{5.57}$$

4) The probability $q(j)$ that the *jth* arrival has to wait in the queue can be represented by:

$$q(j) = \sum_{i=C_2}^{C_2+K-1} y_i(j) \tag{5.58}$$

The mean waiting time density seen by the *jth* arrival is:

$$p_k(j) = \frac{y_{C_2+k-1}(j)}{\sum_{i=C_2}^{C_2+K-1} y_i(j)} \quad while \quad 1 \le k \le K \tag{5.59}$$

Due to the waiting time distribution is a mixture of Erlang distribution with mixing density $p(j)$ [*Meir89*], this can be represented by a phase-type service distribution with parameters $(\alpha(j), T(j))$:

$$T(j) = \begin{bmatrix} -C_2\mu_2 & C_2\mu_2 & & & \\ & -C_2\mu_2 & C_2\mu_2 & & \\ & & \cdots & \cdots & \\ & & & -C_2\mu_2 & C_2\mu_2 \\ & & & & -C_2\mu_2 \end{bmatrix} \tag{5.60}$$

$$\alpha(j) = (p_k(j),\ p_{k-1}(j), \cdots,\ p_1(j)) \tag{5.61}$$

where $\mu_2 = \mu_v + \beta_M = \mu_v + (2V/\pi R_c)$.

Therefore, according to the description of the phase type distribution in [*Neuts80*], we can obtain the waiting time distribution eventually.

5.2.5.2) Results Discussion

1) The Results of MMTs

In this study, if we assume that there is neither arrival nor mobility in the macrocell, we can have the results, which are obtained by using the one-moment, the HA and the three-moment methods as shown in [Figure 5.13]. In this study, we assume the cell radius is $R_c = 1$ km and terminal movement speed is $V = 30$ km/hour. The average call duration is

$1/\mu_v = 120$ seconds. The number of channels in lower layer is $C_1 = 7$ and $C_2 = 3$ in upper layer. The number of finite buffers is $K = 3$.

From the figure, we observe that the use of the IPP three-moment method yields the best result in terms of accuracy for the overflow call blocking probability. On the contrary, compared with using the three-moment method, the use of either the one-moment method or the HA method underestimates the blocking probability. By using the HA method, because the ratio of overflow traffic loads to the peakedness (a/Z) is usually quite small, this leads to the arrival loads in subgroups becoming smaller [*Fred80*]. Therefore, it is hard to maintain a high correlation under a given distribution. More importantly, higher mobility causes higher overflow call blocking probability because of the increased handoff rate.

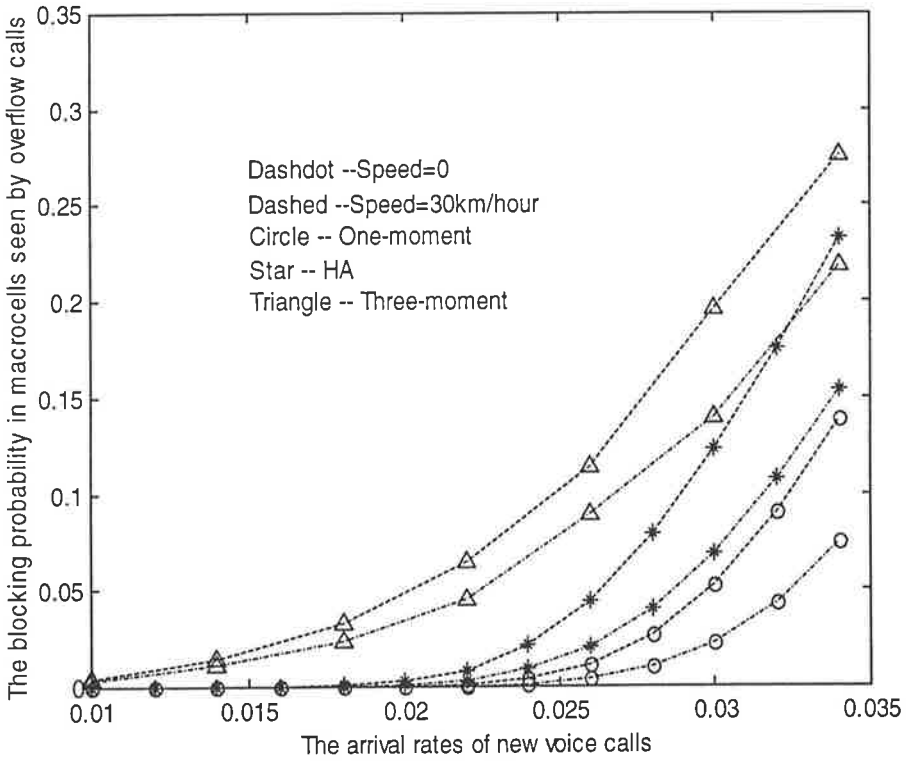


Figure 5. 13: The blocking probability in macrocells seen by overflow traffic

In particular, compared with the one-moment and two-moment estimations used in [*Kim98*], we also observe that there is no discernible advantage in using the two-moment estimation over the one-moment estimation. Therefore, we may have to adopt the three-moment method as the final solution, which is consistent with the results in [Kuczu73]. In

heavy load, we also find that the performance of these three methods tends to converge with the increase of traffic loads. The big difference lies only in the light load and middle load regions. In particular, if the ratio (a/Z) increases, the accuracy of the HA improves. However, this does not pertain to the cellular networks due to the required stringent low blocking probability.

Finally, we conclude that the number of channels in macrocells must increase in order to maintain a given QoS while overflow traffic is more bursty than a Poisson process. Secondly, the higher terminal mobility tends to cause higher overflow call blocking probability, especially for the high traffic load condition.

2) Disposition Policy Results

In this study, we allow the overflow traffic to have an alternative waiting choice when the channels are found to be full in macrocells. In order to distinguish the blocking probability in microcells, we define the rejection proportion of the overflow traffic, which finds the finite buffers full, as loss probability. Then we can solve the equilibrium probability and compute the loss probability P_{loss} as:

$$P_{loss} = \frac{\sum_{j=0}^{n} j\pi_{\infty}^{(C_2+K)}(j)}{\sum_{i=0}^{C+K}\sum_{j=0}^{n} j\pi_{\infty}^{(i)}(j)} \tag{5.62}$$

We have the result as shown in [Figure 5.14].

For example, by using the disposition policy, the blocking probability of an overflow call decreases from 14% to 2% for a still terminal while $\lambda_{v0} = 0.03$ calls/s. This improvement is so significant that the handoff call failure rate can be reduced eventually.

For the terminals with speed $V = 30$ km/hour, we have the results as shown in [Table 5.1]. From the table, we observe that the percentage improvement reduces as the ratio (a/Z) increases. However, this still shows quite a high percentage of improvement even as the arrival rate becomes high. For example, while $\lambda_{v0} = 0.034$ calls/s, the percentage improvement is 66%. If the arrival is $\lambda_{v0} = 0.03$ calls/s, the percentage improvement becomes 77%.

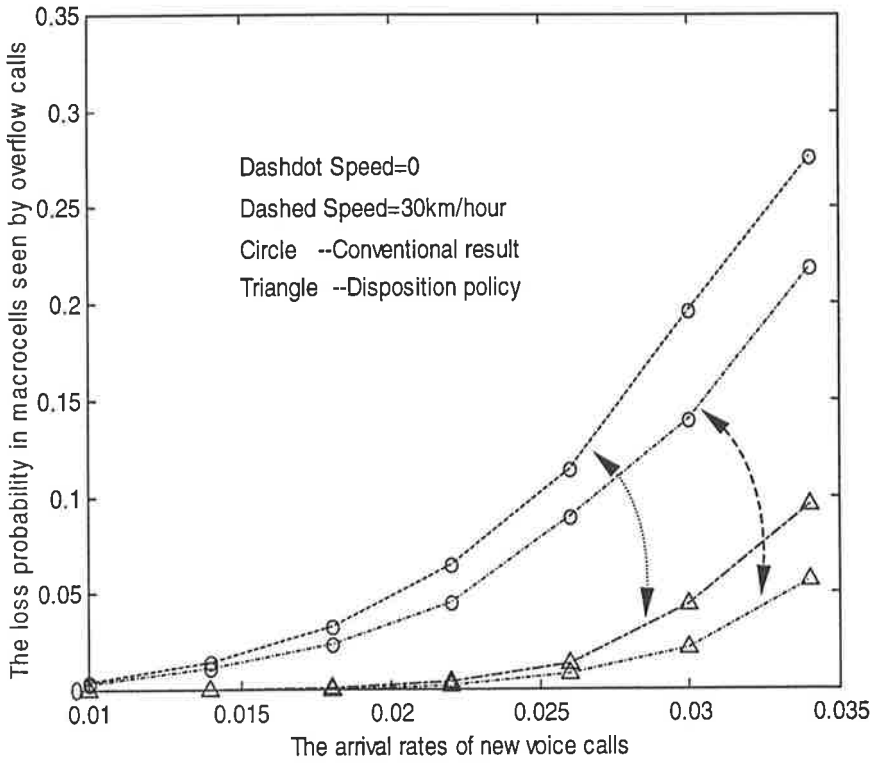


Figure 5. 14: The loss probability for the disposition policy

Apparently, the prices for trading-off this improvement are the additive buffer mechanisms, the prolonged service, and the probable increased signalling loads.

Table 5. 1: The comparison of overflow call handling schemes

a/Z	P_B	P_{loss}	Percent
0.0004	0.0036	0	100%
0.003	0.0145	0.0002	99%
0.0125	0.0325	0.001	97%
0.0351	0.0645	0.004	94%
0.0755	0.1143	0.0137	88%
0.1361	0.1959	0.0446	77%
0.2171	0.2759	0.0951	66%

3) HCS Network Optimisation

It is known that the main challenge in designing a HCS network is how to optimise the resource among different layers. A good design must be able to perform at an optimal cost. Before the actual networks rollout, network planning needs to take the minimum cost into account apart from satisfying the QoS in a given service area. Although many studies have been conducted for network planning in the areas of cell coverage, channel assignment and propagation, relatively few studies concentrate on the network planning in terms of cost-saving system design. In particular, for the HCS system, although the optimisation of the underlaid network has been studied in [*Mihai97*], the overlaid network is not considered. In this study, given the required QoS and offered loads per square kilometre, we can determine the optimal number of base stations in HCSs. The procedure is summarised as follows.

- **For the underlaid systems,** we can use the first algorithm to optimise the relationship between the number of base stations and offered loads in either microcells or macrocells, in which the overflow traffic is excluded. The number of channels in microcells is assumed to be C_1. The required call blocking probability is $P_{C_1}^{(d)}$ and the handoff forced termination probability is $P_{F_1}^{(d)}$.

Algorithm 5.2:

% Microcells or macrocells with terminal mobility

1) Initialise $C_1 = C_{\min}$.
2) While $C_1 < C_{\max}$

   Set the number of microcells as: $n_0 = 1$.

   While $P_{C_1} < P_{C_1}^{(d)}$ and $P_{F_1} < P_{F_1}^{(d)}$

   Increase the number of microcells: $n_0 = n_0 + 1$.

   Compute the new microcell coverage area: A.

   Compute the traffic load in each cell: $\rho^{(new)} = \rho^{(old)} / n_0$.

   Compute the new parameters: $P_{C_1}^{(new)}$, $P_{F_1}^{(new)}$.

   End.

Reset $C_1 = C_1 + 1$.

End.

3) Collect data.

Then the cell density versus the number of channels per cell can be obtained.

- **For the overlaid systems,** we can use the second algorithm to optimise the HCS network, in which the overflow traffic is taken into account. The number of channels in microcells is assumed to be C_1 and the overflow call blocking probability is P_{req}.

Algorithm 5.3:

% Microcells with terminal mobility

1) Initialise

$\lambda_{v0} = \lambda_{\min}$.

2) While $\lambda_{v0} < \lambda_{\max}$

Set the number of channels in a microcell as: $C_1 = C_{\min}$.

Compute the overflow call blocking probability $P_{Blocking}$.

While $P_{Blocking} > P_{req}$

Increase the number of channels in a microcell: $C_1 = C_1 + 1$.

Compute the new parameters: $P_{Blocking}^{(new)}$.

End.

Reset $\lambda_{v0} = \lambda_{v0} + d$.

End.

3) Collect data.

Finally, if the required overflow blocking probability is assumed to be 1% and the cell radius is $R_c = 1$ km, the number of channels versus voice call arrival rates is shown in [Table 5.2].

As a conclusion, we observe that, in order to maintain a required QoS for the overflow calls, the number of channels in microcells needs to keep increasing according to the

increase of new voice arrivals. Meanwhile, low mobility traffic requires fewer channels in the microcells in order to reach the required QoS.

Table 5. 2: The number of optimal channels in microcells

λ_{v0} $(calls/s)$	$V=0$ (km/hr)	$V=15$ (km/hr)	$V=30$ (km/hr)
0.008	4	5	6
0.009	4	5	6
0.01	5	6	6
0.011	5	6	7
0.012	6	7	7
0.013	7	7	8

5.2.6 Summary

In this study, a HCS queuing system is analysed by a two-state MMPP, which is subsequently approximated by an IPP random switch model. The HA method is proposed to approximately analyse the burstiness of overflow traffic. The main contributions of this section are the comparison of three different MMTs and a proposal for a call handling scheme of overflow traffic. From the results, we observe that, compared with the three-moment method, using the one-moment method yields underestimated results. Moreover, the use of the HA method underestimates the performance of overflow traffic as well. Although the overflow process can be approximately estimated by the lower moment methods, this may give rather too optimistic results because the correlative traffic property is not well handled. As a main conclusion, in order to obtain an accurate result, the overflow traffic needs to be modelled with higher moment methods.

In addition, terminal mobility can increase the overflow blocking probability while terminal speed becomes high. The reason lies in the increase of handoff rate due to the terminal movement. More importantly, the use of disposition policy can not only reduce

the blocking probability of overflow calls, but can also offer flexibility to overflow traffic handling. In network optimisation, higher terminal movement will result in an increased optimal number of channels in microcells in order to maintain the required QoS of overflow traffic.

Chapter 6

Conclusions and Future Work

Since personal wireless communications are resource-limited, the design of future wireless multimedia systems must be centred on two categories: **optimal resource allocation** and **performance enhancement**. The realisation of optimal resource allocation must be through the efficient analysis of traffic integration, while traffic integration needs to use effective call control procedures. In this dissertation, we have conducted a comprehensive study for multi-service and multi-layer systems, especially for the performance of high-speed data (HSD) services within time channels. The study of HSD can be used to estimate the performance of popular future data applications, such as wireless access to the Internet or real time video services.

The main contribution of this dissertation is that we have developed some theoretical methodologies for efficient performance analysis and some new call control procedures to enhance data performance with rigidly specified QoS, including both traffic variability and handoff priority characteristics. It has been found that, firstly, the MAMs are robust enough to be applied in analysing both TDMA and R-TDMA integrated services, especially for integrating HSD services into voice services. Secondly, optimal resource allocation for integrated services largely depends on packet data traffic statistics, while data performance is subject to voice traffic. In particular, terminal mobility has a significant effect on call performance under high traffic load conditions. Finally, traffic performance in wireless multimedia services can be enhanced through the enforcement of effective call control procedures, such as the multiple priority control schemes in various systems.

6.1 Dissertation Summary

In the study of traffic performance for multimedia services, there exist two fundamental criteria, i.e., guaranteeing the required QoS for real time traffic and providing best effort QoS for non-real time traffic. **Traffic integration** and **global reach** are seen as key factors in the successful delivery of wireless multimedia communications. Future communication platforms are expected to support a mix of voice, data, and video source traffic as well as providing ubiquitous and tetherless access.

The study of traffic integration is subject to the use of diverse traffic source models and channel access methods. In this dissertation, performance analysis has been conducted using three methodologies, i.e., source traffic modelling, analytical modelling and simulation studies. First, traffic performance for integrating low rate data services into voice service is analysed by a quasi-static approximation method. Subsequently, due to VBR transmission forming the core for the successful delivery of multimedia services, the HSD rate integrated services, either circuit access or packet access, have been analysed by a systematically efficient MAM. Meanwhile, integrated services using the reservation protocol, which can be applied in a short-range radio environment, are analysed by the proposed MAM. Next, a design benchmark in DS-CDMA integrated services is proposed, while call control schemes are adopted to smooth interference fluctuation. The packet delay is numerically analysed by a marginally distributed method. Finally, the investigation of HCSs for bursty overflow traffic, which can be used for the global access of space segment, is contrasted by different moment matching techniques.

The details of the conclusive results can be found at the end of each section or subsection. Specifically, for a single voice service, the use of adaptive radio resource management, which includes combining traffic variability and handoff priority schemes, is investigated in Section §3.1. It shows that, with increased terminal speed, call incomplete probability increases accordingly. In particular, such an effect becomes significant with increased traffic loads. Faster terminal mobility actually reduces the chance of call waiting. In addition, increased reserved channels result in increased incomplete probability. As a suggestion, better radio resource management can be achieved once the individual task is taken into overall consideration. In Section §3.2, we observe that high channel utilisation can be obtained by multiplexing packet-based services into circuit mode services because of the busty nature of multimedia services. Packet performance can be enhanced by the use of priority control schemes and is determined by the arrival loads, average length of message as well as the voice channel condition, which is affected by voice terminal mobility.

In Section §4.1, call performance in HSD circuit access is largely affected by the use of handoff priority schemes while new call blocking probability is found to be insensitive to terminal mobility. Subsequently, we suggest that a hybrid reservation scheme can be adopted to optimise handoff performance. For data packet access as discussed in Section §4.2, resource allocation for packet data is significantly different from that for circuit mode

data services and the behaviour of data performance is subject to input traffic statistics. In particular, larger packets are able to cause longer latency. Terminal mobility and data message length have a decisive impact on resource allocation for each type of traffic, especially under high traffic load conditions. Additionally, variable voice coding schemes can be used to enhance data performance without degrading voice services. For a reservation protocol in Section §4.3, NBS is found to have the best performance of the three channel access strategies, i.e., FBS, MBS and NBS. Moreover, a benchmark parameter for data packets is derived numerically. The waiting time distribution is found to be largely dependent upon the input traffic loads, in particular, the ratio of call duration to packet message length. Moreover, the CAC proposed in our study can be enforced to enhance mixed traffic performance. The multiple priority scheme is a promising technique to ease the data packet congestion condition for traffic integration without the loss of data traffic preference with either reservation or non-reservation protocols. In Section §5.1, an optimal threshold design method is developed for the single voice service so as to satisfy both the required loss and blocking probabilities. For the integrated services, call control schemes are used to smooth interference fluctuation for maximising system capacity. Subsequently, the delay constraint for data packets is numerically derived. In Section §5.2, the disposition call handling scheme in HCS systems is used to reduce loss probability for overflow traffic, especially under a heavy traffic load condition. Meanwhile, the HA method is proposed to analyse the burstiness of overflow traffic. It is found that the HA has a better estimation for overflow traffic than the one-moment matching method. However, by contrasting different moment matching techniques, the result favours the use of the higher moment method. For network optimisation, higher movement will result in an increased optimal number of channels in lower layers in order to maintain the required QoS of overflow traffic.

6.2 Future Work

It is apparent that the study of wireless multimedia services is a tremendous task in PWC. However, the future investigation can be advanced with the following factors:

Firstly, in this dissertation, although we have found that multimedia applications have diverse bursty source models in Chapter 3, we assume a generic traffic source model for data applications. From some recent studies, the actual interarrival times of packets may

not be well confined to the categories of exponential distribution due to their bursty properties. This can happen in interactive data applications with a strongly bursty packet arrival process, such as reading short e-mail messages, transferring large files between the mobile and a stationary host, or downloading WWW contents. Under these circumstances, the conventional Markov chain assumption may not hold and therefore the more realistic distribution functions, such as the Pareto distribution with heavy tail, need to be considered. If the actual traffic has burstiness over a wide range of time scales, other long-range dependent traffic modelling methodologies, e.g., statistically self-similar processes, need to be applied. Because the queuing performance in a long range dependent process is different from that in a short range dependent process, different assumptions for source traffic models may lead to different results. In addition, it is worthwhile noting that, if the assumption of interarrival distribution is relaxed, the previous analysis of virtual waiting time will become invalid because the Poisson Arrivals See Time Averages (PASTA) principle does not hold. Instead, more advantageous techniques like general arrivals will need to be considered. Moreover, further study is required to integrate video traffic modelling into voice services due to the peculiar properties of video traffic. Therefore, it is recommended to conduct further study for the integration of realistic source traffic models, especially at the range of 10 Mbps in a WLAN multimedia environment.

Secondly, it is established that MAMs are robust enough to analyse the performance of HSD integrated services. Taking the bursty source traffic models into account, the generalised MAMs need to be further developed so as to be applicable to the multimedia environment. Moreover, it is worthy to mention that MAMs are imperfect and have some drawbacks as well. For example, the use of MAMs usually yields numerical results rather than closed-form expressions. In addition, for HSD services, the issues of power consumption and network planning need to be taken into account in actual implementation.

Next, for the study of R-TDMA systems, e.g., the PRMA system, an errorless channel is assumed. In other words, collisions resulting in dropped packets are regarded as the only source for degraded performance in a perfect channel. However, the study for the inclusion of channel characteristics is of importance because the channel errors may cause the loss of reserved slots and higher dropping probability. Note that the analysis above for each data call is based upon the assumption of one slot per frame and a single radio

channel. Because PRMA can also transmit multiple packet data messages, the problems of non-contiguous multislots per frame and the integration of video source traffic are worth further investigation.

Furthermore, we deal only with the CAC of voice and low rate data packets under the assumption of perfect power control in DS-CDMA systems. Taking the volatile fading radio environment into consideration, it is known that imperfect power control can lead to a significant capacity reduction. The tradeoff relationship between high rate data integrated services and users' less preferable quality still remains unclear under time-varying propagation and imperfect power control situations. In order to provide for Internet and multimedia services, HSD services must be supported in DS-CDMA systems. Therefore, future work may focus on the effectiveness of the CAC for HSD services under an imperfect channel condition. In addition, the investigation of overlaying CDMA systems may open up another possibility for further research. As for the HCS system, the study of multiservice and interference needs to be taken into account as well.

Lastly, it is worth mentioning that the random access delay in traffic modelling study accounts for a small proportion of the overall packet delay. The addition of higher layer signalling, such as in TCP/IP networks, may significantly contribute to the overall systematic performance in multimedia services. This needs to be taken into account while a system needs to be optimally designed.

References

[Abram94] N. Abramson, "Multiple access in wireless," Proceedings of the IEEE, vol. 82, no. 9, pp.1360-1370, September 1994.

[Anag96] M. E. Anagnostou, J. A. Sanchez-P and I. S. Venieris, "A multiservice user descriptive traffic source model," IEEE Transactions on Communications, vol. 44, no. 10, pp1243-1246, October 1996.

[Bhat76] U. N. Bhat and M. J. Fischer, "Multichannel queueing systems with heterogenous classes of arrivals," Nav. Res. Logist. Quart., vol. 23, pp. 271-282, 1976.

[Brady65] P. T. Brady, "A technique for investigating on-off patterns of speech," Bell System Technical Journal, vol. 44, pp 1-22, January 1965.

[Brady69] P. T. Brady, "A model for generating on-off speech patterns in two-way conversations," Bell System Technical Journal. Vol.48, pp 2445-2472, September1969.

[Brasch97] G. Brasche and Bernhard Walke, "Concepts, Services, and protocols of the new GSM phase 2+ General Packet Radio Service," IEEE Communications magazine, pp. 94-104, August 1997.

[Bright96] L. W. Bright, "Matrix-analytic methods in applied probability," Ph. D thesis, Adelaide University, 1996.

[Cai97] J. Cai and D. J. Goodman, "General Packet Radio Service in GSM," IEEE Communications Magazine, pp. 122-131, Oct. 1997.

[Calin97a] D. Calin and D. Zeghlache, "Priority queuing analysis for voice-data integration in wireless PCS," Proceedings of Multiaccess, Mobility and Teletraffic advance in wireless networks, pp. 273-285, Melbourne, December 15-17, 1997.

[Calin97b] D. Calin and D. Zeghlache, "Performance analysis of high speed circuit switched data (HSCSD) over GSM," Proceedings of Multiaccess, Mobility and Teletraffic advance in wireless networks, pp. 273-285, Melbourne, December 15-17, 1997.

[Calin97c] D. Calin and D. Zeghlache, "Performance and handoff analysis of an integrated voice-data cellular system," IEEE International symposium on personal indoor and mobile radio communications, vol. 2, pp. 386-390, 1997.

[Calin98] D. Calin and D. Zeghlache, "High speed circuit switched data over GSM: potential traffic policies," IEEE Proceedings of Vehicular Technology Conference, pp.1810-1814, Ottawa, Canada, May 18-21, 1998.

[Calleg95] F. Callegati, C. Carciofi, M. Frullone, P. Grazioso and G. Riva, "Call admission control for multi-service packet switched cellular mobile radio system," IEICE Transactions on Communications, vol. E78-B, no. 4, pp. 504-513, April 1995.

[Chang94a] C-J Chang and C-H Wu, "Slot allocation for an integrated voice/data TDMA Mobile radio system with a finite population of buffered users," IEEE Transaction on Vehicular Technology, vol. 43, no. 1, pp. 21-26, February 1994.

[Chang94b] C. J. Chang, T.T. Su and Y. Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping," IEEE/ACM Transactions on Networking, vol. 2, no.2, pp. 166-175, April 1994.

[Cox95] D. C. Cox, "Wireless Personal Communications: What is it," IEEE Personal Communications Magazine, pp. 20-35, April 1995.

[Daigle92] J. N. Daigle and N. Jain, "A queuing system with two arrival streams and reserved servers," Proceedings of IEEE INFOCOM, vol.2, pp.2161-2167, 1992.

[Dailge91] J. N. Daigle and D.M. Lucantoni, "Queuing systems having phase-dependent arrival and service rates," Numerical Solution of Markov Chains, William J. Stewart, Ed., pp. 161-202, Marcel Dekker Inc., New York, 1991.

[Das97] K. Das and S. D. Morgera, "Interference and SIR in integrated voice/data wireless DS-CDMA networks- a simulation study," IEEE Journal on Selected Areas in Communications, vol. 15, no. 8, pp. 1527-1538, October, 1997.

[Dahl98] E. Dahlman, Bjorn Gudmunson, Mats Nilsson and Johan Skold, "UMTS/IMT-2000 based on wideband CDMA," IEEE Communications Magazine, pp. 70-80, September 1998.

[ETSI234] ETSI, "GSM 02.34 High speed circuit switched data (HSCSD), Stage1," version 5.2.0, July 1997.

[ETSI334] ETSI, "GSM 03.34 High speed circuit switched data (HSCSD), Stage2," version 5.0.0, April 1997.

[ETSI360] ETSI, "GSM 03.60 General Packet Radio Service: Service Description," version 1.10, February 1997.

[ETSI364] ETSI, "GSM 03.64 Overall description of the General packet radio service (GPRS) radio interface, Stage 2," v.5.0.0, July 1997.

[ETSI502] ETSI, "GSM 05.02 Multiplexing and multiple access on the radio path," Fifth edition, July 1998.

[Evans67] R. V. Evans, "Geometric distribution in some two-dimensional queueing systems," Operat. Res., no.15, pp. 830-846, 1967.

[Evans99a] J. S. Evans and D. Everitt, " Effective bandwidth-based admission control for multiservice CDMA cellular networks," IEEE Transactions on Vehicular Technology, vol. 48, no. 1, pp.37-46, January 1999.

[Evans99b] J. S. Evans and D. Everitt, "On the teletraffic capacity of CDMA cellular networks," IEEE Transactions on Vehicular Technology, vol. 48, no.1, pp. 153-165, January 1999.

[Everitt83] D. E. Everitt and N. W. Macfadyen, "Analysis of multicellular mobile radio telephone systems with loss," British Telecom Technology Journal, vol. 1, no.2, pp.37-45, 1983.

[Everitt89a] D. Everitt, "Product form solutions in cellular mobile communication systems," In the Proceedings of Fourth Australian Teletraffic Research Seminar, Bond University, pp. 1-8, December 1989.

[Everitt89b] D. Everitt and D. Manfield, "Performance analysis of cellular mobile communication systems with dynamic channel assignment," IEEE Journal on Selected Area in Communications, vol. 7, no. 8, pp. 1172-1180, October 1989.

[Everitt94] D. Everitt, "Traffic engineering of the radio interface for cellular mobile networks," Proceedings of the IEEE. Vol. 82, No. 9, pp.1371-1381, September 1994.

[Falk83] G. Falk, J. S. Groff, W. C. Milliken, M. Nodine, S. Blumenthal and W. Edmond, "Integration of voice and data in the wideband packet satellite network," IEEE Journal on Selected Areas in Communication, vol. 1, no. 6, pp.1076-1083, 1983.

[Fang99] Y. Fang and I. Chlamtac, " Teletraffic analysis and mobility modeling of PCS networks," IEEE Transaction on Communications, vol.47, no.7, pp. 1062-1072, July 1999.

[Feld82] R. M. Feldman, "A note on a computational model for a data/voice communication queuing system," Naval Research Logistics Quarterly, vol. 29, no. 3, pp. 529-534. 1982.

[Fisch76] M. J. Fischer and T. C. Harris, "A model for evaluating the performance of an integrated circuit- and packet-switched multiplex structure," IEEE Transactions on Communications, vol. 24, no. 2, pp. 195-202, Feb 1976.

[Fisch77] M. J. Fischer, "A queuing analysis of an integrated telecommunications system with priorities," INFOR., vol.15, no. 3, pp. 277-288, 1977.

[Fisher79] M. J. Fisher, "Data performance in a system where data packets are transmitted during voice silent periods-single channel case," IEEE Transactions on Communications, vol. 27, no. 9, pp. 1371-1375, September 1979.

[Fred80] A. A. Fredericks, "Congestion in blocking systems - a simple approximation technique," The Bell System Technical. Journal, vol. 59, no. 6, pp.805-827, July-Aug. 1980.

[Frull94] M. Frullone and G. Riva and P. Garzioso and M. Missiroli, "Comparisons of multiple access schemes for performance communication systems in a mixed cellular environment," IEEE Transactions on Vehicular Technology, vol. 43, no.1, pp. 99-109, February 1994.

[Fuka96] A. Fukasawa, et al. "Wideband CDMA system for personal Radio Communications," IEEE Communications Magazine, pp116-123. October 1996.

[Fukuda83] A. Fukuda and S. Tasaka, "The equilibrium point analysis-a unified analytic toll for packet broadcast networks," In the Proceedings of GLOBECOM'83, pp. 33.4.1-33.4.8, 1983.

[Garcia82] A. Leon-Garcia, R. H. Kwong and G. F. Williams, "Performance evaluation methods for an integrated voice/data link," IEEE Transaction on Communications, vol. 30, no. 8, pp. 1848-1858, August 1982.

[Gaver82] D. P. Gaver and J. P. Lehoczky, "Channels that cooperatively service a data stream and voice messages," IEEE Transactions on Communications, vol. 30, no. 5, pp. 1153-1162, May 1982.

[Gilhou91] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, Jr., and C.E. Wheatley III, "On the capacity of a cellular CDMA system," IEEE Transactions on Vehicular Technology, vol. 40, no. 2, pp. 303-312, May 1991.

[Gimpel65] L. A. Gimpelson, "Analysis of mixtures of wide- and narrow-band traffic," IEEE Transactions on Communication Technology, vol. 13. no. 3, pp. 258-266. 1965.

[Girard90] A. Girard, "Routing and dimensioning in circuit-switched networks," Addison Wesley. 1990.

[Gned68] B. V. Gnedenko and I. N. Kovalenko, "Introduction to queueing to theory," Israel Program for Scientific Translations, Jerusalem, 1968.

[Goodm89] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard and B. Ramamurthi, "Packet Reservation Multiple Access for local wireless communications," IEEE Transactions on Communications, vol. 37, no. 8, pp.885-890, Aug. 1989.

[Goodm90] D. J. Goodman, "Cellular packet communications," IEEE Transaction on Communication, vol. 38, no. 8, pp.1272-1280, Aug. 1990.

[Goodm91] D. J. Goodman and S. X. Wei, "Efficiency of packet reservation multiple access," IEEE Transactions on Vehicle Technology, vol. 40, no.1, pp. 170-176, Feb. 1991.

[Grieco94] D. M. Grieco and Donald L. Schilling, "The capacity of Broadband CDMA overlaying a GSM cellular system," IEEE Vehicular Technology Conference, pp. 31-35, 1994.

[Grozev93] J. Grozev and David Everitt, "Reduction in capacity: the price pay for imperfect power control in a cellular CDMA system," In the Proceedings of the eight Australian Teletraffic Research Seminar, Melbourne, pp. 331-340, December 6-8, 1993.

[Gruber83] J. G. Gruber and N. H. LE, "Performance requirements for integrated voice/data networks," IEEE Journal on Selected Areas in Communications, vol. 1. no. 6, pp. 981-1005, December 1983.

[Guer88] R. Guerin, "Queueing-blocking system with two arrival streams and guard channels," IEEE Transactions on Communications, vol. 36, no. 2, pp153-163, February 1988.

[Guerin90] R. Guerin and L. Y. C. Lien, "Overflow analysis for finite waiting room systems," IEEE Transactions on Communications, vol. 38, no. 9, pp.1569-1577, September 1990.

[Gunt98] A. Guntsch, M. Ibnkahla, G. Losquadro, M. Mazzella, D. Roviras and A. Timm, "EU's R&D activities on third-generation mobile satellite systems (S-UMTS)," IEEE Communications magazine, pp.104-110, February 1998.

[Guo96] N. Guo, S. D. Morgera and P. Mermelstein, "Common packet data channel (CPDC) for integrated wireless DS-CDMA networks," IEEE Journal Selected Areas in Communication, vol 14, no. 4, pp. 735-749, May 1996.

[Gusella91] R. Gusella, " Characterizing the variability of arrival processes with indexes of dispersion," IEEE Journal of Selected Areas in Communications, vol. 9, no. 2, pp. 203-211, February 1991.

[Hamal95] J. Hamalainen et al., "Multi-slot packets radio air interface to TDMA system – Variable Rate Reservation Access (VRRA)," In the proceedings of 1995 Sixth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'95), pp. 366-371, September 27 - 29, 1995.

[Hanzo94] L. Hanzo, J. C. S. Cheung, R. Steele and W. T. Webb, "A packet reservation multiple access assisted cordless telecommunication scheme," IEEE Transactions on Vehicular Technology, vol. 43, no. 2, pp. 234-244, May 1994.

[Har80] E. A. Harrington, "Voice/data integration using circuit switched networks," IEEE Transactions on Communications, vol. 28, pp. 781-793, June 1980.

[Heffes80] H. Heffers, "A class of data traffic processes-covariance function characterization and related queuing results," The Bell System Technical Journal, vol. 59, no. 6, pp.897-929, July-August, 1980.

[Hoff98] S. Hoof, M. Meyer and J. Sachs, "A performance evaluation of Internet access via the General Packet Radio Service of GSM" IEEE 48th Vehicular Technology Conference, Ottawa, Canada, pp.1760-1764, 18th-21st, May 1998.

[Hong86] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," IEEE Transactions on Vehicular Technology, vol. VT-35, no. 3: pp.77-92, Aug. 1986.

[Hong89] D. Hong and S. S. Rappaport, "Priority oriented channel access for cellular systems serving vehicular and portable radio telephone," IEE Proceedings, vol. 136, no. 5, pp. 339-346, October 1989.

[Honk94] Z. Honkasalo, H. Honkasalo, J. Hamalainen and H. Jokinen, "GSM/DCS air interface enhancements for high speed data applications," Proceedings of International Conference on Universal Personal Communication, pp. 480-484, Sep 27-Oct 1, 1994.

[Hu95] L.-R. Hu and S. S. Rappaport, "Personal communication systems using multiple hierarchical cellular overlays," IEEE Journal on Selected Areas in Communications, vol. 13, no. 2, pp.406-415, February 1995.

[Ishika97] Y. Ishikawa and N. Umeda, "Capacity design and performance of call

admission control in cellular CDMA systems," IEEE Journal on Selected Areas in Communications, vol. 15, no. 8, pp. 1627-1635.

[Ivan98] M. Ivanovich, M. Zukerman, P. Fitzpatrick and M. Gitlits, "Performance between circuit allocation schemes for half- and full-rate connections in GSM," IEEE Transactions on Vehicular Technology, vol. 47, no. 3, pp. 790-797, August 1998.

[Jabb96] B. Jabbari, "Teletraffic aspects of evolving and next generation wireless communication networks," IEEE Personal Communications, pp. 4-9, December 1996.

[Jabb97] B. Jabbari and W. F. Fuhrmann, "Teletraffic modelling and analysis of flexible hierarchical cellular networks with speed-sensitive handoff strategy," IEEE Journal on Selected Areas in Communications, vol. 15, no. 8, pp.1539-1548, 1997.

[Jager84] D. L. Jagerman, "Methods in traffic calculations," AT&T Bell Laboratories Technical Journal, vol 63, no. 7, pp.1283-1310, September 1984.

[Katze96] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," IEEE Personal Communications, pp.10-31, June 1996.

[Kaya98] H. Kayama, T. Ichikawa, H. Takanashi, M. Morikura and T. Tanaka, "A multi-slot access control for TDMA-TDD packet radio channel-Application to PHS packet data system," IEICE Transactions on Communications, vol. E81-B, no. 5, pp.1115-1125, May 1998.

[Keil95] J. Keilson and O. C. Ibe, "Cutoff priority scheduling in mobile cellular communication systems," IEEE Transactions on Communications, vol. 43, no. 2/3/4, pp. 1038-1045, 1995.

[Kelly91] F. P. Kelly, "Loss networks," The Annals of Applied Probability, vol.1, no.3, pp.319-378, 1991.

[Kim98] B. K. Kim and H. J. Choi, "Teletraffic model considering subscriber and satellite mobility in the terrestrial cell/satellite beam hierarchical structure," IEICE Transactions on Communications, vol. E81-B, No. 3, pp. 647-658, March 1998.

[Klein75] L. Kleinrock, "Queueing system," Vol. 1, John Wiley & Sons. 1975.

[Kohno95] R. Kohno, R. Meidan and L. B. Milstein, "Spread spectrum access methods for wireless communications," IEEE Communications Magazine, pp. 58-67, January 1995.

[Kraime85] B. Kraimeche and M. Schwartz, "Analysis of traffic access control

strategies in integrated service networks," IEEE Transactions on Communications, vol. Com-33, no. 10, pp.1085-1093. 1985.

[Kuczu73] A. Kuczura, "The Interrupted Poisson Process as an overflow process," The Bell System Technical Journal, vol. 52, no. 3, pp.437-448, March 1973.

[Kuczu79] A. Kuczura and D. Bajaj, "A method of moments for the analysis of a switched communication network's performance," IEEE Transactions on Communications, vol. 25, no. 2, pp.185-193, February 1977.

[Lagran95] X. Lagrange and P. Godlewski, "Teletraffic analysis of a hierarchical cellular network," Proceeding of IEEE Vehicular Technology Conference, pp.882-886. 1995.

[Lagran96] X. Lagrange and P. Godlewski, "Performance of a hierarchical cellular network with mobility-dependent hand-over strategies," Proceeding of IEEE Vehicular Technology Conference, pp.1668-1872, 1996.

[Latou93] G. Latouche, "A logarithmic reduction algorithm for quasi-birth-death processes," Journal of Applied Probability, vol. 30, pp. 650-674, 1993.

[Latou98] G. Latouche, C. E. M. Pearce and P. G. Taylor, "Invariant measures for quasi-birth-death processes," Communications in Statistics: Stochastic Models, 14 (1&2), pp. 443-460, 1998.

[Lavery93] B. Lavery and D. Everitt, " On the teletraffic characterisation of cellular CDMA systems," In the Proceedings of Vehicular Technology Conference, pp. 416-419, 1993.

[Li94] F. Li and L. F. Merakos, "Voice/data channel access integration in TDMA digital cellular networks," IEEE Transactions on Vehicular Technology, vol. 43, no. 4, pp. 986-996, November 1994.

[Lin94] Y-B Lin, S. Mohan and A. Noerpel, "Queueing priority channel assignment strategies for PCS hand-off and initial access," IEEE Transactions on Vehicular Technology, vol. 43, no.3, pp.704-712, August, 1994.

[Lin96a] Y-B Lin and W. Chen, "Impact of busy lines and mobility on call blocking in a PCS network," International Journal of Communications Systems, no.9, pp.35-45, 1996.

[Lin96b] Y-B Lin, A. R. Noerpel and Danial J. Harasty, "The sub-rating channel assignment strategy for PCS hand-offs," IEEE Transactions on Vehicular Technology, vol.45, no. 1, pp. 122-130, February 1996.

[Lin97a] Y-B Lin, "Reducing location update cost in a PCS network," IEEE/ACM Transactions on Networking, 5(1), pp. 25-33, February 1997.

[Lin97b] Y-B Lin, "Performance modeling for mobile telephone networks," IEEE Network, pp. 63-68, November/December, 1997.

[Lui94] Z. Liu and M. E. Zarki, "SIR-based call admission control for DS-CDMA cellular systems," IEEE Journal on Selected Areas in Communications, vol. 12, no. 4, pp. 638-644, May 1994.

[McMil93] D. W. McMillan, "Traffic modelling and analysis for cellular mobile networks," Ph. D thesis, the University of Melbourne, 1993.

[Meier89] K. S. Meier-Hellstern, "The analysis of a queue arising in overflow models," IEEE Transactions on Communications, vol. 37, no.4, pp.367-372, April 1989.

[Mermel93] P. Mermelstein, A Jalali and H. Leib, "Integrated services on wireless multiple access networks," IEEE International Conference on Communications. Vol. 2, pp863-867 1993.

[Mihai97] C. Mihailescu, X. Lagrange and D. Zeghlache, "Analysis of a two-layer cellular mobile communication system," Proceeding of IEEE Vehicular Technology Conference, pp.954-958, 1997.

[Mitrou93] N. M. Mitrou, G. L. Lyberopoulos and A. D. Panagopoulou, "Voice and data integration in the air-interface of a microcellular mobile communication system," IEEE Transactions on Vehicular Technology, vol. 42, no.1, pp.1-13, February 1993.

[Muku81] K. Mukumoto and A. Fukuda, "Idle signal multiple-access (ISMA) scheme for terrestrial packet radio networks," Transactions of IEICE, vol. J64-B, no. 10, pp.66-74, Oct. 1981 (in Japanese).

[Naga98] M. Nagatsuka, Y. Ishikawa and S. Uebayashi, "Data traffic control and capacity evaluation for voice/data integrated transmission in DS-CDMA," IEICE Transaction on Communication, vol E81-B, no. 7, pp. 1355-1364, July 1998.

[Nagh96] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," IEEE Journal on Selected Areas in Communications, vol. 14, pp. 711-717, May 1996.

[Nanda91] S. Nanda, D. J. Goodman and U. Timor, "Performance of PRMA: A packet voice protocol for cellular systems," IEEE Transactions on Vehicular Technology, vol. 40, no. 3, pp. 584-598, August, 1991.

[Nanda94] S. Nanda, "Stability evaluation and design of the PRMA joint voice

data system," IEEE Transactions on Communications, vol.42, no. 5, pp. 2092-2104, May 1994.

[Neuts81] M. F. Neuts, "Matrix-geometric solutions in stochastic models-An algorithmic approach," Johns-Hopkins. 1981.

[Nomura89] M. Nomura, T. Fujii and N. Ohta, "Basic characteristics of viable rate video coding in ATM environment," IEEE Journal on Selected Areas in Communications, vol. 7, no.5, pp. 752-760, June 1989.

[Pavlid94a] F. N. Pavlidou, "Mixed media cellular systems," IEEE Transactions on Communications, vol. 4, no. 2/3/4, pp. 848-853, 1994.

[Pavlid94b] F-N Pavlidou, "Two dimensional traffic models for cellular mobile systems," IEEE Transactions on Communications, vol. 42, no. 2/3/4, pp. 1505-1511, 1994.

[Pick91] R. L. Pickholtz, L. B. Milstein and D. L. Schilling, "Spread spectrum for mobile communications," IEEE Transactions on Vehicular Technology, vol. 40, no. 2, pp. 313-322, May 1991.

[Posne95] E. C. Posner and R. Guerin, "Traffic policies in cellular radio that minimize blocking of handoff calls," In Proceeding of 11th International Teletraffic Congress, Kyoto, 1995.

[Posner85] E. C. Posner and R. Guerin, "Traffic policies in cellular radio that minimize blocking of handoff calls," ITC-11, pp. 2.4B-2.1-2.4B-2.5, 1985.

[Prasad96] R. Prasad, "CDMA for wireless personal communications," Artech House, 1996.

[Qi96] H. Qi and R. Wyrwas, "Performance analysis of joint voice-data PRMA over random packet error channels," IEEE Transactions on Vehicular Technology, vol. 45, no. 2, pp. 332-344, May 1996.

[Rama99] P. Ramanathan, K. Sivalingam, P. Agrawal and S. Kishore, "Dynamic resource allocation schemes during handoff for mobile multimedia wireless networks," IEEE Journal on Selected Areas in Communications, vol. 17, no. 7, pp. 1270-1283, July 1999.

[Ramas85a] V. Ramaswami and D. M. Lucantoni, "Stationary waiting time distribution in queues with phase type service and in quasi-birth-and-death process," Stochastic Models, vol. 1, no.2, pp. 125-136, 1985.

[Rapp91] S. S. Rappaport, "The multiple-call hand-off problem in high-capacity cellular communications systems," IEEE Transactions on Vehicular Technology, vol. 40, no. 3, pp. 546-557, August 1991.

[Rapp96] S. S. Rappaport and C. Purzynski, "Prioritized resource assignment for mobile cellular communication systems with mixed services and platform types," IEEE Transactions on Vehicular Technology, vol. 45, no. 3, pp. 443-458, August 1996.

[Ren98] W. Ren, M. Sweeting, C. Fan and J. Paffett, "Integration of speech and data over non-collision integrated packet reservation multiple access protocol," In the proceedings of IEEE GLOBECOM, vol.3, pp.1344-1349, Sydney, Australia, 8-12, November, 1998.

[Ross82] M. J. Ross and O. A. Mowafi, "Performance analysis of hybrid switching concepts for integrated voice/data communications," IEEE Transactions on Communications, vol 30, no. 5, pp.1073-1087, May 1982.

[Rurzy95] C. Purzynski and S. S. Rappaport, "Multiple call hand-off problem with queued hand-offs and mixed platform types," IEE Proceedings-Communications, vol.142, no. 1, pp.31-39, February 1995.

[Sam97] A. Sampath and J. M. Holtzman, "Access control of data in integrated voice/data CDMA systems: benefits and tradeoffs," IEEE Journal on Selected Areas in Communications, vol. 15, no. 8, pp. 1511-1526, 1997.

[Sark97] J. H. Sarker, "A Consistent Time Division Multiple Access (CTDMA) for integrated voice and data wireless networks. International Conference on Telecommunications," Vol. 3, pp.1211-1216. Melbourne, Australia, 2-5 April 1997.

[Schwartz87] M. Schwartz, "Telecommunication Networks: Protocols, Modelling and Analysis," Addition-Wesley Publishing Company, New York, 1987.

[Serres88] Y. D. Serres and L. G. Mason, "A multiserver queue with narrow- and wide-band customers and wide-band restricted Access," IEEE Transactions on Communications, vol. 36, no. 6, pp. 675-684, June 1988.

[Sriram83] K. Sriram, P. Varshney and J. G. Shanthikumar, "Discrete-time analysis of integrated voice/data multiplexers with and without speech activity detectors," IEEE Journal of Selected Areas in Communications, vol. 1, no. 6, pp. 1124-1132, December 1983.

[Sriram86] K. Sriram and W. Whitt, "Charactering superposition arrival processes in packet multiplexers for voice and data," IEEE Journal on Selected Areas in Communications, vol. 4, no. 6, pp. 833-846, September 1986.

[Steele92] R. Steele and M. Nofal, "Teletraffic performance of microcellular

personal communication networks," IEE Proceedings-I, vol. 139, no. 4, pp. 448-461, August 1992.

[Stern90] H. P. Stern, "Design issues relevant to developing an integrated voice/data mobile radio system," IEEE Transaction on Vehicular Technology, vol. 39, no.4, pp. 281-288, November 1990.

[Tan96] L. Tan and Qi Tu Zhang, "A reservation random-access protocol for voice/data integrated spread-spectrum multiple-access systems," IEEE Journal on Selected Areas in Communications, vol.14, no. 9, pp. 1717-1727, December 1996.

[Taylor91] J. T. Taylor and J. K. Omura, "Spread spectrum technology: A solution to the personal communications services frequency allocation dilemma," IEEE Communication Magazine, vol.29, pp. 48-51, Feb. 1991.

[Bout97] T. V. Bout, "Reservation-Time Division Multiple Access protocols for wireless personal communications," Ph. D. thesis, the University of Adelaide, Australia, 1997.

[Boutyu97] T. Buot and S. Yu, "Video transmission over reservation-TDMA protocols," Proceedings of the Third Asia-Pacific Conference on Communications, Sydney, pp. 1453-1456, December 7-10 1997.

[Tekin92] S. Tekinay and B. Jabbari, "A measurement-based prioritization scheme for handovers in mobile cellular networks," IEEE Journal on Selected Areas in Communications, vol. 10, no. 8, pp. 1343-1350, October, 1992.

[Tekin93] S. Tekinay, B. Jabbari and A. Kakaes, "Modeling of cellular communication networks with heterogenous traffic sources," IEEE 2nd International Conferences on Universal Personal Communications, Ottawa, Canada, pp. 249-253, Oct. 12-15, 1993.

[Thomas88] R. Thomas, H.Gilbert, and G. Maziotto, "Influence of the moving of the mobile stations on the performance of a radio mobile cellular network," Proc. 3rd Nordic Seminar on Digital Land Mobile Radio Communications, paper 9.4, Sept 1988.

[Tripa98] N. D. Tripathi, J. H. Reed and H. F. Vanlandingham, "Handoff in cellular systems," IEEE Personal Communications, pp. 26-37, December 1998.

[Tsang90] D. H. K. Tsang and K. W. Ross, "Algorithms to determine exact blocking probabilities for multirate tree networks," IEEE Transactions on Communications, vol. 38, no. 8, pp.1266-1271, August 1990.

[Turina96] D. Turina, B. Per, E. Schoster and A. Andersson, "A proposal for

multi-slot MAC layer operation for packet data channel in GSM," Proceedings of International Conference on Universal Personal Communications, vol.2, pp. 572-576, Sept-Oct, 1996.

[Urie93] A. Urie, "Advanced TDMA Mobile Access (ATDMA)," 2nd International Conference on Universal Personal Communications, vol. 1, pp392-396. Ottawa, Canada, October 1993.

[Viterbi93] A. M. Viterbi and A. J. Viterbi, "Erlang capacity of a power controlled CDMA system," IEEE Journal on Selected Areas in Communications, vol. 11, No. 6, pp. 892-900, August 1993.

[Viterbi94] A. J. Viterbi, A. M. Viterbi and E. Zehavi, " Other-cell interference in cellular power-controlled CDMA," IEEE Transactions on Communications, vol. 42, no. 2/3/4, pp1501-1504, 1994.

[Wallace69] V. Wallace, "The solution of quasi birth and death processes arising from multiple access computer systems," Ph. D. Dissertation, Systems Engineering laboratory, University of Michigan, 1969.

[Wang95] L. Wang, Z. Honkasalo, J. Hamalainen and H. Jokinen, "A physical layer proposal for multi-slot packet radio services in the existing TDMA cellular system," The 1995 Sixth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'95), pp. 876-885, Sept 27-29, 1995.

[Weinst78] C. J. Weinstein, "Fractional speech loss and talker activity model for TASI and for packet-switched speech," IEEE Transactions on Communications. Vol. 26, No. 8, pp. 1253-1257, August 1978.

[Weinst80] C. J. Weinstein, M. L. Malpass and M. J. Fisher, "Data traffic performance of an integrated circuit- and packets-switched multiple structure," IEEE Transactions on Communications, vol. 28, no.6, pp. 873-878, June 1980.

[Wen95a] J-H Wen and J-W Wang, "A new protocol for wireless voice communications—non-collision packet reservation multiple access," Proceedings of Personal, Indoor and Mobile Radio Communications (PRIMRC'95), vol. 2, pp. 638-642, Sept. 27-29, 1995.

[Wen95b] J-H Wen and Jee-Wey Wang, "A Non-collision PRMA protocol for integrated voice and data wireless networks," IEEE Fourth International Conference on Universal Personal Communications (ICUPC'95), Tokyo, pp. 462-466, Nov. 6-10, 1995.

[Wiese95] J. E. Wieselthier and A. Ephremides, "Fixed- and movable-boundary channel access schemes for integrated voice/data wireless networks," IEEE Transaction on Communications, vol. 43, no. 1, pp. 64-74, January 1995.

[Wilki56] R. I. Wilkinson, "Theories for toll traffic engineering in the U.S.A.," The Bell System Technical Journal, vol. 35, pp. 421-514, Mar. 1956.

[Willi84] G. F. Williams and A. Leon-Garcia, "Performance analysis of integrated voice and data hybrid-switched links," IEEE Transactions on Communications, vol. Com-32, no.6, pp. 695-706, June 1984.

[Wilson93] N. D. Wilson, R. Ganesh, K. Joseph and D. Raychaudhuri, "Packet CDMA versus dynamic TDMA for multiple access in an integrated voice/data PCN," IEEE Journal on Selected. Areas in Communications, vol. 11, no. 6, pp. 870-884, August 1993.

[Wong92] W. C. Wong and D. J. Goodman, "A packet reservation multiple access protocol for integrated speech and data transmission," IEE PROCEEDINGS-I, vol.139, no. 6, pp. 607-612, December 1992.

[Wong93a] W. C. Wong and D. S. Goodman, "Integrated data and speech transmission using packet reservation multiple access," In the Proceedings of IEEE International Conference on Communications, Geneva, Switzerland, pp. 172-176, May 1993.

[Wong93b] W-C Wong, "Packet Reservation Multiple Access in a metropolitan microcellular radio environment," IEEE Journal on Selected Areas in Communications, vol. 11, no. 6, pp. 918-925, August, 1993.

[Wongd97] D. Wong and T. J. Lim, "Soft handoffs in CDMA mobile systems," IEEE Personal Communications, pp.6-17, December 1997.

[Wu94] G. Wu, K. Mukumoto and A. Fukuda, "Analysis of an integrated voice and data transmission system using packet reservation multiple access," IEEE Transactions on Vehicle Technology, vol. 43, no. 2, pp. 289-297, May 1994.

[Yoon89] C. H. Yoon and C. K. Un, "Efficient handoff policy without guard channels for mobile radio telephone systems," Electronics Letters, vol. 25, no. 11, pp. 700-701. May 25th 1989.

[Yoon93] C. H. Yoon and C. K. Un, "Performance of performance portable radio telephone systems with and without guard channels," IEEE Journal on Selected Areas in Communications, vol. 11, no. 6, pp. 911-917, August 1993.

[Yu97] S. Yu and Ted Bout, "Analysis of integrated services in GPRS cellular systems," Proceedings of the Third Asia-Pacific Conference on Communications, pp. 474-478, Sydney, December 7-10 1997.

[Yu98a] S. Yu, "Analysis of wideband and narrowband integrated services in cellular systems," Proceedings of IFIP the 3rd workshop on Personal

Wireless Communications, pp. 183-190, Tokyo, April 8-9, 1998.

[Yu98b] S. Yu, "Data traffic control scheme for wideband and narrowband integrated services in PCS," Proceedings of IEEE Globecom'98, Vol. 3, pp. 1437-1442, Sydney, November 8-12, 1998.

[Zhang90] K. Zhang and K. Pahlavan, "An integrated voice/data system for mobile indoor radio networks," IEEE Transactions on Vehicular Technology, vol. 39, no.1, pp. 75-82, February 1990.

[Zonoo97] M. M. Zonoozi and P. Dassanayake, "User mobility modelling and characterization of mobility patterns," IEEE Journal on Selected Areas in Communications, vol. 15, no.7, pp. 1239-1252, September 1997.

[Zuker89] M. Zukerman, "Applications of matrix-geometric solutions for queueing performance evaluation of a hybrid switching system," J. Austral. Math. Soc. Ser., B31, pp. 219-239, 1989.

Bibliography

[Ander95] P.-G Andermo and L-M Ewerbring, "A CDMA-based radio access design for UMTS," IEEE Personal Communications, pp. 48-53, February, 1995.

[Baier93] A. Baier, "Open multi-rate radio interface architecture based on CDMA," In Proceeding 2nd International Conference on Universal Personal Communications, Ottawa, Canada, pp. 985-989, Oct. 12-15, 1993.

[Baier94] A. Baier, U-C Fiebig, W. Granzow, W. Koch, P. Teder and J. Thielecke, "Design study for a CDMA-based third-generation mobile radio system," IEEE Journal on Selected Areas in Communications, vol. 12, no. 4, pp. 733-743, May 1994.

[Bohm96] C. Bohm, Markus Hidell, Per Lindgren, Lars Ramfelt and Peter Sjodin, "Fast circuit switching for the next generation of high performance networks," IEEE Journal on Selected Areas in Communications. Vol. 14, No.2, pp.298-305, February 1996.

[Cameron96] R. Cameron and B. Woerner, "Performance analysis for CDMA with imperfect power control," IEEE Transactions on Communications, vol. 44, no. 7 pp. 777-781, July 1996.

[Chung93] S-P Chung and K. W. Ross, "Reduced load approximations for multirate loss networks," IEEE Transaction on Communications, vol. 41, no. 8, pp.1222-1231, August, 1993.

[Cooper78] G. R. Cooper and R. W. Nettleton, "A spread-spectrum technique for high capacity mobile communications," IEEE Transactions on Vehicular Technology, vol. 27, no. 4, pp. 264-275, November 1978.

[Crow73] W. Crowther, R. Rettberg, D. Walden, S. Ornstein and F. Heart, "A system for broadcast communication: Reservation-ALOHA," In Proceedings of 6th Hawaii Int. Conf. Syst. Sci., pp. 596-603, Jan. 1973.

[Harpan98] Z. Harpantidou and M. Paterakis, "Random multiple access of broadcast channels with Pareto distributed packet interarrival times," IEEE Personal Communications, pp. 48-55, April 1998.

[Honig84] M. L. Honig, "Analysis of a TDMA network with voice and data traffic," AT&T Bell Laboratories Technical Journal, vol. 63, no. 8, pp.1537-1563, Oct 1984.

[Knisely98] D. N. Knisely, S. Kumar, S. Laha and S. Nanda, "Evolution of wireless data services: IS-95 to cdma 2000," IEEE Communication Magazine, pp 140-149, 1998.

[Linnar93] J-P Linnartz, "Narrowband land-mobile radio networks," Artech House, 1993.

[McMill91] D. W. McMillan, "The M/M/c non-preemptive priority queue: a matrix-analytic approach," In Proceedings of the Sixth Australian Teletraffic Research Seminar, Wollongong, pp. 57-64, November 1991.

[Meier94] K. S. Meier-Hellstern, G. P. Pollini and D. J. Goodman, "Network protocols for the cellular packet switch," IEEE Transactions on Communication, vol. 42, no. 2/3/4, pp.1235-1244, 1994.

[Nikula98] E. Nikula, A. Toskala, E. Dahlman, L. Girard and A. Klein, "Frames multiple access for UMTS and IMT-2000," IEEE Personal Communications, pp.16-24, April 1998.

[Pahl94] K. Pahlavan and A. H. Levesque, "Wireless data communications," Proceedings of IEEE, vol. 82, No. 9, pp. 1398-1430, September 1994.

[Sche90] R. G. Schehrer, "On a cut-off priority queuing system with hysteresis and unlimited waiting room," Computer Networks and ISDN systems, no. 20, pp. 45-56. 1990.

[Tekin91] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks," IEEE Communications Magazine, pp. 42-46, 1991.

[Ojan98] T. Ojanpera and Ramjee Prasad, "Wideband CDMA for third generation mobile communications," Artech House, 1998.

[Vieter91] A. J. Viterbi, "Wireless digital communication: A view based on three lessons learned," IEEE Communication Magazine, vol. 29, pp. 33-36, September 1991.

[Wu95] C-N Wu, Y-R Tsai and J-F Chang, "A quality-based birth-and-death queueing model for evaluating the performance of an integrated voice/data CDMA cellular system," In the proceeding of the sixth IEEE International Symposium on Personal, Indoor and mobile radio Communications, vol.2, pp. 451-455. Sept. 27-29, Canada, 1995.

APPENDIX 1

The derivation of waiting time distribution for the R-TDMA integrated systems

The structure of the block-tridiagonal infinitesimal generator Q is given by:

$$Q = \begin{bmatrix} B_0 & A_0 & & & & \\ B_1 & A_1 & A_0 & & & \\ & A_2 & A_1 & A_0 & & \\ & & \cdots & \cdots & \cdots & \\ & & & A_2 & A_1 & A_0 \\ & & & & \cdots & \cdots \end{bmatrix} \qquad \text{(App.1.1)}$$

$$Q = \begin{bmatrix} \overline{Q} - \Lambda_{(p)} & \Lambda_{(p)} & & & & \\ \mathrm{M} & \overline{Q} - \Lambda_{(p)} - \mathrm{M} & \Lambda_{(p)} & & & \\ & \mathrm{M} & \overline{Q} - \Lambda_{(p)} - \mathrm{M} & \Lambda_{(p)} & & \\ & & \cdots & \cdots & \cdots & \\ & & & \mathrm{M} & \overline{Q} - \Lambda_{(p)} - \mathrm{M} & \Lambda_{(p)} \\ & & & & \cdots & \cdots \end{bmatrix}$$

(App.1.2)

where the row sums of Q are equal to zero. If Q is irreducible, all off-diagonal blocks are nonzero and nonnegative matrices. However, the diagonal elements of the diagonal blocks are negative.

The diagonal blocks A_1 correspond to the interphase shifts without consideration of arrivals and departures. The component $\overline{Q}$ of the infinitesimal generator Q can be expressed by:

$$\overline{Q} = \begin{bmatrix} -\lambda_v & \lambda_v & & & \\ \mu_v & -(\lambda_v + \mu_v) & \lambda_v & & \\ & \cdots & \cdots & \cdots & \\ & & (x_N - 1)\mu_v & -(\lambda_v + x_N\mu_v - \mu_v) & \lambda_v \\ & & & x_N\mu_v & -x_N\mu_v \end{bmatrix} \qquad \text{(App.1.3)}$$

The upper diagonal blocks $\Lambda_{(p)}$ of matrix Q are associated with arrival events. The arrival matrix with the dimension of $(x_N+1)\times(x_N+1)$ is given by:

$$\Lambda_{(p)} = \begin{bmatrix} p_k(i)\lambda_d & & & & \\ & \cdots & & & \\ & & p_n(i)\lambda_d & & \\ & & & \cdots & \\ & & & & p_1(i)\lambda_d \end{bmatrix} \qquad \text{(App.1.4)}$$

Note that all the other unmarked elements are equal to zero.

In contrast, the lower diagonal blocks M correspond to departure events with the dimension $(x_N+1)\times(x_N+1)$. The service matrix is given by:

$$\mathrm{M} = \begin{bmatrix} (N-N_0)\mu_d & 0 & \cdots & & 0 \\ 0 & (N-N_1)\mu_d & 0 & \cdots & 0 \\ \vdots & 0 & \cdots & 0 & 0 \\ & \vdots & 0 & (N-N_{N-1})\mu_d & 0 \\ 0 & 0 & 0 & 0 & (N-N_N)\mu_d \end{bmatrix}$$

(App.1.5)

The equilibrium probability π_∞ in the case of a continuous time Markov process will have to satisfy: $\pi_\infty Q = 0$ as well as $\sum_{i=0}^{\infty}\pi_\infty^{(i)} = 1$.

To be brief, these can be rearranged as:

$$\begin{aligned} &\pi_0(\overline{Q}-\Lambda_{(p)})+\pi_1\mathrm{M}=0 \\ &\pi_{j-1}\Lambda_{(p)}+\pi_j(\overline{Q}-\Lambda_{(p)}-\mathrm{M})+\pi_{j+1}\mathrm{M}=0, \qquad j\geq 1 \end{aligned} \qquad \text{(App.1.6)}$$

We then have:

$$\pi_\infty^{(0)}(I-R)^{-1}e = 1 \qquad \text{(App.1.7)}$$

where e denotes a column vector with all components equal to one and I is a diagonal unit matrix of order (x_N+1). R is the solution of non-linear matrix equations.

The equilibrium probability vector can be derived as $\pi_\infty = [\pi_\infty^{(0)}, \pi_\infty^{(1)}, \ldots, \pi_\infty^{(i)} \ldots] = [\pi_\infty^{(0)}, \pi_\infty^{(0)} R, \ldots, \pi_\infty^{(0)} R^i, \ldots]$.

Namely, the successive substitutions can be simply shown as:

$$\pi_\infty^{(i)} = \pi_\infty^{(0)} R^i \quad \forall i, \quad i \geq 0 \tag{App.1.8}$$

For the stability of a queue, it is necessary that the outflow rate from a state in the queue must be larger than the inflow rate into that state. That is, the stability condition of data queue can be written as:

$$\pi_\infty A_2 e > \pi_\infty A_0 e$$

Namely,

$$\pi_\infty \mathrm{M} e > \pi_\infty \Lambda_{(p)} e \tag{App.1.9}$$

where e denotes a column vector with unit elements.

For the waiting time distribution, it is known that both the phase type process and the renewal process can be characterised as a QBD process [*Ramas85*]. The stationary waiting time in the queue is the time till absorption in the Markov process with infinitesimal generator Q_Q.

$$Q_Q = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ S^\circ\beta & S & 0 & 0 & \cdots & 0 \\ 0 & S^\circ\beta & S & 0 & \cdots & 0 \\ 0 & 0 & S^\circ\beta & S & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots & 0 \\ & & & & \cdots & \cdots \end{bmatrix} \tag{App.1.10}$$

where (β, S) represents service time distribution $H(t)$ in the i.i.d. phase type process of order C.

Similarly, the waiting time in this study can be estimated by the time of the QBD process till it reaches the absorption states with the infinitesimal generator Q_Q. This is

equivalent to the probability that an arrival waits more than time t before obtaining service.

$$Q_Q = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ A_2 & A_0 + A_1 & 0 & 0 & \cdots & 0 \\ 0 & A_2 & A_0 + A_1 & & \cdots & 0 \\ 0 & 0 & A_2 & A_0 + A_1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & \cdots \end{bmatrix} \qquad \text{(App.1.11)}$$

Because of the simple boundary, the algorithm developed by Ramaswami [*Ramas85*] can be used. The important feature of this algorithm for the waiting time distribution is to exploit the computation of rate matrix R rather than the complex calculation of the differential equations.

Let $W(t)$ denote the waiting time distribution in the queue under the FIFO rule. The probability that an arrival in an arbitrary time has to wait more than time t can be given by:

$$W(t) = \sum_{n=0}^{\infty} d_n e^{-\theta t} \frac{(\theta t)^n}{n!} \quad while \quad t > 0 \qquad \text{(App.1.12)}$$

where θ is positive and d_n are probabilities.

$$d_n = \pi_{\infty}^{(0)} (I - R)^{-1} R H_n e \quad while \quad n \geq 0 \qquad \text{(App.1.13)}$$

$$H_0 = I, \quad H_n = H_{n-1} P_1 + R H_{n-1} P_2 \quad while \quad n \geq 1 \qquad \text{(App.1.14)}$$

and

$$\theta = Max(-(A_0 + A_1)_{jj}) \qquad \text{(App.1.15)}$$

$$P_1 = \frac{1}{\theta}(A_0 + A_1) + I \quad , \quad P_2 = \frac{1}{\theta} A_2 \quad . \qquad \text{(App. 1.16)}$$

APPENDIX 2

The derivation of the moments of overflow traffic distribution in the busy servers

Let $p(m,k)$ denote the stationary probability with m busy servers in the infinite group and k for the switched states. Forming a two-dimension Markov chain, we can write down the equilibrium equations with the boundary states as follows.

$$(m\mu+\omega)p(m,k)=\gamma p(m,k)+(m+1)\mu p(m+1,k) \quad while \quad k=0$$

$$(m\mu+\lambda+\gamma)p(m,k)=\omega p(m,k-1)+(m+1)\mu p(m+1,k)+\lambda p(m-1,k) \quad while \quad k=1$$

$$(\gamma+\lambda)p(m,k)=\omega p(m,k)+\mu p(m,k) \quad while \quad m=0 \; and \; k=1$$

(App.2.1)

Let the generating function be:

$$G(z)=\sum_{j=0}^{\infty}p(m,k)z^m=\sum_{j=0}^{\infty}(P(m,0)+P(m,1))z^m \qquad \text{(App.2.2)}$$

Namely,

$$G(z)=G_0(z)+G_1(z) \qquad \text{(App.2.3)}$$

Therefore we can have the second order of the differential equations for the $G_0(z)$ and $G_1(z)$ as:

$$\mu(z-1)G_0''(z)+[\mu+\gamma+\omega-\lambda(z-1)]G_0'(z)-\frac{\lambda}{\mu}\omega G_0(z)=0$$

$$\mu(z-1)G_1''(z)+[\mu+\gamma+\omega-\lambda(z-1)]G_1'(z)-\frac{\lambda}{\mu}(\mu+\omega)G_1(z)=0$$

(App.2.4)

The solution of the equations can be expressed by the confluent hypergeometric function. Consequently, the factorial moments of the busy server distribution becomes:

$$G^{(1)}(1) = (\frac{\lambda}{\mu})\frac{\omega}{\omega+\gamma}$$

$$G^{(2)}(1) = (\frac{\lambda}{\mu})^2 \frac{\omega(\omega+\mu)}{(\omega+\gamma)(\omega+\gamma+\mu)}$$

$$G^{(3)}(1) = (\frac{\lambda}{\mu})^3 \frac{\omega(\omega+\mu)(\omega+2\mu)}{(\omega+\gamma)(\omega+\gamma+\mu)(\omega+\gamma+2\mu)} \qquad \text{(App.2.5)}$$

Now let the traffic load a offer to the loss system with c servers. Then the factorial moments $M^{(n)}$ of the Y busy servers in the infinite group can be expressed by the Kosten's moment formula:

$$M^{(n)} = a^n \frac{\sigma_0(c)}{\sigma_n(c)} \quad while \quad n = 0, 1, 2, \cdots \qquad \text{(App.2.6)}$$

If we approximate the factorial moments of m in the IPP switch to the Kosten's moments of Y in the infinite group, we can have:

$$\frac{G^{(n+1)}(1)}{G^{(n)}(1)} = \frac{M_{(n+1)}}{M_{(n)}} = \delta_n \quad while \quad n = 0, 1, 2, \cdots \qquad \text{(App.2.7)}$$

Substituting into the expressions above, we can obtain:

$$\frac{\lambda}{\mu}\frac{(\omega+n\mu)}{(\omega+\gamma+n\mu)} = \delta_n \quad while \quad n = 0, 1, 2, \cdots \qquad \text{(App.2.8)}$$

This implies:

$$\frac{\lambda}{\mu}\frac{\omega}{(\omega+\gamma)} = \delta_0 \quad while \quad n = 0$$

$$\frac{\lambda}{\mu}\frac{(\omega+\mu)}{(\omega+\gamma+\mu)} = \delta_1 \quad while \quad n = 1$$

$$\frac{\lambda}{\mu}\frac{(\omega+2\mu)}{(\omega+\gamma+2\mu)} = \delta_2 \quad while \quad n = 2 \qquad \text{(App.2.9)}$$

In the end, the parameters of IPP can be obtained:

$$\lambda = \mu \frac{(2\delta_2 - \delta_1)(\delta_1 - \delta_0) - \delta_1(\delta_2 - \delta_1)}{2\delta_1 - \delta_0 - \delta_2} = \mu \frac{\delta_2(\delta_1 - \delta_0) - \delta_0(\delta_2 - \delta_1)}{(\delta_1 - \delta_0) - (\delta_2 - \delta_1)}$$

$$\omega = \mu \frac{\delta_0(\frac{\lambda}{\mu} - \delta_1)}{\frac{\lambda}{\mu}(\delta_1 - \delta_0)} = \mu \frac{\delta_0(\lambda - \mu\delta_1)}{\lambda(\delta_1 - \delta_0)}$$

$$\gamma = \omega \frac{(\frac{\lambda}{\mu} - \delta_0)}{\delta_0} = \omega \frac{\lambda - \mu\delta_0}{\mu\delta_0} \qquad \text{(App.2.10)}$$