Protocols for Voice/Data Integration in a CDMA Packet Radio Network

Garth Judge

Submitted in fulfilment of the academic requirements for the degree of

Doctor of Philosophy

in the School of Electrical and Electronic Engineering at the
University of Natal
Durban, South Africa

November 1999

Supervisor: Dr. Fambirai Takawira

ABSTRACT

Wireless cellular communications is witnessing a rapid growth in, and demand for, improved technology and range of information types and services. Future third generation cellular networks are expected to provide mobile users with ubiquitous wireless access to a global backbone architecture that carries a wide variety of electronic services. This thesis examines the topic of **multiple access protocols** and models suitable for modern third-generation wireless networks.

The major part of this thesis is based on a proposed Medium Access Control (MAC) protocol for a Code Division Multiple Access (CDMA) data packet radio network, as CDMA technology is proving to be a promising and attractive approach for spectrally efficient, economical and high quality digital communications wireless networks. The proposed MAC policy considers a novel dual CDMA threshold model based on the Multiple Access Interference (MAI) capacity of the system. This protocol is then extended to accommodate a mixed voice/data traffic network in which variable length data messages share a common CDMA channel with voice users, and where the voice activity factor of human speech is exploited to improve the data network performance. For the protocol evaluation, the expected voice call blocking probability, expected data throughput and expected data message delay are considered, for both a perfect channel and a correlated Rayleigh fading channel. In particular, it is shown that a significant performance enhancement can be made over existing admission policies through the implementation of a novel, dynamic, load-dependent blocking threshold in conjunction with a fixed CDMA multiple access threshold that is based on the maximum acceptable level of MAI.

PREFACE

The research work presented in this thesis was performed by Mr. Garth Judge, under the supervision of Prof. Fambirai Takawira, at the University of Natal's department of Electronic Engineering. The thesis topic was initially sponsored by Telkom SA Ltd as part of the Telkom Teletraffic Initiative Programme (TIP), and later became part of the research programme at the University of Natal's Centre for Radio Access Technologies, which is sponsored by Telkom SA Ltd and Alcatel Altech Telecomms.

Parts of this thesis have been presented by the student at the Teletraffic'96 conference in Durban, South Africa, the COMSIG'97 conference in Grahamstown, South Africa, the IEEE ISSSTA'98 conference at Sun City, South Africa, and the PIMRC'99 conference in Osaka, Japan. A part of this thesis has also been reviewed and accepted for publication in the ACM/Baltzer Publishers journal "Wireless Networks", and is due to be published in 2000/2001. Two other journal papers have been submitted for review to the journals "IEEE Transactions on Vehicular Technology" and "Wireless Networks".

The whole thesis, unless specifically indicated to the contrary in the text, is the student's own work, and has not been submitted in part, or in whole to any other University.

ACKNOWLEDGEMENTS

Firstly, I wish to thank my supervisor, Prof. Fambirai Takawira for his invaluable help, encouragement, motivation and constructive criticism during the past four years of my postgraduate education. His patience and support have been greatly appreciated.

Secondly, a great thanks is owed to my family for their patience, encouragement and support, not only over the past four years, but throughout my tertiary education. In particular, I would like to thank my mother for proof reading parts of this document.

Thanks is also owed to Telkom SA Ltd and Alcatel Altech Telecomms for their valued financial support over the last two years, and for providing the equipment necessary for the completion of this doctoral degree.

Finally, thanks must go to my other postgraduate colleagues who provided hours of entertainment, sporting challenges, criticism (constructive and destructive) and support during the times when work was the last thing on my mind.

TABLE OF CONTENTS

PREFACE iii
ACKNOWLEDGEMENTS iv
TABLE OF CONTENTS v
LIST OF FIGURES/TABLES ix
LIST OF ACRONYMS xii

CHAPTER 1 - Introduction 1-1

1.1 Background 1-1
1.1.1 Integrated Services Networking 1-2
1.1.2 Wireless Access Networks 1-3
1.1.3 The Medium Access Control (MAC) Protocol 1-5
1.2 Focus and Goals of this Thesis 1-7
1.3 Document Overview and Layout 1-9
1.4 Original Contributions of the Thesis 1-12

CHAPTER 2 - Traditional Medium Access Control (MAC) Protocols 2-1

2.1 Introduction 2-1
2.2 Multiple Access Issues in Direct-Sequence Spread-Spectrum Communications 2-2
2.2.1 Multiple channel access in DS/SS systems 2-3
2.2.2 Multiple access capability of a DS/SS channel 2-4
2.3 Fixed Assignment MAC Protocols 2-7
2.4 Random Access MAC Protocols 2-8
2.4.1 The ALOHA protocol 2-9
2.4.2 Collision Resolution Algorithms (CRA) and System Stability Issues 2-10
2.4.3 Carrier Sense Multiple Access (CSMA) 2-14
2.4.4 CSMA with Collision Detection (CSMA/CD) 2-16
2.4.5 Inhibit Sense Multiple Access (ISMA) 2-17
2.4.6 Performance of random access protocols in packet radio networks 2-18
2.5 Spread-Spectrum (CDMA) Random Access MAC Protocols 2-19
2.5.1 Spread-spectrum ALOHA (SS/ALOHA) 2-19
2.5.2 Channel Load Sensing (CLS) 2-23
2.6 Demand Assignment and Reservation MAC Protocols 2-26
2.6.1 Reservation-ALOHA (R-ALOHA) 2-27
2.6.2 Robert's Reservation Protocol 2-28
2.7 Summary 2-30

CHAPTER 3 - Dual-Threshold CDMA (DT/CDMA) MAC Protocol 3-1

3.1 Introduction 3-1
3.2 Network Model 3-2
3.2.1 Network architecture 3-2
3.2.2 Description of the reverse link communications interface 3-3

3.2.3 Physical channel assumptions 3-3
3.2.4 Forward link and signalling assumptions 3-4
3.3 Description of the Dual-Threshold CDMA MAC Protocol 3-5
3.3.1 ISMA/CDMA: A centralised channel state broadcast scheme 3-5
3.3.2 Application of a Dual-Threshold model 3-6
3.3.3 The Dual-Threshold CDMA MAC procedure 3-8
3.4 System Performance Analysis 3-11
3.4.1 Data traffic model 3-12
3.4.2 Description of all pertinent variables and parameters 3-13
3.4.3 Definition of the data network state 3-14
3.4.4 Computation of the data state transition probability matrix 3-16
3.4.5 Distribution of the length of a Transmission Period 3-18
3.4.6 Computing the probability that the reference message begins in slot Z 3-19
3.4.7 The state distribution as seen by the first packet of a reference message 3-20
3.4.8 Computation of the message blocking probability 3-21
3.4.9 Computation of a packet's success probability when MAI is considered 3-22
3.4.10 Computation of the message success probability 3-22
3.4.11 Computation of the message and packet throughputs 3-24
3.4.12 Computation of the average message delay 3-25
3.5 Performance Results 3-27
3.5.1 Network parameters 3-27
3.5.2 Determination of the data collision threshold β 3-29
3.5.3 Effect of varying the data blocking threshold α 3-34
3.5.4 Effect of α on the average message delay 3-41
3.5.5 Effect of varying the mean message length 3-42
3.5.6 Comparison of DT/CDMA with other random access CDMA models 3-44
3.5.7 Downfalls of the proposed data traffic model and DT/CDMA 3-45
3.6 Summary 3-47

CHAPTER 4 - Selective-Repeat Dual-Threshold CDMA (SR-DT/CDMA) MAC Protocol 4-1

4.1 Introduction 4-1
4.2 Description of the SR-DT/CDMA MAC Protocol 4-1
4.3 Finite Population Traffic Model Issues and Assumptions 4-3
4.3.1 Solution to the problem of finding $P_{RML}(l)$ 4-4
4.3.2 Computation of the data message arrival process 4-8
4.3.3 Computation of the combined message length distribution $P_{ML}(l)$ 4-9
4.4 SR-DT/CDMA Protocol Performance Analysis 4-11
4.4.1 Concept of the system information flow model 4-12
4.4.2 Computation of the data system state vector 4-13
4.4.3 Computation of the message blocking probability 4-15
4.4.4 Computation of the message success probability 4-16
4.4.5 Computation of the system throughput (output flow rates) 4-17
4.4.6 Computation of the expected message delay 4-18
4.4.7 Equating the flow equations to obtain a solution 4-19
4.4.8 System performance for an infinite population model 4-20
4.5 Performance Results 4-21
4.5.1 Network parameters 4-21

4.5.2 Determination of the collision threshold β 4-22
4.5.3 Effect of the α threshold on the system performance 4-25
4.5.4 Evaluation of the finite population equilibrium-point analysis 4-27
4.5.5 Comparison of DT/CDMA and SR-DT/CDMA 4-32
4.6 Summary 4-34

CHAPTER 5 - Protocols for Integrating Voice and Data 5-1

5.1 Introduction 5-1
5.2 Exploitation of the Silent Periods in Human Speech 5-2
5.3 Wideband TDMA Frame Protocols for Voice/Data 5-5
5.3.1 Packet Reservation Multiple Access (PRMA) 5-5
5.3.2 Derivatives of Robert's reservation scheme for voice/data integration 5-6
5.3.3 Movable frame boundary techniques 5-7
5.3.4 Narrowband wireless ATM MAC protocols 5-9
5.4 Multimedia and Voice/Data CDMA MAC Protocols 5-11
5.4.1 The pure spread-spectrum ALOHA (SS/ALOHA) approach 5-11
5.4.2 Dual threshold voice/data MAC model with movable "code" boundary 5-12
5.4.3 Wideband-CDMA (W-CDMA) 5-15
5.4.4 WISPER 5-17
5.5 Summary 5-18

CHAPTER 6 - Voice and Data Integration in DT/CDMA 6-1

6.1 Introduction 6-1
6.1.1 Justification for the proposed model 6-1
6.2 System Model 6-2
6.2.1 Network architecture 6-2
6.2.2 MAC protocol for the voice traffic 6-3
6.2.3 Voice traffic model 6-5
6.2.4 Data admission policy 6-7
6.2.5 Data traffic model 6-9
6.2.6 Voice has unconditional priority over data 6-9
6.3 Voice Protocol Performance Analysis 6-12
6.3.1 Definition and computation of the voice network state statistics 6-12
6.3.2 Computation of the expected offered and carried voice loads 6-15
6.3.3 Computation of the expected voice call blocking probability 6-16
6.4 Integrated Data/Voice Protocol Performance Analysis 6-18
6.4.1 Computation of the statistics for the modified data threshold β' 6-18
6.4.2 Computation of the modified data state transition probability matrices 6-19
6.4.3 Computation of the data message blocking probability 6-21
6.4.4 Computation of the data throughput 6-22
6.4.5 Data performance without voice activity exploitation 6-23
6.5 Results for the Integrated Traffic Model 6-24
6.5.1 Network parameters 6-24
6.5.2 Results for the voice network 6-26
6.5.3. Results for the integrated voice/data network 6-28
6.5.4 Discussion on the choice of thresholds and exploitation of the VAF 6-32
6.6 Summary 6-34

CHAPTER 7 - Effects of a Correlated Rayleigh Fading Channel on the Proposed System 7-1

7.1 Introduction 7-1
7.2 Traditional Markov Channel Models 7-2
7.2.1 The Gilbert-Elliot model 7-3
7.2.2 Higher-order Hidden Markov Models 7-5
7.3 Proposed Channel Signal Model 7-5
7.3.1 User Signal Model (U) 7-6
7.3.2 MAI Model (Y) 7-9
7.4 Proposed Markov model for a Multi-user DS-CDMA Channel 7-13
7.4.1 Description of the relationship between the SIR and the channel state 7-13
7.4.2 Description of the Markov model 7-14
7.4.3 Computation of the stationary state occupancy and transition probabilities 7-16
7.4.4 Proposed mapping of the channel state onto the packet error process 7-20
7.4.5 Computations of the packet success probability 7-21
7.4.6 Computation of the message success probability 7-22
7.4.7 Computation of the expected packet throughput 7-25
7.5 Results 7-26
7.5.1 Network Parameters 7-26
7.5.2 Determination of the fading threshold ratio θ. 7-26
7.5.3 Effect of correlated Rayleigh fading on the data network throughput 7-31
7.5.4 Evaluation of the HMM for a correlated channel 7-35
7.6 Summary 7-39

CHAPTER 8 - Conclusions 8-1

8.1 Summary and Conclusions 8-1
8.2 Possible Extensions and Topics of Further Study 8-6

REFERENCES R-1

APPENDIX A-1

A.1 Introduction A-1
A.2 Custom Simulator / Analytical Model Package A-3
A.3 Workings of the simulator A-5

LIST OF FIGURES / TABLES

CHAPTER 2

Figure 2.1: Throughput vs. Offered Load for ALOHA, S-ALOHA, and non-persistent CSMA for different values of delay (d=0,1,5) 2-16

Figure 2.2a: Throughput vs. Offered Load for narrowband S-ALOHA and S-SS/ALOHA for various processing gains (N) 2-22

Figure 2.2b: Normalised Throughput vs Normalised Offered Load for S-ALOHA and S-SS/ALOHA with different values of N 2-22

Figure 2.3: Frame structure for the Reservation-ALOHA protocol 2-27

Figure 2.4: Example of a typical frame structure for the class of DAMA protocols based on Robert's Reservation 2-28

CHAPTER 3

Figure 3.1: Synchronisation and packetisation of messages in the reverse link........ 3-5

Figure 3.2: Illustration of the DT/CDMA MAC protocol 3-9

Figure 3.3: Illustration of a Transmission Period 3-14

Figure 3.4: Simulated data load showing several TP's........ 3-15

Figure 3.5: Markov chain for data users 3-16

Figure 3.6: Effect of β on the data message throughput (values of $\beta \leq 17$) 3-32

Figure 3.7: Effect of β on the data message throughput (β>17) 3-33

Figure 3.8: Effect of varying α on the system throughput........ 3-34

Figure 3.9: Distribution of TP lengths for various values of α at G_d=12 3-36

Figure 3.10: Distribution of message success probabilities for α=17 and α=3 at G_d=12.. 3-38

Figure 3.11: Optimum packet throughput curve obtained by using a dynamic α threshold 3-39

Figure 3.12: Relationship between optimum α value and G_d for $\overline{L}_d$=4 3-39

Figure 3.13: Optimum message throughput curve obtained by using a dynamic α threshold 3-40

Figure 3.14: Average message delay vs. offered load for various values of α........ 3-41

Figure 3.15: Effect of different mean message lengths on the system performance........ 3-43

Figure 3.16: Comparison of proposed data model with other common CDMA models.... 3-44

Table 3.1: Example network parameters used in the analytical evaluation and simulation 3-28

Table 3.2: Probability of bit error versus the number of transmitting users, k as computed using the SGA expression........ 3-29

Table 3.3: Performance measurements for the full range of possible α values ($0 \leq \alpha \leq \beta$) at a fixed offered load value (G_d=12 and β=17). 3-35

CHAPTER 4

Figure 4.1: Illustration of the SR-DT/CDMA MAC policy 4-2

Figure 4.2: Simulation results of the SR-DT/CDMA protocol with arbitrary network parameters showing new, retransmitted and combined message length distributions. 4-6

Figure 4.3: Flow diagram for messages and packets in the network [Raychaudhuri, 1987] 4-12
Figure 4.4: Data system throughput plots for various values of β 4-22
Figure 4.5a: Data system throughput plots for various values of β close to the optimum threshold (infinite population model) 4-24
Figure 4.5b: Data system throughput plots for various values of β close to the optimum threshold (finite population model) 4-24
Figure 4.6a: Effect of the α threshold on the system throughput (finite population) 4-26
Figure 4.6b: Effect of the α threshold on the system throughput (infinite population) 4-26
Figure 4.7: Effect of the α threshold on the average message delay 4-27
Figure 4.8: Example packet and message flow equation contours for some of the equilibrium solutions plotted on figures 4.6a and 4.7. 4-29
Figure 4.9: Comparison between DT/CDMA and SR-DT/CDMA 4-33

Table 4.1: Table of known and unknown variables at the outset of the analysis 4-10
Table 4.2: Example network parameters used in the analytical evaluation and simulation 4-21
Table 4.3: Comparison between analytical model (AN) results and simulation (SIM) results for the points on the α=6 curve of figures 4.6a and 4.7. 4-31
Table 4.4: Comparison between analytical model (AN) results and simulation (SIM) results for the points on the α=14 curve of figures 4.6a and 4.7. 4-31
Table 4.5: Comparison between analytical model (AN) results and simulation (SIM) results for the points on the $\alpha=\beta$=20 curve of figures 4.6a and 4.7. 4-31

CHAPTER 5

Figure 5.1: Illustration of the effects of hangover on talkspurt lengths 5-3
Figure 5.2: Transmission of data messages during silent periods in speech. 5-4
Figure 5.3: Divided frame structure for voice/data integration 5-7
Figure 5.4: Typical moving boundary frame structure 5-8
Figure 5.5: Frame structure for the DTDMA/TDD Wireless ATM MAC protocol 5-9
Figure 5.6: Example of the uplink frame structure used in WISPER (obtained from [Akyildiz *et al*, 1999]) 5-17

CHAPTER 6

Figure 6.1: Illustration of the voice call admission policy 6-4
Figure 6.2: Voice model showing voice terminal states and state transition probabilities . 6-5
Figure 6.3: Illustration of flow of users into and out of the different voice states 6-13
Figure 6.4: Voice model for the case where there is no voice activity factor exploitation (or equivalently, the VAF=100%) 6-23
Figure 6.5: Average carried voice and talkspurt loads ($\bar{v}$ and $\bar{t}$) versus the offered voice call load (G_v), for various voice activity factors (10%, 30%, 60%, and 90%) 6-26
Figure 6.6: Expected voice call blocking probability versus the offered voice call load 6-27
Figure 6.7a: Data packet throughput for a light voice load ($\bar{v}$=6 calls) and various VAF 6-29
Figure 6.7b: Data message throughput for a light voice load ($\bar{v}$=6 calls) and various VAF 6-29

Figure 6.8a: Data packet throughput for a heavy voice load ($\bar{v}$ =15 calls) and various VAF 6-30
Figure 6.8b: Data message throughput for a heavy voice load ($\bar{v}$ =15 calls) and various VAF 6-30
Figure 6.9: Effect of α on the data throughput performance for the case where there is voice traffic 6-32

Table 6.1: Imaginary network parameters used in the analytical evaluation and simulation 6-25

CHAPTER 7

Figure 7.1: The traditional two-state Gilbert-Elliot channel model 7-3
Figure 7.2: Illustration of a narrowband channel being defined as "good" or "bad" 7-4
Figure 7.3: The traditional generic Fritchman model 7-5
Figure 7.4: Simulation sample sequence of values for *j*, *U* and *Y* 7-11
Figure 7.5: Illustration of the relationship between the received signal amplitudes (arbitrary units) from both a reference user (*U*) and the total other-user interference (*Y*), and the resultant transitions between the imagined Good (G) and Bad (B) states. θ is the *fading ratio* 7-13
Figure 7.6: Proposed multi-user Markov chain for the channel model 7-15
Figure 7.7: Probability of packet success vs. the number of transmitting users for various θ (course resolution). 7-27
Figure 7.8: Probability of packet success vs. the number of transmitting users for various θ (fine resolution). 7-28
Figure 7.9: Packet success probabilities for the *MM* approach for the value of θ which provides the closest match to the *EX* and simulation results. 7-28
Figure 7.10a: Packet success probabilities from the *MM* for Q=10000 bits. $\theta \approx$ 22dB provides a satisfactory match to the exact (*EX*) and simulation results. 7-30
Figure 7.10b: Bit error probability (*Q*=1) for the *MM* (θ=44dB) 7-30
Figure 7.11: Data packet throughput for the perfect and Rayleigh fading channels (collision detection only, no blocking : α=β=20) 7-31
Figure 7.12: Data packet throughput for the perfect and Rayleigh fading channels (collision detection, with blocking : α=12, β=20) 7-33
Figure 7.13: Optimum data packet throughput for the Rayleigh fading channel obtained by using a variable α threshold 7-33
Figure 7.14: Comparison of the optimum data packet throughputs for the perfect and Rayleigh fading channels 7-34
Figure 7.15a: Message success probabilities for the low offered load scenario 7-36
Figure 7.15b: Message success probabilities for the medium offered load scenario 7-36
Figure 7.15c: Message success probabilities for the high offered load scenario 7-37
Figure 7.16: Message throughput for the correlated Rayleigh fading channel for several different values of $f_D t_s$ 7-38

APPENDIX

Figure A.1: Example screenshots of the custom simulator and analytical model package A-4

LIST OF ACRONYMS

AMPS	: Advanced Mobile Phone System
ARQ	: Automatic Repeat Request
ATM	: Asynchronous Transfer Mode
AWGN	: Additive White Gaussian Noise
B-ISDN	: Broadband Integrated Services Digital Networking
BPSK	: Binary Phase Shift Keying
BER	: Bit Error Rate
BTMA	: Busy Tone Multiple Access
CDMA	: Code Division Multiple Access
CLS	: Channel Load Sensing
CRA	: Collision Resolution Algorithm
CSMA	: Carrier-Sense Multiple Access
CSMA/CD	: Carrier-Sense Multiple Access with Collision Detection
CT-2	: Cordless Telephone 2
DAMA	: Demand Assignment Multiple Access
DB	: Data Blocking
DC	: Data Collision
DECT	: Digital European Cordless Telephony
DFT	: Delayed First Transmission
DS-CDMA	: Direct Sequence Code Division Multiple Access
DSI	: Digital Speech Interpolation
DSP	: Digital Signal Processing
DS/SS	: Direct-sequence Spread-spectrum
DT/CDMA	: Dual-Threshold CDMA
EPA	: Equilibrium Point Analysis
ETSI	: European Telecommunications Standards Institute
FDD	: Frequency Division Duplex
FDMA	: Frequency Division Multiple Access
FEC	: Forward Error Correction
FH/SS	: Frequency-hopping Spread-spectrum
GPS	: Global Positioning System
GSM	: Global System for Mobile Communications
HMM	: Hidden Markov Model
ICMA	: Idle-Signal Casting Multiple Access
IGA	: Improved Gaussian Approximation
IP	: Internet Protocol
IMT-2000	: International Mobile Telecommunications - 2000
ISDN	: Integrated Services Digital Network
ISMA	: Inhibit-sense Multiple Access
ITU	: International Telecommunications Union
LAN	: Local Area Network
MAC	: Medium Access Control
MAI	: Multiple Access Interference
MAN	: Metropolitan Area Network
MM	: Markov Model
NADC	: North American Digital Cellular

PBX	Private Branch Exchange
PDC	Personal Digital Cellular
PDVD	Packetised Data - Voice Dedicated
PN	Pseudo-random Noise
PRMA	Packet Reservation Multiple Access
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
R-ALOHA	Reservation-ALOHA
S-ALOHA	Slotted-ALOHA
SGA	Standard Gaussian Approximation
SIR	Signal-to-Interference Ratio
SNR	Signal-to-Noise Ratio
SR-DT/CDMA	Selective-Repeat Dual-Threshold CDMA
SRQ	Selective-Repeat Request
SS/ALOHA	Spread-spectrum ALOHA
S-SS/ALOHA	Slotted Spread-spectrum ALOHA
TASI	Time Assignment Speech Interpolation
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TP	Transmission Period
UMTS	Universal Mobile Telecommunications System
UPCN	Universal Personal Communications Network
VAF	Voice Activity Factor
WAN	Wide Area Network
W-CDMA	Wideband Code Division Multiple Access
WIMA	Wireless Integrated Multiple Access

CHAPTER 1

INTRODUCTION

1.1 Background

For many years it has been a somewhat surreal topic amongst communications engineers to provide a global "Universal Personal Communications Network (UPCN)" – a world in which any person can access a vast array of near-instant services and information types, or communicate with any other person, at any place and any time, through any medium using only a single hand-held wireless terminal. This world of modern and future telecommunications is very much driven by the concept of a wired backbone superhighway infrastructure that is accessed via local wired and wireless *access networks*. The wired backbone infrastructure will merely be a natural evolution and merging of the existing local-area, metropolitan-area, wide-area (LAN, MAN, WAN), Internet and telephony networks to support a wide variety of services over a single network infrastructure. In future years, this infrastructure may consist mainly of high bandwidth fiber-optic cables and satellite systems supporting a transmission standard such as IP (Internet Protocol), ATM (Asynchronous Transfer Mode) or B-ISDN (Broadband-Integrated Services Digital Networking).

Until recently, this notion of a global UPCN had been thought to be impractical due to the daunting scale of the system, the extremely large bandwidth requirements, and the challenges associated with integrating different services with contrasting transmission requirements. However, recent advances in the fields of satellite communications, fiber-optic transmission mediums, information coding schemes (e.g. the invention of turbo codes), signal processing, and information compression and modulation techniques that attempt to compress as many bits of information as possible into a unit of bandwidth, have allowed for increased available bandwidth, improved efficiency of bandwidth usage, and reduced burdening on the network resources. These advances have placed a viable solution to the UPCN within a tangible grasp.

1.1.1 Integrated Services Networking

A significant problem that has faced communications engineers over the years is that of efficiently integrating a multitude of different services with diverse traffic profiles and contrasting Quality of Service (QoS) requirements on to the same communications medium. Examples of such services might be: text, e-mail, digitised voice (telephony), video, video telephony and video conferencing, music on demand, World Wide Web pages, data files and control data (signalling). The QoS is the end-user's qualitative evaluation of the ease and quality of communication in the network. The QoS requirements specify factors such as the information transmission rate, maximum acceptable bit error rate (noise), end-to-end delay constraints, delay variation constraints, information retransmission requirements, and the worst case call blocking/dropping probabilities.

Let us consider, for example, a typical digital voice telephony service. We may list the transmission requirements for digital voice as follows. Generally, digital voice requires a relatively low bit transmission rate; typically in the range 4kbits/s to 32kbits/s, depending on the quality requirements and voice coding scheme (voice CODEC). In order to ensure a real-time, fluent conversation, there must be a very low end-to-end delay with a minimal variation in delay. Finally, voice can tolerate moderately high bit error rates without the need to retransmit corrupt information. A typical video telephony or video on demand service would have similar requirements, although the bit transmission rates would need to be a lot higher due to the inclusion of visual information (64kbits/s up to several Megabits/s depending on the image size, image resolution and quality, frame rate, depth of colour, video compression technique etc...). Traditionally, services that have stringent delay constraints, such as voice traffic, have been classified as *real-time services*. *Circuit switched networks* have traditionally been better suited to supporting real-time services.

As a direct contrast, a data service such as e-mail or a Web page can generally tolerate moderately high end-to-end transmission delays with large variations in the end-to-end delay, but requires very low bit error rates and generally requires the retransmission of any corrupt or irrecoverable information. Also, depending on the type of service, data bit rate requirements can vary dramatically. A bit rate of 1kbit/s might be sufficient for the transmission of small text files or e-mail, but is grossly inadequate for the transmission of large data files exceeding several Megabytes in size. Time non-critical data services are usually referred to as *non real-time services*. *Packet switched networks* have traditionally been better suited to supporting non real-time services.

Although circuit switched telephony networks are supportive of low bit rate data communications, the bursty nature and relatively low percentage channel usage of interactive data traffic make grossly inefficient use of circuit switched networks. In standard wired networks, the only inconvenience is having to pay for the large proportion of idle time between data interactions. In a cellular system, the effects of data integration are more serious. Limited cell capacities, increased call rates and the fact that users share the limited number of channels results in increased wastage. Not only is a data user wasting his own paid for bandwidth, but also denying access to other users who could be using the idle channel.

Efforts in the field of integrated services networking have co-ordinated unifying headings such as ISDN (Integrated Systems Digital Networking), ATM, IP and, more recently, IMT-2000 (International Mobile Telecommunications-2000) and UMTS (Universal Mobile Telecommunications System). These standards provide structured protocols for supporting a large variety of information types with diverse characteristics and transmission requirements.

1.1.2 Wireless Access Networks

The evolution of telecommunications towards the mobile (wireless) domain has inspired much research in the fields of radio and satellite communications, cellular networks and packet radio networks. An important component in the overall network hierarchy is the *access network*. As the name implies, wireless access networks provide the initial (or final) “hop”, via a flexible, convenient and ubiquitous wireless subscriber interface, from/to the subscriber’s mobile (or fixed location) wireless terminal to/from the backbone access point (such as a base station, file server, hub, etc…). Wireless access techniques now find application in a variety of systems, including cellular telephony networks, access to satellites, wireless LAN's, and wireless PBX applications (e.g. DECT (Digital European Cordless Telephony)). In evolving towards the UPCN concept, we will be moving into the third generation of wireless access networks.

The *first generation* of mobile communications systems were characterised by analog transmission of voice only, and had severe capacity limitations. An example of a first generation application for wireless cellular telephony is the AMPS (Advanced Mobile Phone System) system, which offered 666 analog FM channels (42 control and 624 voice channels) [Goodman, 1990]. The conversion from analog to digital allowed for improved time division multiplexing techniques, increased network capacity and the inclusion of digital error detection and correction schemes. A variety of *second generation* networks accrued as a result [Goodman, 1991], all dedicated to supporting digital voice only but designed for different operating environments. Throughout the world, several predominant second generation cellular voice standards have emerged such as: IS-54 (Digital-AMPS), GSM (Global System for Mobile Communications, European origin), IS-136 (North American Digital Cellular), PDC (Personal Digital Cellular, Japan), and IS-95 CDMA (USA, Asia Pacific countries). For cordless telephony and wireless PBX applications, the DECT (indoor and outdoor) and CT-2 (indoor cordless telephony standard-2) standards have been successfully deployed throughout the world. An unfortunate consequence of this multiplicity of standards is that they offer very little inter-compatibility, each require specific hardware (separate mobile terminals for each standard), and are designed for different cell dimensions and populations sizes.

The sudden explosion of the Internet and World Wide Web over the last few years has prompted an increased demand for Internet connectivity in both wired and wireless domains. *Third generation* wireless networks are deemed to support a vast variety of information types, operate in all transmission environments and be capable of supporting a large population with a highly dynamic profile. Unlike first and second generation networks which are designed for cell sizes of several kilometers, third generation networks are likely to consist of cells which are several orders of magnitude smaller. This allows densely populated metropolitan areas to be serviced with a higher capacity. In order for third generation access networks to be a success, a standard must be specified to allow global inter-compatibility and inter-networking. Only very recently, with the need and demand for third generation systems continually increasing, have researchers started striving for a common global standard that will be able to meet these requirements. At the time of writing this thesis, the third generation cellular standards of choice appear to be IMT-2000 and UMTS. IMT-2000 has been proposed and is in the process of being standardised by the International Telecommunications Union (ITU) and appears to be the standard of choice for North America and the Asia Pacific countries. The prime candidate access protocol standard for IMT-2000 appears to be CDMA-2000, which will be a wideband evolution of the IS-95 CDMA standard. Concurrently, the European Telecommunications Standards Institute (ETSI) has proposed the UMTS as an alternative. Both systems are being designed with stringent inter-compatibility, a high degree of worldwide design commonality, world roaming, and a large array of high quality services in mind.

In wireless networking comes a vast array of new challenges. In contrast to wired networks, the properties of wireless channels are highly unpredictable and time varying. At low transmission frequencies, signal propagation is most penetrative, but practical antennae sizes become the limiting factor. At high frequencies, radio signals are more easily received but the propagation of the signal is strongly influenced by many factors such as: the distance between the transmitter and receiver; the geographical obstructions and terrain over and through which the signal traverses (responsible for fading, multipath distortion, and shadowing of the signal); and interference from adjacent signals operating in the same frequency band. In a mobile environment, some of these effects are slowly time-varying, while others may be rapidly time-varying.

1.1.3 The Medium Access Control (MAC) Protocol

A key technical issue related to wireless access networks is the selection and implementation of a suitable medium access and medium sharing control protocol. The radio channel is inherently a single ether, or free space channel, and as such cannot support multiple radio signals without separating them in some way. The exact process by which the signal separation is co-ordinated to allow more than one person to simultaneously access the radio medium is known as the *Multiple Access*, or *Medium*

Access Control (MAC) *Protocol*, and is undeniably one of the most important aspects of any multi-user communications system. In addition to being a single ether, the available ether medium which, in a radio environment is, of course, the allocated (or useable) electromagnetic spectrum, is a *limited* resource. The MAC protocol is thus also responsible for invoking a procedure to control, distribute and co-ordinate the use of the limited resource among several users. The topic of MAC protocols is vast, with many inventive and efficient schemes being proposed and implemented. Some of the more noteworthy schemes are reviewed in Chapters 2 and 5 of this thesis.

MAC protocols based on Code Division Multiple Access (CDMA) are emerging as a promising alternative to conventional second-generation TDMA (Time Division Multiple Access) and FDMA (Frequency Division Multiple Access) MAC protocols. Since the commencement of this thesis project, especially over the last two years, the popularity of CDMA has enjoyed a tremendous growth worldwide. For example, subscriber counts and growth rates show that the cdmaOne cellular standard (the official brand name for the IS-95 CDMA standard) is fast becoming the dominant choice for second-generation digital communications in many countries, and has found tremendous commercial application worldwide. For the third generation of mobile telecommunications, CDMA appears to be by far the most popular choice. For the IMT-2000 standard, which is still in the process of being standardised at the time of writing this thesis (although almost fixed), the wideband-CDMA (W-CDMA) standard has been chosen as the air-interface. A key reason, amongst others, why CDMA has become popular is that inherent in CDMA systems are several properties that are highly beneficial to cellular mobile networks. These include: improved frequency reuse (better spectral efficiency), information privacy, lower transmission power requirements (longer battery life) and immunity to a wide variety of channel adversities. Until recently, CDMA had been significantly more difficult to implement than conventional narrowband TDMA and FDMA. However, the advent of sophisticated DSP (digital signal processing) hardware, improved power control schemes, and improved receiver structures (e.g. RAKE receivers) have narrowed this implementation gap significantly.

1.2 Focus and Goals of this Thesis

At the commencement of this research project, CDMA cellular systems were still very much in their infancy and, although CDMA had received much literary attention for first and second generation voice cellular systems as well as packet radio data-only networks, considerably less research had been done on MAC protocols for integrating multiple services using CDMA as an under-lying multiple access protocol in a cellular/LAN environment. The work that had appeared had either focused on simple satellite implementations and protocols, or had described protocols that were somewhat elementary or idealised. The research this thesis presents was motivated by the evolution of mobile communications towards global support of multimedia services and the somewhat lacking amount of work done in this field for CDMA radio networks at the time of commencement of this project.

The focus of this thesis falls on the study of MAC protocols and the derivation and evaluation of a MAC protocol for a CDMA wireless access network supporting data and voice traffic. We focus on integrating voice and data only, as this addresses the fundamental problem of integrating a "long term", real-time service that has traditionally been carried over circuit switched networks with a "short term", non real-time packet orientated service that requires a secure and noiseless channel. The goals of this research are two-fold. The first goal is to contribute to the field of MAC protocols. To this end, we aim to derive a new protocol that improves in some way on the performance of existing schemes, where the specific class of existing schemes we will target is that of *random access techniques* for the shared reverse link of centralised wireless networks employing CDMA. The second goal of this research is to develop analytical models for the performance characterisation of our proposed MAC protocol. Analytical models are useful for obtaining quick predictions of the performance of a system, prior to simulation or actual implementation of the system. In addition to an analytical model, a simulation model is also developed, not only to evaluate the protocol performance, but also to evaluate and validate the results and approach of the analytical model.

It is well known that the practical capacity, or multiple access capability of CDMA is *interference limited* [Gilhousen *et al*, 1991], [Viterbi & Viterbi, 1993]. This is particularly true in a system such as the IS-95 system that uses conventional matched filter receivers instead of optimal multi-user receivers. In theory, optimal multi-user receivers do allow for a desired signal to be completely filtered out from several interfering signals. However, in practical situations where factors such as adverse channel influences (unequal signal delays and multipath etc...) and code acquisition and synchronisation imperfections occur, even theoretically optimal multi-user receivers cannot completely overcome other user interference. Thus, in CDMA, the more simultaneous transmissions carried by the channel, the higher the degree of *multiple access interference* (MAI) in the channel. As alluded to in the above, there is a maximum number of users, often referred to in the literature as the channel's *multiple access capability* [Geraniotis, Soroushnejad, & Yang, 1995] or the channel's *multiple access threshold*, such that this level of MAI is just below some maximum acceptable value [Fapojuwo, 1993], [Gilhousen *et al*, 1991], [Viterbi & Viterbi, 1993], [Geraniotis, Soroushnejad, & Yang, 1995]. The occasion when the number of users exceeds this multiple access threshold is often termed as a channel *overload*, channel *collision* [Lam & O'Farrell, 1992], or channel *outage* [Gilhousen *et al*, 1991], [Fapojuwo, 1993]. A literature review on existing random access techniques for a shared CDMA channel (for both fully-connected and centralised network architectures) revealed two fundamental approaches to avoiding channel overload occurrences. The first method is to allow uncontested access to the channel and abort all transmissions in the event of such an overload case [Resheff & Rubin, 1990], [Lam & O'Farrell, 1992]. The second method is to allow uncontested access to the channel up to a point, and to block further access when the number of users exceeds some maximum threshold (e.g. [Toshimitsu *et al*, 1994], [Abdelmonem & Saadawi, 1989], [Geraniotis, Soroushnejad & Yang, 1995]). In this thesis, we will explain the advantages and disadvantages of these two schemes, and derive a new MAC protocol for a CDMA channel that implements *both* of the above threshold schemes simultaneously. Subsequently, we name our proposed protocol "Dual-Threshold CDMA" (abbreviated to DT/CDMA).

In our initial evaluation of DT/CDMA, we will consider a data-only network. Justification for this decision is as follows. As discussed in section 1.1.1, the components of a multi-service system must accommodate the transmission requirements of all the services it supports in the best way it can, and the MAC protocol component is no exception in this regard. In presenting a more thorough review of MAC protocols in this thesis, it should become obvious to the reader that MAC policies for data networks are inherently, and have traditionally been, more complicated than policies for voice-only networks. This phenomenon arises primarily from the fact that data traffic consists of short messages that are generated in an inherently bursty fashion, and that generally require retransmission if corrupt. This, coupled with the fact that our dual-service MAC protocol is derived from a data-only MAC protocol, results in much of the focus of this thesis falling on the data MAC policy. In fact, the nature of our protocol is such that the data admission policy can be applied in the absence of any voice traffic, thus making it highly suitable for a wireless LAN or cellular network scenario that supports data traffic only. In addition, our data-only MAC protocol contributes a novel concept that improves on the performance of similar existing CDMA data-only protocols. These same performance improvements also apply in the presence of voice traffic. This allows us to isolate and compare the effects that our proposed dual-threshold concept has on the data network performance. Our second motivation for isolating the data component at first is based on the analytical modelling of the MAC protocol. The protocol analysis is complicated by the integration of voice traffic, but is easier to follow if presented for the data-only network. Once the basis analytical concepts and procedure have been derived for the data-only case, it is easier to extend the model to incorporate both voice and data traffic.

1.3 Document Overview and Layout

There are eight chapters in this thesis document. Chapter 1, this chapter, introduced the concepts, challenges and future requirements of a global information infrastructure incorporating the next generation of wireless networks required to provide ubiquitous access to such an infrastructure. The focus and goals of, and the motivation for this thesis research were also presented.

Chapter 2 is a literature survey on traditional MAC protocols, with particular focus on those relevant for single-service (e.g. data only) wireless applications. A discussion on the multiple access issues associated with direct-sequence spread-spectrum communications is presented first, since these issues are pertinent to most of the work this thesis presents, our proposed MAC protocol in particular. Thereafter, a discussion on existing MAC techniques is presented, starting with the fixed channel assignment schemes (TDMA, FDMA and CDMA), moving on to the random access protocols (ALOHA, CSMA, CSMA/CD, ISMA and their spread-spectrum equivalents), and ending with the demand assignment/reservation MAC schemes (R-ALOHA, PRMA and such like). We defer a review of MAC protocols for integrating voice and data, as well as some of the more recent and more advanced multi-service MAC protocols, until Chapter 5. This has been done to break down the literature review and perhaps relax the reader by creating an insert of relatively easier reading into the somewhat mathematical and technical analyses that appear in Chapters 3, 4, 6 and 7.

In Chapter 3, we present in detail our proposed MAC policy for a CDMA data-only packet radio network, called "Dual-Threshold CDMA" (DT/CDMA). Particular emphasis is placed on the derivation of an analytical model (using a Markov modelling approach) of the performance of DT/CDMA. As will be discussed, we assume an infinite population traffic model with a Poisson arrival process in order to simplify the analytical performance characterisation of DT/CDMA. The modelling of DT/CDMA becomes analytically infeasible for more complicated finite population models that consider modelling the traffic component generated by retransmitted messages. The results of the derived Markov model are compared to results obtained from a custom designed software simulation, for a typical set of imaginary network parameters. Finally, we present a comparison of our proposed protocol with other similar protocols proposed in the literature, and quantify the performance enhancements that can be gained through the application of our dual-threshold scheme.

In Chapter 4, we evaluate the performance of DT/CDMA in which an improved *selective-repeat retransmission scheme* is implemented. As a result of this modification, (1) the MAC protocol becomes more efficient, and (2) it allows us to analyse the network performance for a more realistic finite population traffic model that incorporates both new and retransmitted messages. We call this improved protocol "Selective-Repeat Dual-Threshold CDMA" (SR-DT/CDMA). A combined Equilibrium-Point/Markov analysis is derived to evaluate the performance of our SR-DT/CDMA MAC protocol for a finite population model. This model allows us to consider, amongst other things, the stability of the data network and a more accurate representation of the combined traffic arrival process of the network. Again, analytical results are compared to simulation results.

Chapter 5 presents a literature review on relevant traditional and modern protocols that have been proposed for multi-service networks (e.g. voice and data).

In Chapter 6, we extend our proposed DT/CDMA (or SR-DT/CDMA) protocol to incorporate the integration of voice and data traffic over the same shared CDMA channel. We present the network model, traffic models for both services and our proposed service admission policies. Our proposed model is loosely based on the scheme proposed by the group of authors Geraniotis, Yang and Soroushnejad (e.g. [Soroushnejad & Geraniotis, 1995] discussed in Chapter 5). We diverge from the aforementioned model in that: (1) we base our voice/data scheme on our DT/CDMA (or SR-DT/CDMA) protocol, (2) we model and include an evaluation on the effects of exploiting the silent periods inherent in human speech in order to improve the multiple access capacity for the data traffic. A literature survey suggested that, although much work has been done on exploiting voice silence periods for the benefit of data traffic in narrowband TDMA based systems, few (if any) analytical models have appeared for joint voice/data CDMA protocols of the nature proposed in this thesis. A significant emphasis is thus placed on the derivation of a joint-state Markov analysis for the purpose of modelling the silence periods and their effects on the data protocol. Results of the joint voice/data network performance, from both the Markov model and simulation model, are presented and compared.

In Chapters 3,4 and 6, an ideal channel is assumed in all analyses and simulation results, for the purpose of isolating the performance of our proposed MAC protocols. In Chapter 7 we investigate the effects that a correlated multipath Rayleigh fading channel has on the performance our MAC protocol. To aid in doing this, we generate a simple

Hidden Markov Model (HMM) to simplify the modelling of the fading process and resultant error process that occurs in the multi-user CDMA channel. This HMM is based on the popular two-state narrowband Gilbert-Elliot Markov channel model although, in our CDMA case, we model the channel interference as a time varying value (due to the fact that it is a function of the number of users in the channel), rather than some time invariant value (as assumed by Gilbert to be associated with the time-invariant AWGN component of a narrowband channel). We present the details of our proposed HMM channel model, and derive approximations for the parameterisation of the model for a channel in which the fading is relatively slow and the sequence of fading samples is highly correlated. We show that our proposed HMM is surprisingly accurate considering the gross simplification that is made of the fading process and the MAI in the channel. Finally, we evaluate the effect that Rayleigh fading has on the system performance, when compared to a perfect, non-fading channel.

Finally, Chapter 8 presents the research conclusions drawn in this thesis and some recommendations for further study.

The appendix contains a brief discussion on the topic of simulators and presents details of the simulation models and algorithms that are pertinent to this dissertation.

1.4 Original Contributions of the Thesis

The original contributions of this thesis may be summarised as follows:

1. The derivation of a novel MAC protocol, called DT/CDMA, for the reverse link of a CDMA data packet radio network (presented in Chapter 3). The originality of our protocol lies in the implementation of a centralised channel load sensing/broadcast scheme that operates in conjunction with a dual threshold user admission scheme - a fixed channel-overload detection (collision) threshold and a variable, load-dependent channel access regulation (blocking) threshold.

2. The derivation of an accurate statistical Markov analysis (presented in Chapter 3) to predict the performance of the proposed DT/CDMA protocol for a simple Poisson traffic process.

3. The derivation of a second MAC protocol called SR-DT/CDMA (presented in Chapter 4), that improves on our original DT/CDMA protocol by implementing a selective-repeat retransmission scheme.

4. The derivation of an Equilibrium Point/Markov analysis (presented in Chapter 4) to predict the performance of the proposed SR-DT/CDMA protocol for a more realistic finite population traffic model.

5. The derivation of a joint voice/data access protocol based on our proposed SR-DT/CDMA (or DT/CDMA) MAC policy (presented in Chapter 6). This includes the suppression of voice transmissions during the silence periods in the speech stream to improve the multiple access capacity of the data channel.

6. The derivation of an accurate Markov analysis to predict the performance of the integrated voice/data MAC protocol. In particular, we derive a joint-state Markov model to evaluate the performance improvements that can be gained from exploiting the voice activity factor of human speech to improve the multiple access capacity for the data protocol. This appears never to have been done for a combined voice/data random access CDMA protocol.

7. The derivation of a simple hidden Markov channel model for a multi-user CDMA channel, adapted from the traditional narrowband Gilbert-Elliot Markov model, to predict the effect that correlated Rayleigh fading has on the performance of our protocol. This technique may be applied to any slotted protocol in which Rayleigh fading is considered.

Parts of the research presented in this thesis have been presented by the author at the following local and international conferences:

1. Judge, G & Takawira, F, (1996) "A protocol for voice/data integration over a CDMA packet radio network", *Proceedings of the local TELETRAFFIC'96 Symposium*, Durban S.A., pp. 214-220, September 1996

2. Judge, G & Takawira, F, (1997), "Performance analysis of a multiple-access protocol for a packet switched CDMA packet radio network using channel load sensing", *Proceedings of IEEE COMSIG'97 Conference*, Grahamstown S.A., pp. 121-126, September 1997

3. Judge, G & Takawira, F, (1998a), "Performance analysis of a channel load sensing protocol for CDMA packet radio networks with finite population and variable length messages", *Proceedings of IEEE ISSSTA'98 Conference*, Sun City S.A., pp 282-286, September 1998

4. Judge, G & Takawira, F, (1999a), "Performance Analysis of a Novel Channel Load Sensing Protocol for CDMA Packet Radio Networks Supporting Voice and Data Traffic Integration", *Proceedings of PIMRC'99 Conference*, Osaka, Japan, pp 1295-1299, September 1999

The following three journal papers have been submitted for review, with the first one being accepted for publication and the other two being submitted/reviewed at the time of writing this thesis document:

1. Judge, G & Takawira, F, (1998b), "Spread-Spectrum CDMA Packet Radio MAC Protocol Using Channel Overload Detection and Blocking", *To appear in Wireless Networks* (ACM/Baltzer Publishers)

2. Judge, G & Takawira, F, (1999b), "A Markov Analysis of a Voice/Data MAC Protocol for a CDMA Packet Radio Network", *Submitted to IEEE Transactions on Vehicular Technology*

3. Judge, G & Takawira, F, (1999c), "A Simple Hidden Markov Model for a CDMA Channel with Correlated Rayleigh Fading", *Submitted to Wireless Networks* (ACM/Baltzer Publishers)

CHAPTER 2

TRADITIONAL MEDIUM ACCESS CONTROL (MAC) PROTOCOLS

2.1 Introduction

It is a well known fact that a receiver is unable to separate two or more signals if they are received simultaneously, in the same frequency band, using the same modulation technique and with similar powers (the capture effect does allow signals with overly dominant power levels to be separated from significantly weaker signals). Furthermore, any communications medium resource (whether it be a copper cable, fibre-optic cable, micro-wave wave-guide or a radio channel) is invariably limited due to limitations on the usable (available) bandwidth of the transmission medium and limitations on the transmittable signal bandwidth imposed by the system hardware (e.g. practical antenna sizes, DSP processing speed limitations etc...). The goals of the *Medium Access Control* (MAC) *protocol*, also often referred to as the *Multiple Access Protocol*, are two-fold. Firstly, the MAC protocol is responsible for employing a technique(s) to separate several simultaneous transmissions in order to allow *multiple access* to a shared channel. Secondly, the MAC protocol should invoke a resource sharing procedure that is responsible for co-ordinating and controlling access to the medium in a manner that is hopefully fair to all users in the channel.

This chapter presents a brief literature survey on traditional MAC techniques, focusing particularly on random access protocols for wireless networks. We begin in section 2.2 by discussing multiple access issues in direct-sequence spread-spectrum communications pertinent to this dissertation. In section 2.3 we present traditional fixed assignment schemes such as TDMA, FDMA and CDMA. Sections 2.4 and 2.5 present traditional random access protocols for both narrowband (ALOHA, CSMA, ISMA etc...) and spread-spectrum (SS/ALOHA, CLS etc...) systems. Section 2.6 concludes with the class of protocols based on demand assignment/reservation.

2.2 Multiple Access Issues in Direct-Sequence Spread-Spectrum (DS/SS) Communications

A multiplicity of benefits accrue from applying spread-spectrum techniques [Yang & Geraniotis, 1994], [Prasad, 1996], [Glisic & Vucetic, 1997]. Of these advantages, the most important are: an immunity to inband interference, multipath distortion, and frequency selective fading; anti-jamming capabilities; and multiple user random access communications with selective addressing. Of the two traditional methods by which to create a spread-spectrum signal: Direct-Sequence Spread-Spectrum (DS/SS) and Frequency-Hopping Spread-Spectrum (FH/SS) (see [Prasad, 1996] and [Glisic & Vucetic, 1997] for details on these methods), we choose to focus on DS/SS multiple access issues in this thesis.

2.2.1 Multiple channel access in DS/SS systems

In DS/SS, the original message bit stream is *amplitude modulated* by a significantly higher rate *pseudo-random noise* (PN) sequence. If the PN sequence is such that there are N "chips" per message bit, then the "processing gain" of the system is simply equal to N. DS/SS is known as an interference averaging technique, as interference is reduced by averaging the interference over a larger time interval. One of the most important advantages obtained from spreading the spectrum using pseudo-random noise sequences is that more than one user can transmit at the same time over the same frequency range. In order to be able to separate signals however, there has to be a domain in which signal separation can occur. In spread-spectrum schemes, this is done in the *code domain.* To allow this, each user must use a different PN sequence and each PN sequence must have very specific auto-correlation and cross correlation properties with every other PN sequence in the PN code set [Prasad, 1996], [Glisic & Vucetic, 1997].

The manner in which PN codes are allocated to users, as well as the manner in which terminals are addressed is also an important consideration in a wireless CDMA system. The code allocation scheme is an important parameter in the overall design of a multi-user spread-spectrum system (i.e. a CDMA system). The aim of a code allocation scheme is to avoid collisions in the code domain, where two or more terminals attempt to transmit using the same spreading code. This inevitably results in a receiver being unable to separate the two signals. A collision in the code domain is analogous to the collision of two narrowband signals in the time and frequency domains. Several code allocation and addressing schemes exist, such as: Common code, Receiver-Based code, Transmitter-Based code, Common-Transmitter-Based protocol and Receiver-Transmitter-Based protocol. We refer the interested reader to [Sousa & Silvester, 1988] for a comprehensive description and comparison of these methods.

The majority of second and third generation mobile systems are assumed to employ a base station of some kind. The base station is generally assumed to allocate codes to the users in its cell using a separate signalling channel. A base station might have a code pool of say *M* unique PN codes in anticipation of supporting a maximum population of *M* users. As users enter the cell they are allocated a unique code from the pool, and when they leave they return the code to the pool. The base station is then aware of each user in the network, and the code allocated to them. All communications are done via the base station and are performed using a mobile transmitter-based code protocol [Sousa & Silvester, 1988]. The base station implements a multi-user detector (with up to *M* channels, each synchronised to one of the *M* codes) in order to receive multiple transmissions. No two terminals can ever collide in the code domain, since each user is assumed to have a unique code. In mobile and cellular CDMA systems, inter-cell mobility implies dynamic population changes and the process of handover generally requires code allocation changes and updates. Another consideration of which to take cognisance is the fact that the size of a PN code set is limited in size due to the following reasons: firstly, PN sequences have a finite period and are hence limited by the number of possible combinations of binary values. Secondly, of the number of possible binary combinations, only a small subset satisfies the very specific auto-correlation and low-cross correlation properties required for PN sequences.

Furthermore, in practice a set of PN codes can never be perfectly orthogonal and as such each PN code will have a finite non-zero cross-correlation with every other code. This non-zero cross-correlation results in a receiver being unable to totally separate the wanted signal from the other received spread signals. Each unwanted signal appears as background interference. This process is mutual since each user in the network effects every other user in the network. The total summation of interference from all users is called the *Multiple Access Interference* (MAI), and is a time varying value as users enter and leave the channel. In spread-spectrum systems, one usually defines the *signal-to-interference ratio* (SIR) as the ratio between the power of the wanted signal to the combined power of the MAI as received from all other users in the network. As the number of transmitting users increases, the SIR decreases as each new user contributes more background noise to the wanted signal. This property of spread-spectrum systems is known as *graceful degradation*, where the system performance degrades gracefully as more users are added to the channel. Immediately, one can see that the capacity of a spread-spectrum system is *interference limited* rather than bandwidth limited. This means that the limit on the number of allowed simultaneous spread-spectrum transmissions is directly related to the maximum acceptable amount of MAI (or equivalently, the minimum acceptable SIR value). Computation of the relationship between the number of users and the quantity of total MAI is a highly complicated and multi-dimensional problem. The level of MAI, as observed by a certain receiving terminal, is dependent on a vast array of variables, the main contributing factors being: the number of received signals, the processing gain used for each signal, the received power of each signal, the modulation technique used in the system, the type of PN codes used, and the level of PN code correlation and synchronisation between users.

2.2.2 Multiple access capability of a DS/SS channel

As mentioned previously, an exact computation of the MAI is complex. However, several approximations have been derived. A popular approximation that is commonly used is the Standard Gaussian Approximation (SGA) [Pursley, 1977]. In the SGA it is assumed that the level of interference from all interfering users (those with codes unlike the desired code) is treated as additive Gaussian noise which has a uniform power spectral density over the band in which the system operates. The SGA then approximates the probability of bit error as experienced by a receiving user when there are K users in total transmitting as [Pursley, 1977], [Sunay & McLane, 1996]

$$P_{BE}^{SGA}(K) = Q\left(\left(\frac{1}{N\sigma}\sum_{k=2}^{K}\frac{E_{bk}}{E_{b1}} + \frac{N_0}{2E_{b1}}\right)^{-1/2}\right) \tag{2-1}$$

where E_{b1} is the energy per bit received from the desired user, E_{bk} is the energy per bit received from undesired user *k*, N_0 is the power spectral density of the channel Additive White Gaussian Noise (AWGN), *N* is the processing gain (number of chips per bit) and *Q*(*x*) is the complementary error function given by

$$Q(x) = \frac{1}{2}erfc\left(\frac{x}{\sqrt{2}}\right) = \frac{1}{\sqrt{2\pi}}\int_{x}^{\infty} e^{-u^2/2}du \tag{2-2}$$

The variable σ in equation 2-1 depends on the level of PN code synchronisation in the system. Pursley [1977] and Morrow & Lehnert [1989] compute the levels of MAI based on various chip and phase offsets, and arrive at the following set of observations: $\sigma = 1$ for chip and phase aligned systems; $\sigma = 1.5$ for phase aligned systems with random chip delays; $\sigma = 2$ for chip aligned systems with random phases; and $\sigma = 3$ for systems with random chip and phase delays. We refer the reader to [Pursley, 1977] and [Morrow & Lehnert, 1989] for a discussion on the accuracy and derivation of these results.

In star-networks (cellular systems) which use power control, the ratios E_{bk}/E_{b1} are all reduced to near unity (if perfect power control is achieved) since all signals arrive with near equal power at the central receiver. Equation 2-1 then becomes

$$P_{BE}^{SGA}(K) = Q\left(\frac{K-1}{\sigma N} + \frac{N_0}{2E_{b1}}\right)^{-1/2} = Q\left(\frac{2E_{b1}}{I_0 + N_0}\right)^{1/2} \tag{2-3}$$

where

$$I_0 = \frac{2E_{b1}(K-1)}{\sigma N} \tag{2-4}$$

is the power spectral density of the MAI due to the *K*-1 other users.

If we define $\mathrm{SIR} = 2E_{bl}/(I_0 + N_0)$ as representing the total signal-to-interference ratio (i.e. signal-to-AWGN plus signal-to-MAI) experienced by a user, then the maximum number of users allowed in the channel is

$$K_{\max} = 1 + \frac{\sigma N}{\mathrm{SIR}_{\min}} - \frac{\sigma N.N_0}{2E_{bl}} \qquad \textbf{(2-5)}$$

where $K_{\max}$ is the maximum number of users allowed before the SIR drops below some minimum acceptable value defined as $\mathrm{SIR}_{\min}$. $K_{\max}$ is often referred to as the *multiple access threshold*, *multiple access soft capacity* [Fapojuwo, 1993], [Gilhousen *et al*, 1991] or *multiple access capability* [Geraniotis, Soroushnejad & Yang, 1995] of a DS-CDMA channel. The SGA provides reasonable accuracy when the number of users K is large. At low values of K however, the SGA provides very optimistic bit error probabilities [Morrow & Lehnert, 1989]. Ignoring the AWGN, equation 2-3 becomes

$$P_{BE}^{SGA}(K) = Q\left(\sqrt{\frac{\sigma N}{K-1}}\right) \qquad \textbf{(2-6)}$$

In addition to the SGA, several more accurate approximations have been developed. Examples are: the Improved Gaussian Approximation (IGA) [Morrow & Lehnert, 1989]; a class of highly accurate models that are based on a Fourier series expansion for the error function $Q(x)$ [Holtzman, 1992], [Sunay & McLane, 1996]; and several other accurate computations of the expected upper and lower bounds of the BER (bit error rate), and an accurate approximation of the BER itself [Pursley, Sarwate & Stark, 1982], [Geraniotis & Pursley, 1982]. Generally, it is found that the computational complexity of these alternative approximations, although more accurate over a large range of K, tends to increase with increasing K. Despite the availability of these more accurate expressions, the SGA is still very popular and finds common application, especially when all that is required is a simple (course) approximation of the BER (for example, when studying the performance of a MAC protocol where the focus is more on the performance of the protocol rather than the physical channel conditions).

2.3 Fixed Assignment MAC Protocols

The *fixed assignment*, or *contentionless* MAC protocols are based on a fixed channel assignment process and are designed to share a fixed bandwidth (B) between a *fixed* population of terminals (K). Traditionally, there have been three main fixed assignment MAC methods: Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), and Code Division Multiple Access (CDMA).

- In FDMA, bandwidth B is divided into K smaller, disjoint sub-bands. Each user has exclusive (uncontested) use of only one of these sub-bands, regardless of their need to transmit anything. The adjacent, unwanted, sub-bands are removed (separated) through spectral filtering at the user's transceiver.

- In TDMA, time is split into consecutive frames and each frame is subdivided into K non-overlapping time slots. Users access the slots in a round robin fashion, where each slot is dedicated to a single user only. If a user has nothing to transmit, their time slot is wasted. Thus each user transmits, using the entire bandwidth B, for one K^{th} of the time.

- In CDMA, it is generally assumed that the available bandwidth B refers to the bandwidth used after the spreading process has been applied. In fixed assignment CDMA, there are $K \leq K_{\max}$ unique PN codes, one assigned on a permanent basis to each of the K users in the population ($K_{\max}$ is the multiple access threshold). All transmissions are done in the same frequency range and can occur simultaneously.

The efficiency of fixed assignment schemes is a function of the percentage channel usage of each user in the system. That is, they are most efficient during periods of high load when each user is transmitting for a large percentage of time. In FDMA, this means that a user would transmit continuously on their allocated frequency, in TDMA a user would always be transmitting in their allocated slot within a frame and in CDMA, a user would always be transmitting on their allocated PN code. At low loads however, these schemes are very inefficient since unused slots, frequencies or PN codes are not available for other users and are effectively wasted. In CDMA, a small number of simultaneous transmissions do not make efficient use of the processing gain obtained through spreading.

Fixed assignment schemes can be implemented in a wireless system, but the user capacity is always limited to K and these schemes are not flexible with regards to dynamic network populations. In order to add a new user to the system, either the bandwidth available (B) must increase, or the bandwidth allocated to each user must decrease to accommodate the new user. In CDMA, the addition of a new user is easily implemented by assigning a new, unique PN code to that user. This is done at the expense of the signal quality of the other users in the network, since the addition of another user results in a slight increase in the MAI. Alternatively, one can increase the system processing gain (this requires an increase in the available total bandwidth, B), hence decreasing the level of MAI (as can be seen in equation 2-4).

2.4 Random Access MAC Protocols

The advent of the computer and, in particular, the information age brought about a new form of traffic, i.e. data. Circuit switched techniques employing FDMA and TDMA, although very efficient for synchronous, "long term" traffic forms such as voice, were found to be very inefficient when applied to computer networks with large populations, bursty data traffic and chaotic arrival processes. Random access protocols are based on channel-seizure techniques, where a single user attempts to temporarily transmit using *all* of the available channel resources. Users contend for the channel in an uncoordinated, or *random*, manner. This randomness and uncoordination results in channel conflict, as multi-user signals are no longer orthogonal in frequency or time. When two or more users decide to transmit at the same time, the signal transmissions are said to "*collide*" in the time and frequency domain - a process of mutual destructive interference which prohibits the successful separation of the individual signals. Random access MAC protocols attempt to achieve channel sharing by employing signal separation in the *time domain*, where the basic assumption is that the random nature in which terminals access the channel occurs in such a manner that the probability of two or more terminals attempting to simultaneously transmit is small.

The success of a message depends, not only on the quality of the transmission medium, but also on the probability of being able to transmit the entire message over the channel without conflict from other users. Channel collisions are the primary degrading factor with regards to the performance of random access protocols. The following sub-sections briefly review the main concepts of narrowband random access protocols.

2.4.1 The ALOHA protocol

The first random access protocol, aptly named ALOHA, was developed by a group of radio researchers from the University of Hawaii [Abramson, 1970]. A single channel is shared by a population of terminals in a totally asynchronous and uncoordinated fashion. In pure, asynchronous ALOHA, a ready user with a packet to transmit does so immediately, using *all* of the channel bandwidth available. All messages or packets generated in the network are assumed to be of equal length and are transmitted in a fixed time duration t_d. In slotted-ALOHA (S-ALOHA) time is divided into fixed length slots of duration t_s which are equal to the size of one packet, i.e. $t_s = t_d$. Some form of channel access co-ordination occurs in that users are now synchronised such that packet transmissions can only commence at the start of a slot, as in TDMA.

The performance of ALOHA and S-ALOHA have been studied exhaustively by many authors. The classical approach (commonly referred to as Abramson's model) is to assume that the point process representing the packet arrival times is Poisson in nature [Abramson, 1970]. This assumption allows for easy modelling of the traffic arrival statistics and, in particular, the use of equation 2-7 to determine the probability of having m users attempting to transmit during a time period τ (seconds), given that the mean message arrival rate from all users in the network is λ (packets per second)

$$P(m,\tau) = \frac{e^{-\lambda\tau}(\lambda\tau)^m}{m!} \qquad \textbf{(2-7)}$$

In computing the performance of such a protocol, it is standard practice to compute various performance probabilities and expressions of interest for a single "reference" message in the channel and, under the assumption that all messages are subjected to the same conditions, use these probabilities and expressions to obtain the overall measurements of the system performance. For example, if the expected probability of successful transmission of an arriving message is fifty percent and there are, on average ten messages arriving per unit time, then we expect the number of successful messages leaving the channel per unit time to be five. For a Poisson traffic model, if the total average packet arrival rate is G packets per unit packet duration (i.e. $G=\lambda t_d$), then the throughput S in packets per unit time t_d is given by [Taub & Schilling, 1986]

$$S = \begin{cases} Ge^{-2G} & ALOHA \\ Ge^{-G} & S-ALOHA \end{cases} \tag{2-8}$$

with maximum throughputs of $S_{\max} = 0.5e^{-1} \approx 0.184$ when $G = 0.5$ for ALOHA and $S_{\max} = e^{-1} \approx 0.368$ when $G = 1.0$ for S-ALOHA. It should be noted that the throughput gained by using slots is at the expense of increased complexity and cost due to the need for network synchronisation. Packets involved in collisions are not received correctly and need to be retransmitted. Transmitters are usually notified that transmission was unsuccessful via some form of ARQ (Automatic Repeat Request) system or acknowledgement receipt time-out mechanism. These unsuccessful packets are rescheduled for retransmission according to a *Collision Resolution Algorithm* (CRA).

2.4.2 Collision Resolution Algorithms (CRA) and System Stability Issues

The aim of a CRA is to spread the starting instances of retransmitted packets over time to avoid the certainty of them conflicting with each other again. Immediate retransmission of collided packets is not an effective solution as it inevitably results in the same packets continually colliding. Random waiting times ensure that packet retransmission starting times are different. This waiting time (defined as W) is commonly referred to as the *backoff* time, and nodes that are waiting to retransmit are commonly referred to as being *backlogged*. In asynchronous ALOHA, the CRA is usually based on a process known as *non-persistence*. In non-persistence, the backoff time is explicitly chosen in a random fashion according to some distribution function, the most common being the uniform and truncated negative exponential distributions [Gallager, 1985]. In slotted systems such as S-ALOHA, the CRA is usually based on a *p-persistence* process. Instead of explicitly choosing a backoff time period, the terminal attempts to retransmit in the next slot with probability *p*, or delays retransmission until the following slot with probability 1-*p*. This process is repeated (with *persistence*) until the packet is transmitted. The parameter *p* should be chosen small enough to ensure that the number of retransmitted packets that a retransmitted reference packet finds across the channel from one collision to the next are virtually independent [Resheff & Rubin, 1990]. For a constant value of *p*, this results in a backoff time *W* (measured in slots) which is geometrically distributed with mean $\overline{W} = 1/p$.

In a random access protocol such as ALOHA, the probability of packet success is $P_{PS} = S/G$. The average number of transmissions required until success is then given by $1/P_{PS}$. If the mean backoff time ($\overline{W}$) is known, then the expected packet delay $\overline{D}$ for ALOHA can be shown to be [Taub & Schilling, 1986]

$$\overline{D} = \begin{cases} t_p + \dfrac{(\overline{W} + A + t_d)(1 - P_{PS})}{P_{PS}} & ALOHA \\ \dfrac{1}{2} + t_p + \dfrac{(\overline{W} + A + t_d)(1 - P_{PS})}{P_{PS}} & S-ALOHA \end{cases} \tag{2-9}$$

where t_d is the packet length, A is the delay while waiting for a message acknowledgement and t_p is the signal propagation delay.

Especially evident in ALOHA and S-ALOHA is the sudden decrease in network performance when loaded beyond a certain point (refer to figure 2.1). This turning point is due to the increased rate of collisions and is also closely attributed to the stability of the network. In the low load, linear region of the load-throughput curve, most message transmissions are successful. As the load increases, more and more users become backlogged due to blocking or collisions. If the collision resolution algorithm is not chosen correctly, the network becomes unstable in that backlogged users attempting to retransmit collide with other idle users who are transmitting for the first time. This effectively adds more users to the backlog population in a "snowball effect". In the worst case scenario, the entire network population is backlogged and the channel becomes *saturated* with retransmissions, very few of which succeed due to the high channel access rate. The collision resolution algorithm is thus a decisive component in the overall access protocol, as retransmissions have to be catered for in the arrival process.

The problem with fixed CRA schemes is that there is no flexibility with regards to the retransmission delay, since $\overline{W}$ (the average backoff time) is fixed for all nodes at all load values. At low channel loads, the collision rate is small and the number of terminals involved in each collision is likely to be low (perhaps 2 or 3 at the most). It is intuitive then that the CRA need not spread these 2 or 3 users over as large a time span as it would need to do for say 10 or 20 users involved in a collision at a much higher load.

Many researchers have studied the stability problems inherent in random access protocols and many collision resolution algorithms associated with improved system stability have been proposed ([Jenq, 1980], [Carleial & Hellman, 1975], [Thomopoulos, 1988] and the references cited therein). Before presenting stabilising CRA's, we first review some terms used to analyse the stability of a system. In [Jenq, 1980] and [Carleial & Hellman, 1975] for example, the concepts of population drift and system equilibrium points are explored in the context of the stability of ALOHA systems. These concepts provide network designers with a measure of the rate of change of the backlog population size, and the rate at which users enter and leave the backlogged state.

There are three main schemes that improve system stability through dynamic adaptation of the CRA backoff time: *load* dependence, *retries* dependence and *backlog* dependence. These three factors are closely related however. At high loads one would expect both the number of retransmissions (retries) per packet and the fraction of terminals in the backlog state to increase, although perhaps not in a linear manner. The aim of a stabilising CRA is to regulate the channel load generated by backlogged terminals. When the load becomes too high, the collision rate increases and the system automatically compensates by decreasing the channel access rate from retransmitted packets. The three schemes are discussed below:

1. **Load dependent:** Each terminal measures the network load (channel access rate G) and adjusts the average backoff time accordingly. If the load increases, the distribution of W is expanded to obtain a larger $\overline{W}$. In p-persistent schemes, high values of p can be used at low channel access rates. High values of p imply a small delay between retransmissions and hence a lower overall average message delay. At high loads, p must decrease to ensure that the system remains stable. Low values of p are unfortunately associated with larger average message delays however.

2. **Backlog dependent**: Terminals estimate the number of users in the backlog state in the current slot by using feedback (*idle*, *success,* or *collision*) from the previous slot (some methods for doing this are described in [Thomopoulos, 1988]). The estimated backlog is updated on a slot by slot basis and the backoff delay adjusted accordingly. At low backlog values, short backoff durations are more efficient. As the backlog population increases, $\overline{W}$ should increase accordingly.

3. **Retries dependent:** Each time a terminal fails to transmit a packet, $\overline{W}$ is increased according to a suitable algorithm. Packets that find themselves continually colliding with other packets, whether they be new or retransmitted packets, extend the length of their backoff each time in an attempt to spread themselves over a larger period of time. An example of a retries dependent scheme is the truncated binary exponential backoff scheme used in the popular ETHERNET p-persistent CSMA/CD protocol (see [Chen & Li, 1989] for a detailed description of this scheme). This method has proved to achieve a high level of system stability, and is implemented in many LAN's throughout the world.

Another common set of sophisticated collision resolution schemes is the set of *splitting algorithms*. The simplest forms of splitting algorithms are the *tree algorithms* developed by Capetanakis [1979]. In fact, it has been shown that Capetanakis' tree algorithm has an improved maximum throughput of 0.43 and is stable for all input rates less than 0.43 (as opposed to the maximum throughput of 0.37 for S-ALOHA).This scheme involves subdividing the collided users into two groups (sub-branches) and resolving the collisions within each branch. Successive collisions create new branches, and the process repeats until all collisions have been resolved and no more can occur. A detailed discussion is deemed to be beyond the scope of this review, but we point the reader to [Capetanakis, 1979] for further treatment of tree algorithms.

2.4.3 Carrier Sense Multiple Access (CSMA)

In ALOHA and S-ALOHA, no co-ordination between users exists and packet transmissions are performed at the user's convenience. At high loads, collisions become the primary performance degrading factor. In an attempt to minimise destructive collisions, the Carrier Sense Multiple Access (CSMA) scheme was derived to add some degree of user co-ordination. Two versions of CSMA exist, based on whether the system in slotted or unslotted:

1. **Non-persistent CSMA:** In *non-persistent* CSMA, time is unslotted. A user wishing to transmit first senses the channel for a transmission carrier. If the channel is sensed to be idle (not in use), the user seizes the channel for their exclusive use and starts transmitting immediately. If the channel is sensed to be busy, the user enters a backlogged state (is blocked) and reschedules transmission using a *non-persistence* scheme. In this way collisions are avoided as, once a user has seized the channel, no further conflict can occur. The randomness property of the CRA avoids the event of all waiting nodes immediately transmitting when the channel becomes idle.

2. ***p*-persistent CSMA:** In *p-persistent* CSMA, time is divided into miniature slots of size $t_p/2$, where $t_p/2$ is the maximum one-way propagation delay that exists between the furthermost two terminals in the network [Kleinrock, 1975b]. Packets can only begin transmission at the start of a slot, and are transmitted over many contiguous slots. Upon detecting an idle channel, a ready terminal attempts to

transmit with probability *p*, or delays transmission by one minislot ($t_p/2$) with probability 1-*p*. If the channel is sensed to be busy, the terminal waits until the channel becomes idle, and then continues with the same *p*-persistence process.

Ideally, once the channel has been successfully seized, no collisions can occur since other terminals will sense that the channel is in use. In real networks, the end-to-end signal propagation delays and system processing delays result in small periods of conflict. During these periods, the delays are such that a user (A) may sense the channel to be idle and begin transmission when in fact another user (B) had seized the channel a short time earlier and B's signal has not yet reached A due to the propagation delay.

Arguably, the definitive analysis of CSMA was done by Kleinrock & Tobagi [1975]. Extensive descriptions and analyses of variants of the basic CSMA technique such as slotted-CSMA, non-persistent CSMA, 1-persistent CSMA and *p*-persistent CSMA are given. The interested reader is encouraged to refer to [Kleinrock & Tobagi, 1975] and [Kleinrock, 1975b] for a detailed treatment of CSMA. The throughput and delay of CSMA are non-linear functions of the propagation delay and are somewhat more involved to compute than for ALOHA. Nevertheless, Abramson's assumptions of fixed packet lengths and a Poisson arrival process allow for simple closed form solutions to be found. We present the final results for throughput and delay of *non-persistent* CSMA (and refer the interested reader to [Kleinrock & Tobagi, 1975] and [Kleinrock, 1975b] for a complete analysis).

$$S = \frac{Ge^{-Gd}}{G(1+2d)+e^{-Gd}} \quad \textbf{(2-10)}$$

and

$$\overline{D} = t_p + (A + \overline{W} + t_d)\left(\frac{G}{S} - 1\right) \quad \textbf{(2-11)}$$

where $d = (t_p/2)/t_d$ is the one-way normalised propagation delay and the other delays are as explained for ALOHA. Kleinrock [1975b] also provides results, more complicated than the scope of this brief review, for the average throughput and delay for *p-persistent* CSMA. Figure 2.1 compares the throughput performance of ALOHA, S-ALOHA and *non-persistent* CSMA for various values of *d*.

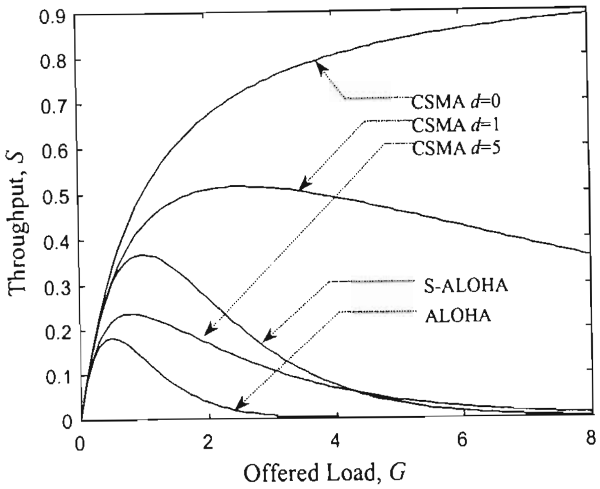


Figure 2.1: Throughput vs. Offered Load for ALOHA, S-ALOHA, and non-persistent CSMA for different values of delay (*d*=0,1,5)

As the offered load *G* approaches and starts to exceed 0.5 (ALOHA) or 1.0 (S-ALOHA), collisions start to degrade the system performance and the throughput decreases. CSMA is observed to be far more efficient than the ALOHA systems, but only when the signal propagation delay is small in comparison to the system packet length (i.e. low values of *d*). It is obvious from figure 2.1 that the maximum obtainable throughput for any narrowband random access protocol is unity, even for channel loads far exceeding unity.

2.4.4 CSMA with Collision Detection (CSMA/CD)

The CSMA protocol was further extended by Tobagi & Hunt [1980], by implementing a Collision Detection (CD) scheme. Since CSMA cannot totally avoid collisions because of propagation delays, a collision detection scheme enables users involved in a collision to abort transmission immediately and avoid continued wasteful transmission of a packet that has collided and is hence corrupt. In CSMA/CD a user who has seized the channel monitors it to see if what they are transmitting is what is actually on the channel. If not, a collision is assumed and the user aborts transmission. It is generally assumed that all users involved in the collision are aware of its occurrence. In some cases, a collision declaration signal is broadcast to ensure that the duration of the collision is long enough to allow all nodes in the network to become aware of its occurrence [Chen & Li, 1989]. The CSMA/CD protocol has been found to be extremely efficient, and has been employed worldwide in LAN's under the ETHERNET brand-name.

2.4.5 Inhibit Sense Multiple Access (ISMA)

In packet radio networks with vast topographical variations, the performance of CSMA is greatly reduced by the "hidden terminals" problem. Geographical obstructions can cause some nodes to be unaware of the transmitting states of certain other nodes in the network. Two transmitting nodes, A and B, may be within range and line-of-sight of a receiving node C, or a base station, but be out of range of each other or have no direct line of sight of each other. In this case, neither A nor B will be able to detect whether the other is using the channel and may attempt to simultaneously transmit to C, resulting in conflict at C's receiver. ISMA (Inhibit Sense Multiple Access) has been proposed and studied as a centralised form of CSMA in packet radio networks [Zdunek *et al*, 1989], [Prasad, 1991]. ISMA has been studied under several aliases, the more common being ICMA (Idle-signal Casting Multiple Access) [Lee & Un, 1996] and BTMA (Busy-Tone Multiple Access). In ISMA, a centralised base station, which is within range and line-of-sight of all nodes in the network, senses the shared inbound communications channel and broadcasts the channel information via a separate out of band signalling channel. In most cases, a simple narrowband on-off tone is used as a collision declaration signal.

ISMA protocols have been studied by Zdunek *et al* [1989], Lee & Un [1996] and Prasad [1991]. Both Lee & Un and Prasad show that the inhibit delay fraction is a decisive parameter in the design of the network. This inhibit delay is measured as a fraction of the average packet duration and represents the maximum round trip propagation delay required to detect the channel state and notify the network thereof and, as in CSMA, degrades the system performance when too large.

2.4.6 Performance of random access protocols in packet radio networks

Wireless networks which use ALOHA, S-ALOHA, CSMA and ISMA have received considerable attention [Abramson, 1970], [Prasad, 1991], [Habbab et al, 1989], [Jeong & Jeon, 1995], [Lutz, 1992], [Stavrakakis & Kazakos, 1989], [Kleinrock & Tobagi, 1975], [Tobagi & Hunt, 1980], [Raychaudhuri, 1984]. Some of these studies have been done to determine the effects of the radio channel on the performance of these protocols, with particular interest in the near-far effect and how capture can improve the system capacity [Prasad, 1991], [Habbab et al, 1989]. Capture provides a non-zero probability of success in the event of channel collisions and hence increases the system throughput, although expressions for throughput are now much more difficult to obtain due to the complex and time varying nature of the channel, as well as the spatial distribution and signal power distribution of the population. Prasad [1991] showed that the combined effect of Rayleigh fading, log-normal shadowing and capture with spatial diversity can improve the capacity of S-ALOHA by as much as 80%. Habbab *et al* [1989] produced a similar result when analysing the effects of capture over slow and fast fading channels with FEC (Forward Error Correction) coding schemes and spatial diversity. These packet contention protocols thus perform better in radio networks than in wired common-bus or common-cable type systems, due to capture. As stated before, ISMA is a better alternative to CSMA in star-networks with "hidden terminals".

In contrast to fixed assignment protocols (TDMA and FDMA), random access protocols are most efficient at low network load values where user traffic is bursty in nature. From figure 2.1, we see that the ALOHA schemes both offer very limited capacity. CSMA is quite clearly the more efficient protocol, but only in cases where the worst case propagation delay is small. It is intuitive then that CSMA and ISMA are not practical in satellite systems where the propagation delay is in the order of tenths of a second (e.g. 140ms for geostationary satellites at orbits of 36000km above the earth). Because of this, satellite systems traditionally used TDMA and FDMA techniques, although reservation (demand assignment) and CDMA systems are currently being employed due to their improved performance and flexibility.

2.5 Spread-Spectrum (CDMA) Random Access MAC Protocols

As mentioned in section 2.3, fixed assignment CDMA is inefficient for variable size populations with bursty traffic and low percentage channel usage per user. This arises from the fact that at low channel loads, inefficient use is made of the large processing gain that is implemented in order to provide a satisfactory bit error rate at high loads. As in the narrowband case, random access to the channel offers an alternative for networks which have populations greatly in excess of the network capacity and bursty traffic arrival processes.

The main limiting factor of narrowband random access protocols is that, at any one time, the channel guarantees errorless support for one single user only (assuming an ideal channel). As soon as two or more users attempt to transmit, the probability of packet success becomes zero (assuming no capture). Spread-spectrum systems offer the advantage of allowing more than one user to access the channel at any one time, but with some error rate. The cost for this "luxurious", almost contentionless, channel is a vastly increased bandwidth. By "almost contentionless", we refer to the fact that each user does interfere with each other user in some form, although the interference added by each user is minimal, and is insufficient to cause message errors by itself. When summed with the interference from all other users in the network however, the result may be such that packet errors can occur, depending on how many users are involved in the summation. The probability of success of a packet is thus almost entirely dependent on *the number of users accessing the channel at that time*. In this section, we consider some schemes that combine the concepts of CDMA with the concepts of narrowband random access protocols.

2.5.1 Spread-spectrum ALOHA (SS/ALOHA)

In spread-spectrum ALOHA (often abbreviated to SS/ALOHA), users follow the same access algorithms and CRA's as in ALOHA. That is, when a terminal has a packet ready to transmit, it does so immediately. Upon receipt of a negative acknowledgement, it reschedules transmission according to a CRA. The difference is that "collisions" in spread-spectrum systems are not associated with a 100 percent packet error probability, but rather an error probability which is dependent on the number of users. Not all users are equally affected by the MAI - some of the transmissions may be successful, while others may not. The packet error probability, which is almost zero when few users are transmitting, approaches one as the number of transmitting users tends towards infinity.

As in narrowband ALOHA, much research has been done in the field of SS/ALOHA protocols (examples are: [Raychaudhuri, 1981], [Storey & Tobagi, 1989], [Yin & Li, 1990], [Joseph & Raychaudhuri, 1993], [van Nee *et al*, 1995], [Toshimitsu *et al*, 1994], [Resheff & Rubin, 1990], [Abdelmonem & Saadawi, 1989], [Polydoros & Silvester, 1987]). Authors have studied slotted and unslotted systems, centralised systems and distributed systems, systems with fixed length packets and variable length packets - all possible evolutions of the traditional narrowband ALOHA scheme. The effects of channel fading, capture and the use of FEC coding schemes have also been exhaustively investigated.

Raychaudhuri [1981] presents a comprehensive analysis of an ideal Slotted-SS/ALOHA (S-SS/ALOHA) system with fixed length packets (the same size as a slot). He considers both an infinite population model with a Poisson arrival process and a finite population model with a Markov arrival process. For the finite population case, Raychaudhuri applies the concept of population drift (following the logical structure of [Jenq, 1980]) to a p-persistent CRA and shows the system to be unstable for fixed values of p which are too large. The results of his work reveal that SS/ALOHA is in fact no different to narrowband ALOHA systems with regards to system stability and channel efficiency. The throughput-delay curves so characteristic of ALOHA systems were obtained for SS/ALOHA as well, the only difference being that SS/ALOHA was effectively a scaled up version of the narrowband case with respect to the average throughput per slot. Narrowband ALOHA can effectively be thought of as a special degenerate case of SS/ALOHA. Although SS/ALOHA improves upon the narrowband throughput of 0.184 (ALOHA) or 0.368 (S-ALOHA), it does so at the expense of the increased bandwidth due to spreading. If the bandwidth expansion factor N (or spreading gain) is divided into the throughput S, then the normalised efficiency (β) of SS/ALOHA is not as good as it originally appears [Raychaudhuri, 1981]. β represents the throughput per "unit" of bandwidth:

$$\beta = S / N \qquad \textbf{(2-12)}$$

The analysis of a spread-spectrum system is not unlike the narrowband case. For this review we present a simplified analysis (for comparative purposes) of a slotted-SS/ALOHA system. We ignore channel effects such as fading, capture and AWGN, and assume the chip and phase asynchronous SGA model for MAI (σ=3 in equation 2-6). The population model is assumed to be infinite and the combined arrival process is Poisson with mean λ packets per second. The system is time slotted with slot size t_s. Packets are of fixed length L bits, and are transmitted in exactly one slot. Since the number of users is constant during a slot, it is assumed that the level of MAI during a slot is also constant. If no FEC scheme is employed, then the probability of success of a reference packet during k other spread-spectrum packet transmissions, $P_{PS}(k)$, is conditioned on the fact that all L bits in the packet are successful:

$$P_{PS}(k) = [1 - P_{BE}(k)]^L \qquad \textbf{(2-13)}$$

where $P_{BE}(k)$ is a suitable bit error probability (e.g. the SGA approximation provided in equation 2-6). The overall probability of packet success must include all possible values of k. In an infinite population, the probability exists, albeit extremely small, of k being infinite. If the load offered to the slotted-SS/ALOHA system is $G = \lambda t_s$, then the throughput S is given by

$$\begin{aligned} S &= \sum_{k=1}^{\infty} k.P(m = k, \tau = t_s).P_{PS}(k) \\ &= Ge^{-G} \sum_{k=0}^{\infty} \frac{G^k}{k!}.P_{PS}(k+1) \end{aligned} \qquad \textbf{(2-14)}$$

where $P(m = k, \tau = t_s)$ is the Poisson arrival process given by equation 2-7. The throughput S and normalised throughput β for S-SS/ALOHA are plotted in figure 2.2 for various processing gains, along with the throughput of narrowband S-ALOHA. As the processing gain N increases, the throughput improves because $P_{BE}(k)$ decreases for all values of k. The normalised load/throughput curves in figure 2.2b reveal however, that S-SS/ALOHA is more spectrally inefficient than narrowband S-ALOHA.

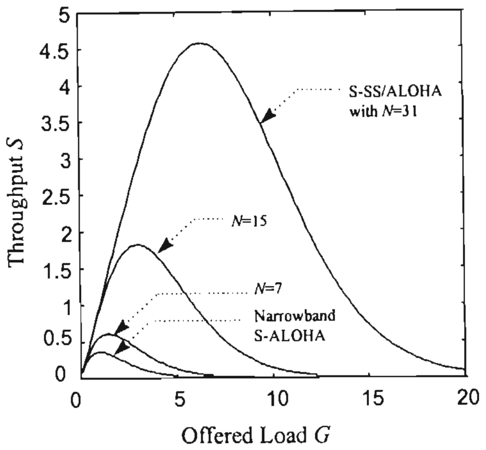


Figure 2.2a: Throughput vs. Offered Load for narrowband S-ALOHA and S-SS/ALOHA for various processing gains (N)

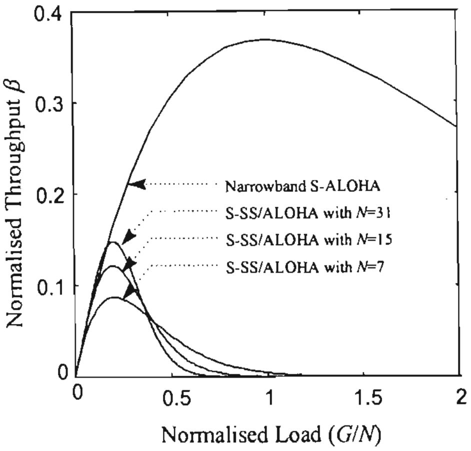


Figure 2.2b: Normalised Throughput vs. Normalised Offered Load for S-ALOHA and S-SS/ALOHA with different values of N

Similar observations were made for unslotted SS/ALOHA by Storey & Tobagi [1989] and Yin & Li [1990]. Storey and Tobagi analysed the effects of having variable packet lengths with exponential distribution in an infinite population model. They show that SS/ALOHA is far more supportive of longer packets than narrowband ALOHA. This is somewhat intuitive since narrowband ALOHA only supports one packet at a time, and longer packets have a longer period of vulnerability. Joseph & Raychaudhuri [1993] complement the above work by analysing the effect of variable length packets in a finite population model. They analyse the packet arrival process from idle and backlogged users and confirm that longer packets have a lower success probability than short packets. [van Nee *et al*, 1995] provide a good comparison between narrowband S-ALOHA and slotted-SS/ALOHA in various fading conditions and show that both systems provide similar responses.

Most authors agree that an accurate performance measure for SS/ALOHA is far more difficult to obtain than for narrowband ALOHA. Factors such as fading [van Nee *et al*, 1995], signal capture, receiver types (e.g. RAKE receivers combat the negative effects of multipath distortion [Cheun, 1997]), FEC coding schemes, different processing gains and even different spreading techniques can greatly affect the performance of a CDMA system. From a relative performance viewpoint however, most researchers show that SS/ALOHA shares many common conceptual performance characteristics with narrowband ALOHA.

2.5.2 Channel Load Sensing (CLS)

In narrowband CSMA, channel sensing is used to avoid destructive collisions. The result obtained from the narrowband sensing mechanism can have two states (idle or busy) and is used to allow access or block access to ready packets. By observing the throughput gained from this technique, it has been of interest to see if a similar improvement can be made to spread-spectrum networks. Since blocking of access when one user is transmitting defeats the object of spread-spectrum, a measurement of the channel *load* (number of transmitting users) is of much greater interest than whether or not the channel is being used. Based on this load value, steps can be made to regulate the channel access to avoid cases of excessively large MAI that occur when too many users are transmitting.

Unfortunately, an accurate physical measurement of the channel load is more complicated to obtain than in narrowband CSMA. Two main techniques exist for measuring the number of transmissions, one based on power level measurements and one based on multiple receivers to demodulate all signals in the channel.

1. **Estimation of the channel load using power levels:** Power level measurement schemes involve measuring the total power received from all stations and dividing this value by the expected power received from each user. These power level measurements are only valid in centralised spread-spectrum systems that implement fast and near-perfect power control, since each user is received at the base station with near equal power. In such systems, power load level estimations are generally not used to determine the channel load however, since the base station is usually capable of obtaining an estimate of the number of transmitting users purely from the fact that takes part in all communications and will know how many users it is

dealing with. In distributed systems, where each mobile is responsible for measuring the channel load for itself, non-equal power levels from different terminals (due to the different distances between terminals) cause this form of channel load estimation to be extremely inaccurate. In addition, each mobile will inevitably arrive at a different estimate of the channel load.

2. **Measurement of the channel load through multiple receivers:** A more accurate channel load measurement can be obtained by using multiple receivers at each terminal to despread all signals in the channel. Each receiver is tuned to one of the PN codes in the code population. In distributed, fully connected schemes where power control is impossible to implement, this measurement technique provides good results. The main disadvantage is that each terminal needs to have at least as many receivers as there are terminals in the population. In centralised systems, all nodes communicate via a base station so the base station generally implements a multiple receivers structure anyway.

Intuitively, channel load sensing (CLS) techniques are only useful in asynchronous systems or slotted systems in which packets are transmitted over multiple slots. In slotted-SS/ALOHA systems where the packet lengths are equal to the slot size, the channel load is completely independent from slot to slot. A packet which enters the channel in slot t, will not be in the system in slot t+1. It makes little sense to measure the channel load in slot t and base the admission decision in slot t+1 on this measurement. A system in which there can be partial packet overlaps can implement CLS however, since the channel load is no longer independent from one slot to the next [Toshimitsu *et al*, 1994]. Once a value for the channel load has been obtained, it can be used to control channel usage in two ways:

1. **Channel load regulation through blocking**: This form of channel load sensing is analogous to the blocking mechanism in narrowband CSMA. It attempts to avoid excessive MAI by regulating the channel load around a safe operating level. If the channel load is less than the multiple access threshold $K_{\max}$ (defined in section 2.2.2) then access to the channel is allowed, otherwise access is denied until the load

drops below $K_{\max}$. Packets that are denied access follow a rescheduling algorithm in the same manner as narrowband CSMA (i.e. *p*-persistence, non-persistence etc.). Narrowband CSMA is a degenerate case of spread-spectrum CLS, with $K_{\max} = 1$.

2. **"Collision"/Overload Detection:** As in narrowband CSMA/CD, collisions or overload cases in the channel may be sensed and action taken to abort packets. The assumption can be made that once the channel load exceeds $K_{\max}$, all packets involved have a high probability of being corrupt. Aborting these packets may then result in a higher throughput, as observed in the narrowband equivalent. The term "overload detection" is perhaps more suitable than "collision detection" when referring to spread-spectrum systems.

The improvement in system performance gained through the use of channel load sensing has been studied for both the blocking scheme [Toshimitsu *et al*, 1994], [Abdelmonem & Saadawi, 1989] and the overload detection scheme [Resheff & Rubin, 1990], [Lam & O'Farrell, 1992], although no work could be found which implements both schemes together. As mentioned before, CLS is only useful in asynchronous networks or slotted networks with large packets (larger than a slot). Abdelmonem and Saadawi consider an asynchronous system with variable length packets while Toshimitsu *et al* consider a slotted system with fixed length packets transmitted over multiple slots. Both studies show that throughput enhancement occurs if packets are blocked when the load exceeds a channel threshold.

The effect of CLS in overload detection schemes has also been studied. Resheff & Rubin [1990] provide an exact statistical analysis of a slotted system with variable length packets in which collision detection is employed. The model uses slotted-SS/ALOHA with an infinite population and Poisson arrival process. All messages are aborted when the channel load is measured to exceed the overload threshold $K_{\max}$. Lam & O'Farrell [1992] provide similar results obtained from an OPNET simulation of a similar CDMA network with overload detection. Both Resheff & Rubin and Toshimitsu *et al* observe that the propagation delay is an important factor which degrades system performance, as it results in late channel information which may not be representative of the current channel state. An overload case may occur in slot t-t_p, but only be detected by nodes in slot t, where t_p is representative of the signal propagation delay. This delay effect is negligible when the average packet size is much larger than t_p, but becomes increasingly adverse as the mean packet size tends towards t_p.

In the literature reviewed on this subject, little attention was given to the exact process by which channel state information is distributed to the terminals. Most authors assume that each node is capable of obtaining an exact load value, and focus only on the effects of utilising the result.

2.6 Demand Assignment and Reservation MAC Protocols

Random access protocols offer the desirable feature of being able to transmit immediately (no delay) over the full channel bandwidth available. A single user can seize the channel for use when other users do not need it. Fixed assignment protocols offer desirable features such as fixed delay and a guarantee that once transmission is in progress, no collisions can result. Ultimately, it is desirable to be able to transmit as soon as possible without the possibility of colliding with other users, and without allocating wasted channel bandwidth on users who have no need for it. Demand Assignment Multiple Access (DAMA) protocols attempt to combine the advantageous properties of contentionless protocols and random access protocols by employing *channel reservation techniques*. Random access schemes are usually used to contend for channel reservations, while fixed assignment schemes are used to co-ordinate reserved channel bandwidth. In this manner, portions of the available bandwidth are still contended for, but once bandwidth is reserved the user has exclusive use of it until they no longer require it.

In this section, we review the two main classes of channel reservation schemes. Since a multitude of slightly modified protocols have been proposed for each class, we focus almost entirely on the two original schemes that brought about each derivative, namely Reservation-ALOHA and Robert's Reservation scheme.

2.6.1 Reservation-ALOHA (R-ALOHA)

The earliest form of packet reservation is based on improving the efficiency of the S-ALOHA protocol for networks which support variable data packet lengths. This protocol is known as Reservation-ALOHA (R-ALOHA) and was original proposed by Crowther *et al* [1973] and studied for data networks (e.g. [Lam, 1980], [Crisler & Needham, 1995]). In R-ALOHA, time is divided into slots and slots grouped into frames as in a conventional TDMA system, as illustrated in figure 2.3. Each slot in the frame is recognised as being either "*available*" or "*reserved*". Terminals with ready messages to transmit use S-ALOHA to contend for available slots in a frame. When a terminal successfully manages to transmit its first packet in a slot, it reserves that particular slot number for *exclusive* use in all subsequent frames until all packets in the message have been transmitted. Upon completion of the message, the terminal releases the slot for availability to other terminals. If contention occurs when attempting to place a reservation, the terminal attempts in the next unreserved (available) slot with probability p (or other suitable CRA). In this way, collisions are confined to the first packet in the message only. The system can either be centrally controlled by a base station which is responsible for signalling which slots in a frame are available for contention, or distributed. In a distributed system, terminals have to sense for themselves which slots are available for contention.

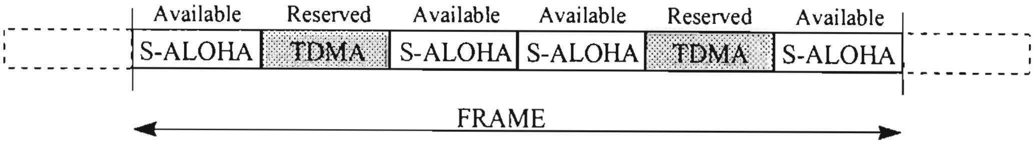


Figure 2.3 : Frame structure for the Reservation-ALOHA protocol

Analysis of such protocols is somewhat more involved than for simple ALOHA protocols and as such we deem it unnecessary to present details of such analyses in this brief review. We point the interested reader to [Lam, 1980] for a comprehensive analysis of R-ALOHA. An obvious necessity for R-ALOHA is messages which require more than one slot for transmission. For variable length packets, Lam [1980] shows that the maximum throughput of R-ALOHA is dependent on the average message length (v). When the average packet length is v = 1 slot, then the maximum throughput of R-

ALOHA is simply that of S-ALOHA (i.e. 0.368). As $v \to \infty$, the maximum throughput approaches the theoretical maximum obtained for TDMA. This is somewhat intuitive since long messages transmitted over many frames reduce the collision rate in the slot they have reserved. An interesting claim made by Crisler & Needham [1995] is that the performance of R-ALOHA is identical to that of slotted non-persistent (CSMA/CD) when the reservation time and propagation delay are equal.

Another common protocol which is based on R-ALOHA is Packet Reservation Multiple Access (PRMA) as proposed and analysed by Goodman *et al* [1989]. Whereas R-ALOHA was initially proposed for data networks, PRMA was proposed to support packetised voice traffic in which only talkspurts are transmitted. We review PRMA in further detail in Chapter 5 as it also includes protocols to integrate voice and data into the same frame structure.

2.6.2 Robert's Reservation Protocol

Roberts [1973] proposed the first of the second main class of DAMA protocols which are based on divided frame techniques. His original protocol, commonly referred to as Robert's Reservation Protocol, forms the basis of many modern DAMA reservation techniques. As with R-ALOHA and PRMA, time is divided into slots and slots are grouped into frames. Each frame is then further partitioned into a reservation subchannel and a data subchannel. The reservation subchannel consists of a single slot in the frame (usually the first) and is subdivided into short reservation minislots. Figure 2.4 illustrates a typical DAMA frame structure employing Robert's scheme.

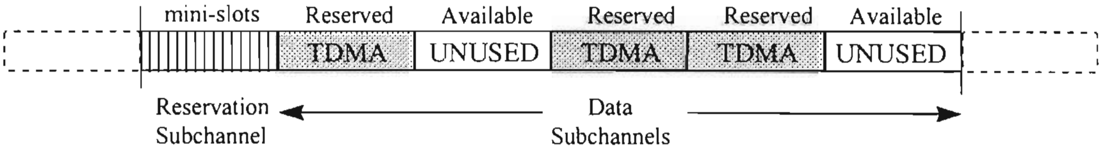


Figure 2.4: Example of a typical frame structure for the class of DAMA protocols based on Robert's Reservation

Using S-ALOHA, each ready station first sends a short reservation request packet (the size of a minislot) over the reservation subchannel to join a global queue which schedules reserved slots on the data subchannel. Once a terminal has successfully reserved a slot in the data subchannel, it has exclusive and uncontested use of that slot in all subsequent frames until message completion. A central controller is responsible for recognising whether slots are reserved or available, as well as resolving requests made in the reservation subchannel. This form of multiple access is more efficient in that channel contention is now limited to the reservation subchannel only and thus throughput losses are not as serious as in S-ALOHA and R-ALOHA. In Robert's Reservation protocols, the data subchannel slots which are available are unused, unlike in R-ALOHA where users contend in available slots using S-ALOHA.

A plenteous amount of channel reservation schemes based on Robert's reservation and R-ALOHA (PRMA) have been proposed. Each variant modifies the original to suit a specific application, the most explicit modifications being to the bandwidth reservation access protocol. Derivatives of Robert's Reservation scheme have considered multiple access protocols other than conventional S-ALOHA for the reservation subchannel. An example is the WIMA (Wireless Integrated Multiple Access) protocol [Wieselthier & Ephremides, 1995] which implements a conventional TDMA system for the reservation subchannel. In WIMA, the population size (M) is assumed to be fixed and there are M TDMA mini-slots in the reservation subchannel. The number of data subchannels should obviously be significantly less than M in order to justify the purpose of implementing channel reservation.

Demand assignment is essentially a TDMA based system that supports more users (with bursty traffic profiles) than there are available slots in the frame. The frame structure and channel transmission mechanisms are essentially TDMA in that a terminal has to wait for its reserved slot. The advantage however is that when a user does not require the slot, it can release it to other users in the system. Reservation protocols also offer the possibility of a single terminal being able to reserve more than one slot in a frame. A shortcoming of the original Robert's Reservation scheme is the fixed time duration of the complete frame, data subchannel and reservation subchannel. Roorda & Leung [1996] point out that when the frame size is fixed, slots in the data subchannel could lie idle during periods of low channel load. This decreases the spectral efficiency and implies an end-to-end message delay which is larger than it could be, since idle slots represent both wasted bandwidth and wasted time. To address this problem, Rooda & Leung [1996] propose an improved model which dynamically adjusts the subframe partition sizes based on the network load state, and shows that optimal partition strategies can be applied to obtain vastly superior performances over fixed frame techniques. Reduction in frame size at low channel reservation rates has several advantages for data traffic. Firstly, it reduces the number of wasted slots in a frame and secondly, it reduces the end-to-end delay because corresponding slots in each frame are closer together in time. Wieselthier & Ephremides [1995] point out however, that fixed sized frames are better suited to satisfying the real-time and fixed delay requirements of synchronous voice traffic.

2.7 Summary

Chapter 2 was a literature review on existing multiple access techniques for modern single service networks. Firstly, we discussed pertinent multiple access issues in spread-spectrum communications; with particular focus on the multiple access capability (capacity) of a CDMA channel and some approximations for the probability of bit error in the presence of multiple access interference. We then discussed the three main classes of traditional MAC protocols: the fixed assignment schemes (TDMA, FDMA and CDMA), the random access schemes (narrowband: ALOHA, S-ALOHA, CSMA, CSMA/CD and ISMA; and spread-spectrum: SS/ALOHA and CLS), and the demand assignment/reservation schemes (R-ALOHA, Robert's Reservation). The performances of these protocols were characterised and compared, and each protocol was reviewed in the context of it's suitability to a wireless packet radio environment.

CHAPTER 3

DUAL-THRESHOLD CDMA (DT/CDMA) MAC PROTOCOL

3.1 Introduction

In this chapter we present our proposed MAC protocol, called "Dual-Threshold CDMA" (abbreviated to DT/CDMA), for the reverse link of a centralised wireless CDMA data-only packet network. Our DT/CDMA protocol uses a data-only MAC protocol presented and analysed by Resheff & Rubin [1990] as a basis, and improves on this scheme by exploiting a phenomenon noted in the study of a MAC protocol proposed by Toshimitsu *et al* [1994].

The layout of this chapter is as follows. In section 3.2, we present details of the imagined network architecture and the reverse link interface, and state various assumptions we will make regarding the channel characteristics. In section 3.3 we describe our proposed DT/CDMA MAC procedure in detail. Section 3.4 is devoted to an analytical modelling of the DT/CDMA protocol. To model the protocol performance, we use a renewal theory based Markov analysis derived by Resheff and Rubin [1990] as a basis for the model and derive the modifications necessary to incorporate our proposed MAC procedure.

The outputs of the Markov analysis are accurate expressions for the expected data throughput, data message blocking statistics and the expected data message delay. In section 3.5, results from the analytical model are compared to results obtained from a software simulator, for an imaginary network with a set of arbitrary (yet representative of a possible real world scenario) parameters. The advantages and contributions of our model and its results are discussed throughout the latter part of the chapter. The chapter concludes with a comparison of our proposed model to other similar existing models.

3.2 Network Model

In this section, we present the proposed system model for the data network. We detail the network architecture, describe the characteristics of the reverse link communications interface and present several assumptions that are made regarding the channel conditions.

3.2.1 Network architecture

The network architecture under consideration is a single-hop, star-network architecture in which a population of identical data terminals[†] communicate with each other, or with a global backbone network, via a centralised base station[‡]. This architecture is suitable for a cellular environment, or a stand-alone wireless LAN environment. For a cellular environment, we focus our attention on the performance of a single cell only in this thesis. User mobility issues and handoff traffic from surrounding cells (if any) are ignored.

All communications are done via a frequency division duplex (FDD) channel. In other words, transmissions from the base station to the mobiles (*forward link* or *downlink*) and from the mobiles to the base station (*reverse link* or *uplink*) are assumed to occur on separate, non-overlapping frequency bands. As is customary in the study of multiple access protocols for centralised systems, we concentrate on analysing the reverse link only, as it is inherently less efficient than the forward link. This arises from the fact that all forward link communications are co-ordinated by the base station and hence channel usage can be optimised, while the reverse link communications for a random access based protocol (such as the one proposed in this research) are generally uncoordinated and hence non-optimal. As such, it is in the reverse link that most researchers of MAC protocols are interested.

[†] We also refer to terminals as "users" or "mobiles".

[‡] For a wireless LAN scenario, the base station might be a hub or a file server.

3.2.2 Description of the reverse link communications interface

All communications on the reverse link implement direct-sequence CDMA (DS-CDMA), using our proposed random access Dual-Threshold CDMA MAC protocol which we will present in detail in section 3.3. Each data terminal communicates with the base station only, and is completely unaware of the state of transmission of any other terminal in the network. As discussed in section 2.2.1, the base station is assumed to allocate a unique PN code to each terminal in the network and is aware of each user in the network, and the code allocated to them. All data terminals transmit at the same fixed bit rate, B_d, and use the same processing gain, N. Each bit is thus spread using a terminal unique "long" PN sequence containing N chips, or M repetitions of a "short" PN sequence containing N/M chips. For a bit rate of B_d, this implies a "chip" rate of NB_d. In this thesis, we do not focus on the details of the spreading, PN sequences or code allocation scheme. Rather, we assume that the code pool is large enough to satisfy a reasonably large number of users.

Time is divided into fixed length slots of duration t_s, and it is assumed that all terminals in the network are exactly slot synchronised with the base station. The network is capable of supporting variable length data messages of any size. Each message is sub-divided into fixed length packets containing exactly Q bits, such that the transmission of a packet corresponds exactly with one slot (the packet or slot duration is thus equal to $t_s = Q / B_d$). Messages with more than one packet are transmitted contiguously over as many slots as there are packets in the message.

3.2.3 Physical channel assumptions

In our initial analysis of the DT/CDMA MAC protocol we will assume that, apart from the CDMA multiple access interference, the channel is perfect. Alternatively, we can assume that a perfect power control scheme is implemented. Effectively, this means that channel effects such as the near-far effect, shadowing, fading and multipath distortion are assumed to be completely negated by power control. We also ignore the effects of AWGN. These channel related problems are all very interesting by themselves and the degree to which they are overcome determines the success and quality of service the protocol is able to provide. In this chapter, we ignore these negative channel adversities and assume that *all reverse link signals are received with exactly equal power at the base station receiver*. By doing so, we can isolate and focus entirely on the performance of our DT/CDMA MAC protocol. Thereafter, the aforementioned adverse (or beneficial) channel effects can be considered.

Each mobile terminal is geographically displaced from all other terminals, and from the base station. The total transmission delay, defined as t_p, which incorporates both the signal propagation delay due to the physical separation between mobile and base station, and the packet processing delay, is assumed to be always less than half the duration of a system time slot (i.e. $t_p < t_s/2$). It is thus assumed that the total maximum round trip delay (mobile to base station and back again) is always less than the duration of a slot. This assumption is not at all invalid for small micro-cellular/packet radio networks in which cell radii seldom exceed a few kilometres.

The base station is assumed to have a permanent "line-of-sight" with each and every terminal. In this context, we refer to "line-of-sight" as the fact that a terminal's signal can always be "heard" by the base station. No terminal is ever "hidden" from the base station. Even in the presence of high signal-to-interference ratios when the bit error rate is high, we assume that the base station is still capable of the occasional PN sequence acquisition and is thus aware of the attempted transmission of each terminal, as well as the terminal identity. In knowing this, the base station is capable of determining the exact number of transmitting users, even during periods of high SIR.

3.2.4 Forward link and signalling assumptions

The transmission of synchronisation timing, power control, code allocation, and any other forward link signalling information does not fall under the focus of this thesis and, for the purposes of isolating and evaluating the reverse link MAC protocol, we assume that the forward link (or perhaps a separate signalling channel) carrying this information is "perfect" (errorless).

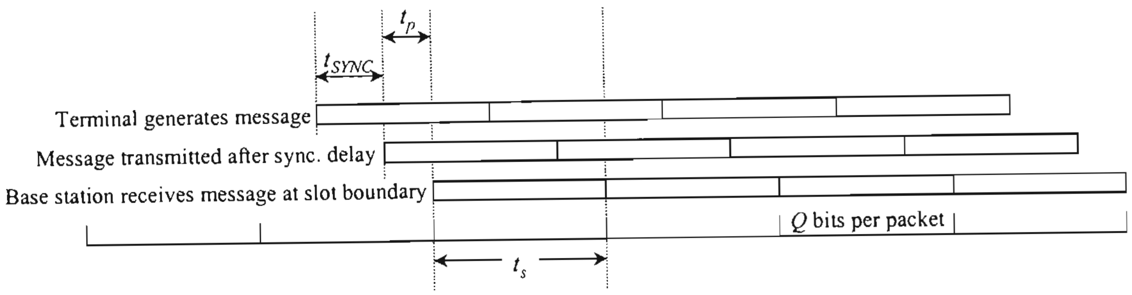

t_p
t_{SYNC}
Terminal generates message
Message transmitted after sync. delay
Base station receives message at slot boundary
Q bits per packet
t_s

Figure 3.1 : Synchronisation and packetisation of messages in the reverse link

Figure 3.1 illustrates the concept of message packetisation and slot synchronisation in the reverse link. A terminal synchronises itself by delaying its transmission for time t_{SYNC} (obtained via the signalling and timing information) in order for each packet to be received, after transmission delay t_p, in time with the next base station slot boundary.

3.3 Description of the Dual-Threshold CDMA MAC Protocol

Dual-Threshold CDMA is a centralised MAC protocol that is based on a combined ISMA/CDMA protocol that works in conjunction with *two* multiple access thresholds. In this section we will describe DT/CDMA in terms of the following three aspects: (1) Centralised ISMA/CDMA, (2) Application of two multiple access thresholds, (3) The MAC procedure description.

3.3.1 ISMA/CDMA: A centralised channel state broadcast scheme

In section 2.4.5, we discussed the concept of Inhibit Sense Multiple Access as an alternative to CSMA for centralised networks. In a packet radio environment, centralised channel load measurement and channel state broadcast schemes (such as ISMA [Prasad, 1991] and ICMA [Lee & Un, 1996]) have been shown to be superior to the distributed CSMA schemes in which each terminal senses the channel for itself and makes its own admission decisions. In narrowband ISMA, the base station detects if anyone is transmitting on the channel and, if so, broadcasts a “busy tone” to alert users of this fact, hence avoiding collisions by inhibiting any further access until the channel becomes free again. In other words, the base station reacts when the number of transmitting users equals or exceeds the maximum number of users allowed on the channel (which, for a shared narrowband channel, is obviously equal to one user).

In ISMA/CDMA the concept is the same, with the base station reacting to the number of users in the channel although, in CDMA, obviously more than one user is allowed in the channel. The base station measures the load (number of users) on the shared inbound CDMA channel and broadcasts the state of the network on separate out-of-band narrowband signalling tones. These signalling tones may be none other than unmodulated carrier frequencies, the presence or absence of which may signify a certain channel state. For example, a certain tone might be broadcast when the number of users exceeds the CDMA multiple access threshold. It is on this centralised ISMA/CDMA principle that we base our MAC protocol, and we justify this decision with the following reasons:

1. There is only one decision-maker (base station) in an inherently noisy environment.
2. The base station is assumed to have a "line-of-sight" connection with all terminals. No terminal is ever "hidden" from the base station.
3. As a result of (2) and the assumptions made in section 3.2.3, the base station is able to detect all incoming signals (even in poor signal-to-interference cases) and is thus assumed capable of obtaining an *exact* measurement of the channel load.
4. All packets arrive in a slot synchronised fashion at the base station and hence load levels are assumed to be constant during a slot.

3.3.2 Application of a Dual-Threshold model

As the name implies, the Dual-Threshold CDMA MAC protocol implements two thresholds, and herein lies our main contribution.

The first threshold we implement is the multiple access threshold which reflects the maximum acceptable multiple access interference (or error probability) as discussed in section 2.2.2. The proposed DT/CDMA is a random access protocol that attempts to avoid users exceeding this multiple access threshold. Throughout the description and analysis of DT/CDMA, we will use symbol β to represent this multiple access threshold for data traffic ($\beta = K_{\max}$ as described and computed in section 2.2.2). A more formal definition of β is: *the maximum acceptable number of simultaneously transmitting data users on the reverse link, such that the following criteria for data traffic is satisfied*:

$$\begin{aligned} P_E^{data}(x) &\le P_{E(\max)}^{data} && \textit{for} \quad x \le \beta \\ P_E^{data}(x) &> P_{E(\max)}^{data} && \textit{for} \quad x > \beta \end{aligned} \qquad \textbf{(3-1)}$$

where $P_{E(\max)}^{data}$ is some maximum acceptable error probability for data traffic due to the maximum acceptable MAI, x is the number of transmitting data users, and $P_E^{data}(x)$ is some probability of error given that x users are transmitting. Of course, this does not mean that errors always occur if the number of users exceeds β, just that the probability of an error occurring exceeds some predetermined acceptable maximum. A useful way to visualise this threshold is to imagine that there is a maximum of β virtual "channels" available on the reverse link, and only one user is allowed per channel.

Determination of this multiple access threshold, or this maximum MAI, is subjective and may depend on many factors such as: the network traffic type (voice, video, data etc...), the maximum tolerable bit error rate, the system's ability to recover from channel errors, etc... We defer any further discussion on the determination of β until the results section of this chapter (section 3.5) where, after having presented the MAC protocol and various other factors pertinent to the discussion, we are better prepared for the discussion.

The originality of DT/CDMA lies in the proposal and implementation of a second threshold, defined as α, which is *less than* the multiple access threshold β. The purpose of and justification for this α threshold will become clear after the MAC procedure has been presented and discussed. In DT/CDMA, we define and refer to these two thresholds, including the actions the base station performs with respect to them, as follows:

- The multiple access threshold β will hereafter be referred to as the "*collision threshold*". The base station broadcasts a "*Data Collision*" (**DC**) signal (tone) if the number of users exceeds β.

- Threshold α is referred to as the "*blocking threshold*". The base station broadcasts a "*Data Blocking*" (**DB**) signal if the number of users exceeds α.

3.3.3 The Dual-Threshold CDMA MAC procedure

Each data terminal can exist in one of three states: *idle*, *active* or *backlogged.* Idle terminals are terminals that have no messages to transmit. Active terminals are terminals that are in the process of transmitting a message. Backlogged terminals are terminals that are waiting to retransmit a message that was blocked or corrupted due to MAI. Based on these state definitions, the two thresholds and their associated signalling tones, we can describe the MAC protocol that each terminal in the network adheres to, as follows (refer to figure 3.2):

1. A data terminal that is ready to (re)transmit a message in the current slot n, first senses for the presence of the **DB** tone. If the **DB** tone is present, the terminal defers transmission (the message is *blocked*) and (re)enters the backlog retransmission routine (step 6).
2. If the **DB** tone is not present, the terminal initiates transmission immediately at the start of slot n and is known as a *contending data user*. This time slot, n, in which the first data packet of the message is transmitted, is known as the message's *arrival slot*.
3. If, after all *contending users* have started transmission in slot n, the data load exceeds β, the **DC** tone is emitted. Upon detecting the **DC** tone, ***all*** transmitting users immediately abort transmission and enter the retransmission routine (step 6). A slot in which the **DC** tone becomes active due to a β crossover is referred to as a *collision slot*. Due to the maximum round trip propagation delay, all terminals should have aborted by time $2t_p$ after the start of any collision slot (see figure 3.2).

4. If no collision occurs in slot n, then the *contending users* become *admitted users* in slot n+1. That is, a message is only admitted if neither the **DB** nor **DC** tones were present during its *arrival slot n*. Once admitted however, a user does not secure the channel until message completion, since all data users, including admitted users, abort upon activation of the **DC** tone.
5. Terminals are notified via an ARQ (Automatic Repeat Request) or acknowledgement time-out system on the forward or signalling link, whether a message was successfully received (received without errors) or not. If a message is corrupt, the terminal enters the backlog retransmission routine (step 6). In DT/CDMA we assume that corruption of even a single packet requires retransmission of the entire message.
6. Aborted, blocked or corrupt messages are retransmitted after a random time period, the duration of which is geometrically distributed with parameter ρ_{dR} [§] and hence a mean retransmission delay (backoff time) equal to $\overline{B} = \rho_{dR}^{-1}$ (measured in slots).

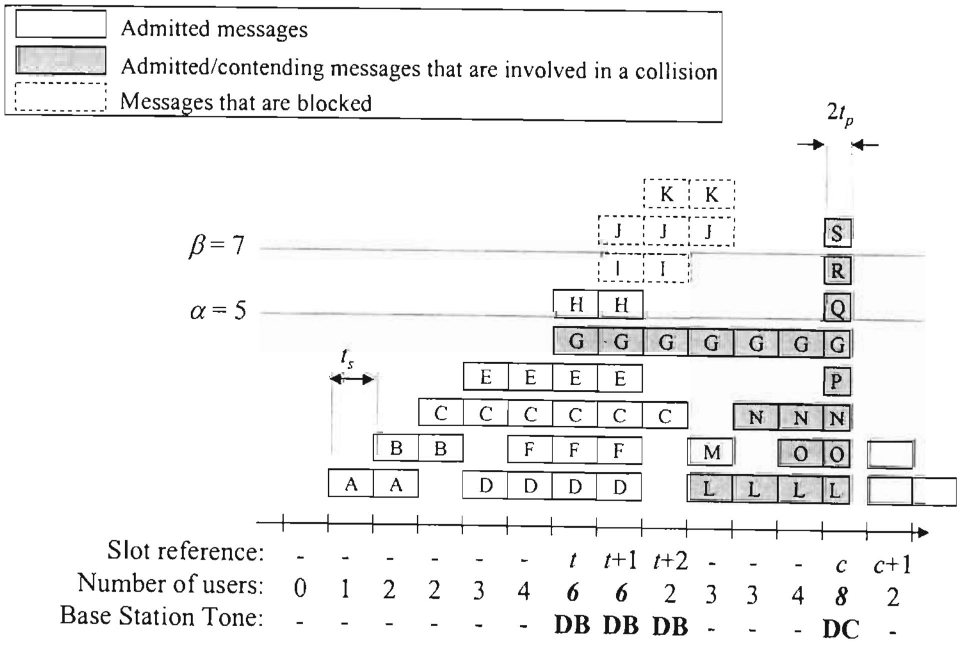

Admitted messages
Admitted/contending messages that are involved in a collision
Messages that are blocked
$2t_p$
$\beta = 7$
$\alpha = 5$
t_s
K K
J J J
I I
H H
G G G G G G G
E E E E
C C C C C C
B B
F F F
D D D D
A A
M
L L L L
N N N
O O
P
Q
R
S
Slot reference: - - - - - - t t+1 t+2 - - - c c+1
Number of users: 0 1 2 2 3 4 6 6 2 3 3 4 8 2
Base Station Tone: - - - - - - DB DB DB - - - DC -

Figure 3.2 : Illustration of the DT/CDMA MAC protocol

[§] Throughout this thesis, we will use the symbol ρ to represent a message (or call) arrival probability and symbol μ to represent a message (or call) termination probability. The subscript denotes the nature of the message or call. The "d" subscript is used to denote data traffic while the "R" subscript implies that the data message is a retransmission. Later, a "v" subscript will be used for voice traffic.

The purpose and justification of the blocking threshold α, which is lower than the collision threshold β, might not be obvious initially. A "collision" occurs when the number of users exceeds β. For example, in figure 3.2, contending messages P, Q, R and S arrive in slot c and, together with previously admitted messages L, N, O and G, cause the number of users to exceed the example multiple access threshold of $\beta = 7$ users. During a collision slot, we assume that all involved packets are corrupt with a high, unacceptable probability. Since we have assumed in our model that any corruption requires retransmission of the entire message, it is obvious that continued transmission of a message that is almost certainly corrupt is wasteful of time and bandwidth. Collisions prove to be a significant degrading factor to the system performance and it is thus a priority that collisions in the data protocol be avoided to as large a degree as possible. One way to achieve this is to block users before the maximum threshold is reached. Collisions cannot be totally avoided however, since the probability still exists of having many users attempting access in the same slot. Should a collision occur however, the decision to abort all transmissions, under the assumption that these messages are corrupted with a high probability, will allow all terminals to start again.

The purpose of the α threshold is thus to regulate the data load at a safe operating region below β and to decrease the probability of collisions. In figure 3.2 for example, we see that the arrival of message H in slot t causes the channel load to exceed the α threshold and the **DB** tone is transmitted by the base station. In slot t+1, we have avoided a collision since arriving messages J and I will not transmit in the presence of the **DB** tone. In slot t+2, we see that the **DB** tone is still present even though the number of users has dropped back to below α. The reason for this is that, at the start of slot t+2, the base station measures the channel load to be less than α and terminates transmission of the **DB** tone. Due to the signal propagation delay (t_p) however, the users will only "hear" the **DB** tone switch off at time t_p after the start of slot t+2. This means that any arriving message, which wishes to initiate transmission in slot t+2 (e.g. message K) will be blocked. The consequences of this delay are well documented for systems which implement distributed or centralised channel load sensing and broadcast schemes [Prasad, 1991], [Lee & Un, 1996], [Kleinrock, 1975b], [Tobagi & Hunt, 1980], [Chen & Li, 1989].

If the maximum time delay $t_{p(\max)}$ is negligible in comparison to the slot length t_s, then one may consider implementing a short guard time, of length $t_{p(\max)}$, at the beginning of each slot. During this time, contending messages wait to see if the **DB** tone is going to switch off before attempting to transmit. This method is not feasible if $t_{p(\max)}$ is such that the throughput advantage gained by having an accurate representation of the network load is less than the throughput lost due to the wasted bandwidth incurred by implementing the guard time.

In this thesis, we will not consider implementing such a guard time scheme, and analyse the protocol for the more general case in which $t_{p(\max)}$ can be as large as half the duration of a system slot. In other words, we do not dismiss the case where DT/CDMA may be used in an application where the packet duration is of the order of magnitude of the signal propagation delay.

In the results section of this chapter (section 3.5), we will discuss the determination of α, and show in fact that α should be a variable value that is based on the expected message arrival rate.

3.4 System Performance Analysis

When considering the performance of a data protocol, the two values most representative of the protocol performance are the average data throughput and the average message delay. Other important performance measures are often the average data message blocking probability and the average number of required retransmissions before a message is correctly received. In this section, we contribute a mathematical and statistical analysis of the DT/CDMA protocol, the outputs of which are accurate expressions for these performance measures. The basis for the analytical model is obtained from [Resheff & Rubin, 1990], and the model itself is an exact Markov analysis. We modify the analysis of [Resheff & Rubin, 1990] in order to accommodate our dual threshold scheme and associated MAC protocol.

3.4.1 Data traffic model

In modelling the performance of the DT/CDMA protocol, we will require a model for the data traffic generated in the network. For our purposes, this model will need to characterise the arrival process of messages in the network, as well as the statistics of the length of messages in the network. We define N_d as the number of data terminals in the network. With respect to the message arrival (generation) process, we will assume the popular Poisson model. The Poisson model assumes that the message arrival process appears to be generated by an infinite number of terminals ($N_d = \infty$), and that the composite data message arrival process (from all data users) is a Poisson process. If the mean message arrival rate is known to be λ messages per second, then we can express the expected number of messages being generated by the population per slot to be $G_d = \lambda t_s$, where t_s is the slot duration. The probability density function for the occurrence of having *m* messages being generated in a slot is given by the Poisson distribution

$$f_\infty(m) = \frac{e^{-G_d} {G_d}^m}{m!} \tag{3-2}$$

where the "∞" subscript denotes an infinite population. Parameter G_d is also known as the *expected offered data load.*

In this thesis, we will assume that data messages can be of any length. For the statistical description of messages, we will assume that the distribution of data message lengths is geometric in terms of the number of packets per message. This means that the probability of a data message completing transmission of a message in a slot is a fixed value, which we define for our use as μ_d. If L_d is a random variable representing the length of a data message, then the probability that a data message containing *l* packets is generated is given by

$$\begin{aligned} P_{ML}(l) &= P\{L_d = l\} \\ &= (1-\mu_d)^{l-1} \mu_d \end{aligned} \tag{3-3}$$

The mean length of data messages is defined as $\overline{L}_d = \mu_d^{-1}$. Throughout this thesis, we will try to use subscripts that easily describe the content of the distribution or probability. For example, $P_{ML}(l)$ denotes the probability distribution for the **M**essage **L**ength, P_{MS} will be used for the probability of **M**essage **S**uccess, P_{MB} the probability of **M**essage **B**locking, P_{BE} the probability of **B**it **E**rror etc...

3.4.2 Description of all pertinent variables and parameters

Before beginning the analysis, we list the variables and network parameters defined so far, as well as those pertinent ones that will be defined within the analysis that follows:

N_d	: the size of the data population
B_d	: the data bit rate
N	: the processing gain (number of chips per bit)
x	: the total number of transmitting data users (state of the data network)
$\hat{x}$	: the number of transmitting data users as seen by a reference user
β	: the data collision threshold (based on the maximum MAI)
α	: the data blocking threshold $(0 \le \alpha \le \beta)$
t_s	: the length of a slot
t_p	: one-way signal propagation delay + processing delay
τ	: the length of a Transmission Period
Q	: the total number of bits per packet
b_c	: the number of correctable bits per packet (due to block FEC)
λ	: the expected message arrival rate (messages per second)
G_d	: the expected message arrival rate (messages per slot)
ρ_{dR}	: the probability of a user retransmitting its message in the current slot
$\overline{B} = 1/\rho_{dR}$	: the average backoff delay for retransmitted data messages
μ_d	: the probability of a user completing a message in the current slot
L_d	: the length of a data message
$\overline{L}_d = 1/\mu_d$	: the average data message length

To simplify the analysis notation, we will use $b(m,M,p)$ to denote the binomial distribution with parameters M, m and p:

$$b(m,M,p) = \begin{cases} \binom{M}{m} p^m (1-p)^{M-m} & m \geq 0, M \geq 0, m \leq M \\ 0 & m < 0 \\ 0 & M < 0 \\ 0 & m > M \end{cases} \tag{3-4}$$

Furthermore, we also define $(A,B)^+ = \max(A,B)$ and $(A,B)^- = \min(A,B)$.

3.4.3 Definition of the data network state

Let the state of the data network, x, be defined as the number of transmitting data users in the network during a system slot. As in [Resheff & Rubin, 1990], we also define a *Transmission Period* (TP) to be the time between two successive data collisions (i.e. the number of slots between those slots in which a β crossover occurs). Since all data users abort transmission on detecting a collision, the state of the data network returns to a known state, i.e. $x = 0$ in the last slot of the TP. The following parameters are associated with a TP, and are illustrated in figure 3.3.

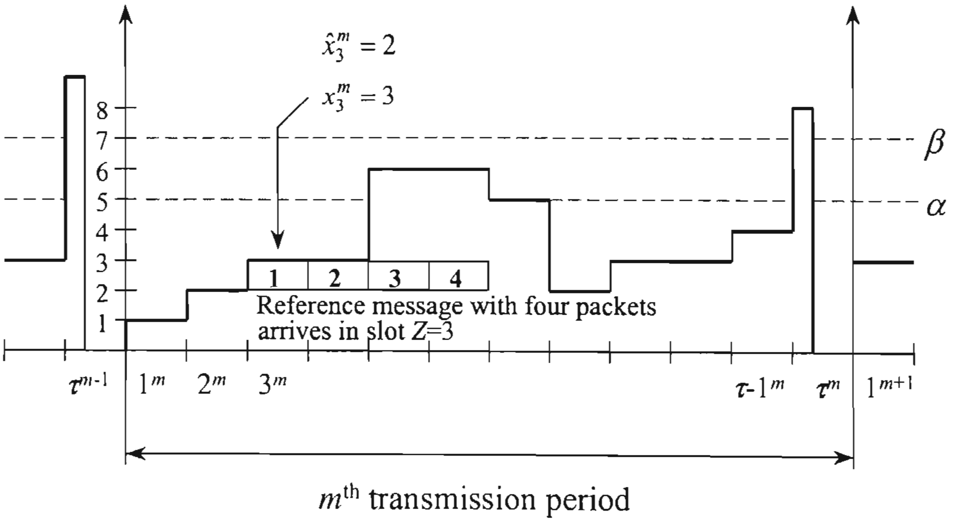


Figure 3.3 : Illustration of a Transmission Period

x_n^m : the state of the network in the n^{th} slot of the m^{th} TP

$\hat{x}_Z^m$: the state of the network as "seen" by the *first* packet of a reference message that arrives at slot Z within the m^{th} TP

τ^m : the length of the m^{th} TP (in slots)
this value represents the slot number in which the β crossover occurs

Throughout the remainder of this thesis, the "hat" symbol (^) indicates that the variable, expression, or value of interest is as "seen" by a certain reference message (or terminal) in the channel. For example, whereas x is the number of transmitting users as seen by the base station, $\hat{x}$ will be the number of transmitting users as seen by the reference message (***excluding itself***). Obviously, in this example, $\hat{x} = x - 1$.

The first slot after which the users abort transmission due to a β crossover marks the start of a new TP and is numbered slot 1. The last slot of the TP, τ, is the one in which the next β crossover occurs. The discrete time point process which marks the start of each TP is a renewal process, since each TP starts from the zero state. The global state x, when viewed over all TP's, is then a regenerative stochastic process, where the regeneration points occur at the start of each TP. Since we are interested in steady state conditions we, for the remainder of the analysis, ignore the superscript m, as we assume that all TP's have equal statistics at steady state conditions.

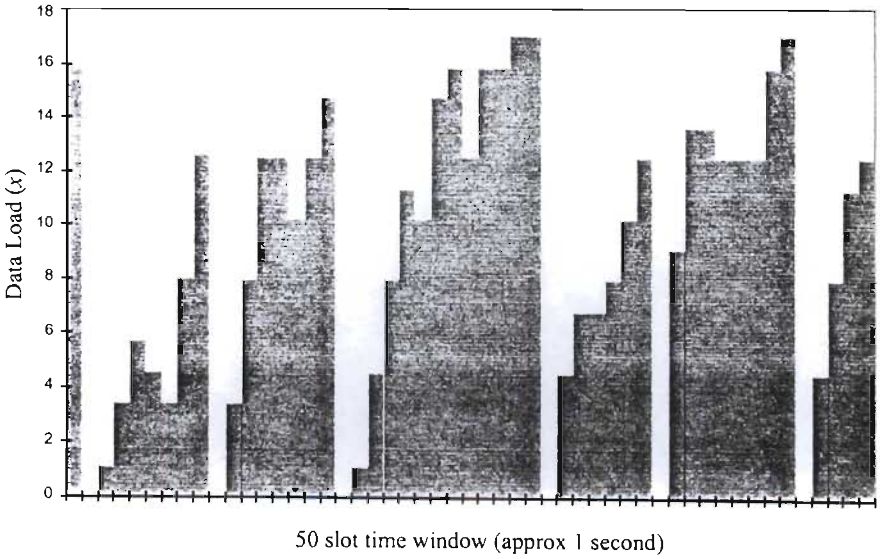


Figure 3.4 : Simulated data load showing several TP's

Figure 3.4 shows an example sequence of data load levels obtained from the software simulator for an arbitrary set of network parameters ($\beta = \alpha = 17$, $\overline{L}_d = 4$, $t_s = 16\text{ms}$). The concepts and details of the simulation model are discussed in the Appendix of this dissertation. The concept of TP's can be clearly seen, with data message abortions taking place whenever the data load exceeds $\beta = 17$.

3.4.4 Computation of the data state transition probability matrix

We begin by recalling that the state of the data network, x, is the number of transmitting data terminals in the current time slot. During a TP, this data state x changes on a slot-by-slot basis, and is represented by a *memoryless process*. The reasons for it being a memoryless process are:

1. The point process representing the message arrival times is a Poisson process.
2. The point process representing the end of a message is a simple Bernoulli trial.
3. The channel admission scheme is only dependent on the number of transmitting users in the current slot, and in the previous slot.

The number of messages arriving in the current slot n is completely independent of the channel history (i.e. the state of the network in all slots prior to slot n). Similarly, the number of terminating messages in the current slot is also completely independent of

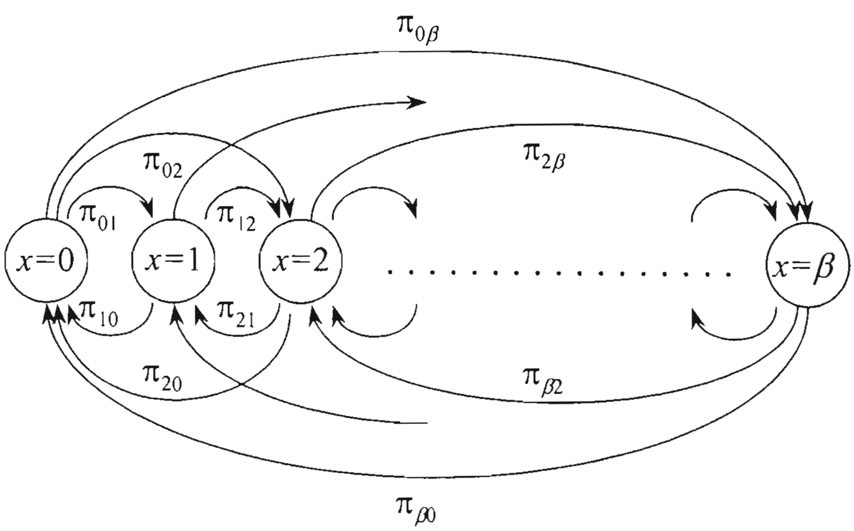


Figure 3.5 : Markov chain for data users

the channel history. Since the entire process for x can be obtained from the information contained in the current slot and the previous slot, we can represent it as a simple Markov chain with a two-dimensional state transition probability matrix. This Markov chain is illustrated in figure 3.5. Throughout the remainder of our analysis, we will only require the states up to and including β (i.e. $\beta+1$ states).

The term π_{ij} is defined as the steady state transition probability of moving from the current state $x=i$ in slot n to the next state $x=j$ in slot n+1. That is, $\pi_{ij} = P\{x_{n+1} = j | x_n = i\}$. The elements π_{ij} can be structured as a two-dimensional $(\beta+1)\times(\beta+1)$ state transition matrix, $\boldsymbol{\pi}$

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{00} & \pi_{10} & \cdots & \pi_{\beta 0} \\ \pi_{01} & \pi_{11} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{0\beta} & \pi_{1\beta} & \vdots & \pi_{\beta\beta} \end{bmatrix} \qquad \textbf{(3-5)}$$

The elements of $\boldsymbol{\pi}$ for $0 \le (i,j) \le \beta$ are computed as follows:

$$\pi_{ij} = \begin{cases} \sum_{k=(i-j,0)^+}^{i} f_\infty(j-i+k)\cdot b(k,i,\mu_d) & i \le \alpha \\ b(i-j,i,\mu_d) & i > \alpha, j \le i \\ 0 & i > \alpha, j > i \end{cases} \qquad \textbf{(3-6)}$$

If the current state i is greater than α, then the **DB** tone will become active in this slot and all new messages in the next slot will be blocked. Thus the only state transitions that may occur are due to messages which terminate in the following slot (i.e. $i-j$) with probability μ_d. If the current state i is less than or equal to α, then the **DB** tone will be off and new data messages will be admitted into the channel. The dummy variable k represents the number of messages (out of the i that are transmitting) that terminate with probability μ_d. The binomial term $b(k,i,\mu_d)$ is the probability of this occurrence. The number of new messages that arrive is then the difference $j-i+k$, where the probability of this occurring is $f_{\infty}(j-i+k)$.

3.4.5 Distribution of the length of a Transmission Period

The length of a TP is defined as random variable τ (measured in slots). The probability of having a transmission period of length τ is simply dependent on the fact that a collision occurs in slot τ of the TP. For this to occur, the system state x must be such that

$$P\{\tau = t\} = P\{x_1 \leq \beta, x_2 \leq \beta, \ldots\ldots, x_{t-1} \leq \beta, x_t > \beta\}. \qquad \textbf{(3-7)}$$

For the first $\tau - 1$ slots of the TP, we require that the stochastic process governed by equation 3-6 is such that $x \leq \beta$. In the τ^{th} slot we must have $x > \beta$. Equation 3-7 can then be expanded as

$$P\{\tau = t\} = \begin{cases} C(0) & t = 1 \\ \sum_{i=0}^{\beta} \sum_{j=0}^{\beta} \pi_{0i} \cdot [\boldsymbol{\pi}^{t-2}]_{ij} \cdot C(j) & t > 1 \end{cases} \qquad \textbf{(3-8)}$$

where $C(j)$ is the probability of having a β crossover in slot t given that there were j transmissions in slot t-1.

$$C(j) = 1 - \sum_{k=0}^{\beta} \pi_{jk} \qquad j \leq \beta \qquad \textbf{(3-9)}$$

The expected length of a transmission period is now given by

$$\bar{\tau} = E\{\tau\} = \sum_{t=1}^{\infty} t.P\{\tau = t\} \qquad \textbf{(3-10)}$$

As discussed in section 2.4.1, we are interested in computing the performance probabilities for a single reference message in the channel. Since all messages are subjected to the same conditions, these probabilities and expressions can then be used to obtain the overall measurements of the system performance. The success of an arriving reference message is dependent on three factors:

1. The reference message is not blocked upon arrival by the **DB** tone
2. A collision does not occur during transmission of the message
3. All packets in the reference message are received uncorrupted at the base station

We begin finding an expression for the message success probability by obtaining an expression for the success of the *first* packet in the message. For this purpose, we use the previously defined state variable $\hat{x}_Z$ (section 3.4.3). This represents the number of transmitting users (excluding itself) that the first packet of the arriving reference message "sees" given that the reference message is ready to transmit its first packet in slot Z of the TP.

3.4.6 Computing the probability that the reference message begins in slot Z

The first step is to compute the probability that the reference message begins in slot Z of the TP. For this purpose, we use a property of renewal processes known as the "age" of the process [Kleinrock, 1975a], [Resheff & Rubin, 1990]. In the context of our renewal process, random age variable Z represents the number of slots that have passed since the previous β crossover. If the distribution and mean of the renewal process are known (equations 3-8 and 3-10), then the distribution of the "age" variable Z is given by [Kleinrock, 1975a]

$$P\{Z = z\} = \frac{P\{\tau \geq z\}}{\bar{\tau}} \qquad \textbf{(3-11)}$$

Equation 3-11 also represents the probability that an arriving message finds the age of the TP being equal to Z slots. The "residual life" of a TP (Z') represents the expected time to the next β crossover. The distribution of Z' is given by

$$P\{Z' = z\} = \frac{P\{\tau \le z\}}{\overline{\tau}} \tag{3-12}$$

3.4.7 The state distribution as seen by the first packet of a reference message

We are now in a position to compute the probability that a reference message, arriving in slot Z, finds $\hat{x} = j$ other transmitting data users. This probability, defined as $\hat{X}_Z(j) = P\{\hat{x}_Z = j\}$, is solved using the following recursive algorithm:

1. If the reference message arrives in the first slot of the TP (i.e. Z=1), then we define

$$\hat{X}_1(j) = P\{\hat{x}_1 = j\} = P\{\hat{x} = j \mid Z = 1\} = \pi_{0j} \tag{3-13}$$

2. For Z=2, we apply the transition probability matrix $\boldsymbol{\pi}$ to the distribution obtained in step 1 to obtain the next state distribution. We also condition on the fact that the reference message arrives during the current TP, i.e. before or during the β crossover (this occurs with probability $P\{\tau \ge 2\}$). If it arrives after the β crossover, it will obviously not affect the state vector for the current TP. This condition was not explicitly included in equation 3-13 since $P\{\tau \ge 1\} = 1$. Thus

$$\hat{X}_2(j) = P\{\hat{x} = j \mid Z = 2\} = \sum_{i=0}^{\beta} \frac{\hat{X}_1(i) \cdot \pi_{ij}}{P\{\tau \ge 2\}} \tag{3-14}$$

3. Solving recursively, we can compute the state probability distribution for any slot R in the TP,

$$\hat{X}_R(j) = \sum_{i=0}^{\beta} \frac{P\{\tau \ge R-1\} \cdot \hat{X}_{R-1}(i) \cdot \pi_{ij}}{P\{\tau \ge R\}} \tag{3-15}$$

Since a recursive procedure is used, we have to account for the condition that a β crossing occurs at or after slot R, hence the term $P\{\tau \ge R-1\}$ is included in the numerator to cancel the denominator condition in the previous recursive operation for slot R-1. Using equation 3-11, we uncondition the state distribution on R to give

$$\hat{X}(j) = \sum_{R=1}^{\infty} \hat{X}_R(j) \cdot P\{Z = R\} \qquad \textbf{(3-16)}$$

We have now obtained the probability that the reference message sees j other data packets transmitting in its arrival slot. The next step is to compute the probability that the arriving reference message is not blocked in its arrival slot by the **DB** signal.

3.4.8 Computation of the message blocking probability

The decision to block messages is always based on the state of the previous slot, as discussed in section 3.3.3 and illustrated in figure 3.2. When the base station measures the channel load to exceed the α threshold, the **DB** tone is emitted. From figure 3.2, it can be seen that new messages are only blocked in the following slot due to the short propagation and processing delay t_p. In figure 3.2, message K is blocked because the **DB** tone is only heard to turn off after the delay t_p, even though the load was below α in the arrival slot of the message. Thus new messages are only blocked if the state of the network as seen by the reference message in the slot prior to its arrival exceeds α, and is less than or equal to β. We can compute the probability of blocking the reference message at its arrival slot, given that it sees $\hat{x} = j$ other transmissions on arrival, as

$$P_{MB}(j) = \sum_{i=\alpha+1}^{\beta} \frac{\hat{\hat{X}}(i) \cdot \pi_{ij}}{\hat{X}(j)} \qquad \textbf{(3-17)}$$

where $\hat{\hat{X}}(i)$ is defined as the probability of seeing i users transmitting in the slot prior to the reference message's arrival. Since equation 3-16 has been unconditioned on the number of the slot in the TP, we can safely assume that $\hat{\hat{X}}(i) = \hat{X}(i)$ for all i. Unconditioning equation 3-17 on j gives the overall average data message blocking probability

$$P_{MB} = \sum_{j=0}^{\beta} \hat{X}(j) \cdot P_{MB}(j) \tag{3-18}$$

3.4.9 Computation of a packet's success probability when MAI is considered

If the reference message is not blocked, then the probability of success of its first packet is conditioned only on the fact that the MAI is insufficient to cause any bit errors in the packet. If each packet is assumed to contain Q bits, then the probability of Packet Success, denoted by subscript *PS*, given $j+1$ users (j other data users plus the reference user) is conditioned on the fact that all Q bits are received correctly, and is defined as:

$$P_{PS}(j+1) = \left[1 - P_{BE}^{SGA}(j+1)\right]^{Q} \tag{3-19}$$

where $P_{BE}^{SGA}(j+1)$ is the probability of bit error during the $j+1$ spread-spectrum transmissions. In this analysis, we will use the SGA model for MAI. We ignore the effects of AWGN and use equation 2-6 with σ=3 (random chip and phase model). If we take cognisance of the fact that block error coding can be used to ensure that at most b_c bit errors in the packet can be corrected, then we can use the following expression to compute the probability of packet success

$$P_{PS}(j+1) = \sum_{b=0}^{b_c} \binom{Q}{b} . \left[1 - P_{BE}^{SGA}(j+1)\right]^{Q-b} . \left[P_{BE}^{SGA}(j+1)\right]^{b} \tag{3-20}$$

Equation 3-20 assumes that the packet is transmitted over a memoryless binary symmetric channel where the average probability of bit error is $P_{BE}^{SGA}(j+1)$ during $j+1$ transmissions [Morrow & Lehnert, 1992].

3.4.10 Computation of the message success probability

We define $R_n(j)$ as the probability of success of all packets in the message up to and including the n^{th} packet given that there are $\hat{x} = j$ other data users transmitting during the n^{th} packet. The probability of success of the first packet in the message is conditioned on the fact that the message is not blocked, and that all bits are received correctly. Thus for n=1, we have

$$R_1(j) = \hat{X}(j).[1 - P_{MB}(j)].P_{PS}(j+1) \qquad j < \beta \tag{3-21}$$

In order to compute the probability of the entire message's success, we solve the following equation recursively:

$$R_n(j) = \sum_{i=0}^{\beta-1} R_{n-1}(i).\hat{\pi}_{ij}^{PS} \qquad n > 1, j < \beta \tag{3-22}$$

Equation 3-22 computes the probability of success of the first n packets in the message. Each iteration conditions on the success of the previous n-1 packets, the fact that the network state remains below β for the n^{th} packet, and the fact that the n^{th} packet is not corrupted by MAI. For this purpose, we define and use the state transition probability matrix $\hat{\boldsymbol{\pi}}^{PS}$ as seen by the reference message. Element $\hat{\pi}_{ij}^{PS}$ represents the probability that the n^{th} packet in the reference message is received correctly, given that it sees $\hat{x} = i$ users during the transmission of its (n-1)$^{\text{th}}$ packet and $\hat{x} = j$ users during transmission of its n^{th} packet:

$$\hat{\pi}_{ij}^{PS} = \begin{cases} \displaystyle\sum_{k=(i-j,0)^+}^{i} f_\infty(j-i+k)\cdot b(k,i,\mu_d).P_{PS}(j+1) & i \le \alpha - 1 \\ b(i-j,i,\mu_d).P_{PS}(j+1) & i > \alpha - 1, j \le i \\ 0 & i > \alpha - 1, j > i \end{cases} \tag{3-23}$$

Apart from the fact that the transition probability now includes the condition that the n^{th} packet is correctly received during the presence of j other users, matrix $\hat{\boldsymbol{\pi}}^{PS}$ differs from $\boldsymbol{\pi}$ in that the thresholds α and β now appear to be one user less than their actual values. This occurs because the reference packet is not included in the ij subscript. The reference packet may see no other transmitting packets ($\hat{x} = 0$) when, in fact, the actual system state during that slot is $x = 1$ user (the reference user). The probability of success of a message containing L packets is defined as $P_{MS}(L)$, where

$$P_{MS}(L) = \sum_{j=0}^{\beta-1} R_L(j) \tag{3-24}$$

The total average message success probability can be found by averaging over all possible message lengths

$$P_{MS} = \sum_{l=1}^{\infty} P_{MS}(l) \cdot P_{ML}(l) \tag{3-25}$$

If RT_L represents the number of times that a message of length L requires to be transmitted before it is successfully received then, given $P_{MS}(L)$, we can compute the expected value of RT_L to be:

$$\overline{RT_L} = \frac{1}{P_{MS}(L)} \tag{3-26}$$

3.4.11 Computation of the message and packet throughputs

Finally, the message throughput is simply the fraction of the offered load G_d that is successfully transmitted across the channel per time slot or, in other words, the number of successfully received messages per slot. We thus define the *message throughput* as

$$S_M = G_d.P_{MS} \tag{3-27}$$

We also compute the expected *packet throughput* S_P as the average number of successfully received packets per slot. This is given simply by counting the number of packets per successful message:

$$S_P = G_d.\sum_{l=1}^{\infty} l.P_{MS}(l).P_{ML}(l) \tag{3-28}$$

The expected number of transmitting data users per slot, or *average carried data load*, $\bar{x}$, can be computed as

$$\bar{x} = \sum_{k=1}^{\beta} k.X(k) \tag{3-29}$$

3.4.12 Computation of the average message delay

Finally, we compute the expected delay that a message is expected to experience. This delay is defined as the time difference, measured in slots, between the message's very *first* arrival slot and the final slot in which the entire message is correctly received. In order to compute the expected message delay, we need an expression for the expected number of slots that the reference message spends while in transmission. Since collisions imply that messages abort during transmission, a message of length L will only transmit all L packets if there is no collision during the message. For this purpose, we define two more state distributions, $T_n(j)$ and Y_n, where $T_n(j)$ is the probability that a reference message has transmitted n of its packets, while Y_n is the probability that a reference message has transmitted n-1 of its packets and experiences a collision during its n^{th} packet. $T_n(j)$ differs from $R_n(j)$ in that it does not matter whether or not the packet was successfully transmitted, just that it was transmitted. We thus divide by $P_{PS}(j+1)$ to negate the condition that all bits in the packet are received correctly.

$$T_1(j) = \frac{R_1(j)}{P_{PS}(j+1)} \tag{3-30}$$

$$T_n(j) = \sum_{i=0}^{\beta-1} \frac{T_{n-1}(i).\hat{\pi}_{ij}^{PS}}{P_{PS}(j+1)} \qquad n > 1 \tag{3-31}$$

For Y_n, we condition on the fact that n-1 packets are transmitted without collision and that a collision occurs during the n^{th} packet:

$$Y_1 = \sum_{i=0}^{\alpha} X(i).\hat{C}(i) \tag{3-32}$$

$$Y_n = \sum_{i=0}^{\beta-1} T_{n-1}(i).\hat{C}(i) \qquad n > 1 \tag{3-33}$$

where $\hat{C}(i)$ is the probability that the n^{th} packet in the reference message experiences a collision given that there were i users transmitting in the slot prior to the collision.

$$\hat{C}(i) = 1 - \sum_{j=0}^{\beta-1} \frac{\hat{\pi}_{ij}^{PS}}{P_{PS}(j+1)} \tag{3-34}$$

If variable A_L is defined to represent the number of packets, from a message of length L, that are transmitted before a collision occurs, then the probability that l out of the L packets are transmitted before a collision occurs is defined as

$$P\{A_L = l\} = \begin{cases} Y_l & l < L \\ \\ Y_L + \sum_{j=0}^{\beta-1} T_L(j) & l = L \end{cases} \tag{3-35}$$

In equation 3-35 there are two cases for l=L. Firstly, a collision may occur in the last packet of the message, with probability Y_L. In the second term, no collision occurs during the entire message duration. The average number of slots, $\overline{A_L}$, that a reference message of length L is expected to be transmitted for is then given by

$$\overline{A_L} = E\{A_L\} = \sum_{l=1}^{L} l.P\{A_L = l\} \tag{3-36}$$

The expected message delay for a message of length L packets, $\overline{D_L}$, is then conditioned on the expected number of retransmissions required ($\overline{RT_L}$), the average backoff delay between each retransmission ($\overline{B} = \rho_{dR}^{-1}$), and the amount of time that the message is expected to be in transmission ($\overline{A}_L$). These factors combine to give a mean delay for a message of length L of [Resheff & Rubin, 1990], [Kleinrock, 1975a]

$$\overline{D}_L = \frac{\overline{A}_L + \overline{B}.[1 - P_{MS}(L)]}{R\overline{T_L}} \tag{3-37}$$

since each retransmission occurs with probability $[1\text{-}P_{MS}(L)]$. The overall expected message delay, $\overline{D}$, is then given by

$$\overline{D} = \sum_{L=1}^{\infty} \overline{D}_L . P_{ML}(L) \tag{3-38}$$

3.5 Performance Results

This section provides example results for the DT/CDMA protocol. We compare the results obtained from the analytical models with results obtained from a custom-built software simulation of the network. The concept and workings of the simulator algorithm are presented in the Appendix.

3.5.1 Network parameters

The network parameters listed in table 3.1 were used in the simulation and analysis of the network. These values were chosen to reflect a possible real world scenario. The data bit rate of 64kbps and PN code length of 63 implies a spread-spectrum chip rate of 4.032Mcps and a processing gain of 18dB. The mean length of data messages is chosen to be 4096 bits, or exactly 4 slots (Q=1024 bits per packet). In section 3.5.5, we quantify the effects of varying the aggregate data message length. The backoff delay has been chosen to be geometrically distributed with mean length $\overline{B}$ =5 slots.

Item (*units*)	**Symbol**	**Value**
Data information bit rate (***kbps***)	B_d	64
Packet size (***bits***)	Q	1024
Number of correctable bits per packet (***bits***)	b_c	0
Data population (***terminals***)	N_d	∞
Spreading/Processing gain (***chips/bit***)	N	63 (18dB)
System slot size (***ms***)	t_s	$64kbps/1024bits = 16$
Signal propagation and processing delay (***ms***)	t_p	$<0.5\,t_s$
Data collision threshold (***transmissions***)	β	To be determined
Data blocking threshold (***transmissions***)	α	To be determined
Average data message length (***slots***)	$\overline{L}_d$	4
$\therefore$Data message termination probability	μ_d	0.25
Average data message arrival rate (***messages/slot***)	G_d	Variable
Average data backoff delay (***slots***)	$\overline{B}$	5
$\therefore$Data message retransmission probability	p_{dR}	0.2

Table 3.1 : Example network parameters used in the analytical evaluation and simulation

The choices of bit rate and packet size (slot size) will only start to affect the system performance when the round trip signal propagation delay $2t_p$ becomes comparable to the slot size t_s, and the base station to terminal feedback delay becomes larger than a slot. In assuming that the feedback delay is always less than a slot, we can always model the system as a memoryless process. As soon as the delay exceeds a slot, we lose the Markovian property and the analysis becomes significantly more complicated. For the given slot duration, the round trip propagation delay $2t_p$ only becomes comparable when the distance from the base station to the terminal exceeds approximately 2400km. The slot size is thus more than sufficient for a terrestrial micro- or pico-cellular network, or a wireless LAN.

For the purposes of evaluating the efficiency of the proposed protocol, we set the number of correctable bits per packet, b_c, to be zero. As has been shown in many other works, block FEC coding improves the overall network performance - the relationship between system performance and choice of FEC scheme being well known and documented. Equation 3-20 reveals that block FEC increases the overall probability of packet success, and hence the message throughput. All throughput and delay performance measures are thus scaleable based on the choice of b_c. We arbitrarily choose $b_c = 0$ in order to isolate and evaluate the system performance in the context of the proposed dual-threshold scheme.

3.5.2 Determination of the data collision threshold β

It is not obvious what the value of β should be for a wireless network supporting data traffic. What exactly is the maximum acceptable level of MAI for a channel that supports data traffic? Traditionally, it has been common practice to associate maximum tolerable bit error probabilities of less than say 10^{-6} for data. If we consult table 3.2 for an approximation of the BER (as found using the SGA approximation in equation 2-6 with $\sigma = 3$) versus the number of users, *k*, it can be seen that a bit error rate of 10^{-6} occurs when there are nine simultaneous transmissions. Does a threshold of β=9 provide the optimum system performance however?

K	$P_{BE}^{SGA}(k)$	*k*	$P_{BE}^{SGA}(k)$	*k*	$P_{BE}^{SGA}(k)$	*k*	$P_{BE}^{SGA}(k)$
0	0.0000e+00	10	2.2964e-06	20	8.0540e-04	30	5.3417e-03
1	0.0000e+00	11	6.8876e-06	21	1.0557e-03	31	6.0369e-03
2	2.6274e-43	12	1.6984e-05	22	1.3499e-03	32	6.7714e-03
3	1.2255e-22	13	3.6144e-05	23	1.6892e-03	33	7.5436e-03
4	1.0335e-15	14	6.8664e-05	24	2.0745e-03	34	8.3517e-03
5	3.1243e-12	15	1.1928e-04	25	2.5061e-03	35	9.1939e-03
6	3.9191e-10	16	1.9287e-04	26	2.9839e-03	36	1.0068e-02
7	9.9720e-09	17	2.9417e-04	27	3.5073e-03	37	1.0973e-02
8	1.0173e-07	18	4.2755e-04	28	4.0755e-03	38	1.1907e-02
9	5.8528e-07	19	5.9687e-04	29	4.6874e-03	39	1.2868e-02

Table 3.2 : Probability of bit error versus the number of transmitting users, *k* as computed using the SGA expression (equation 2-6 with σ=3)

For a data network, the definitive measure of the system performance is the system *throughput*, or the rate of flow of correct information across the channel. In our network, this throughput is the number of successfully transmitted *messages* (or *packets*) per unit time. If β is chosen to maximise the system performance, then it obviously becomes a function of all the factors and parameters that affect the system performance. The list of contributing factors is large and includes, amongst many other things: the traffic statistics, the error detection/correction scheme (if any), the channel conditions (the degree of fading, shadowing etc...), the information retransmission scheme, the type of information and the MAC protocol.

Let us, for instance, consider the information retransmission scheme. If a data network does not support any retransmission of information, then we would expect the multiple access threshold value to reflect the maximum tolerable bit error rate such that *all* packets are received correctly the *first* time. However, in order for such a scheme to be feasible, one would surely need to implement some form of FEC coding scheme that is capable of detecting and correcting any bit errors that do occur. In the context of a CDMA protocol, a data system that does not support retransmissions would thus require a very low value of β used in conjunction with a good FEC scheme.

If the network does not support retransmissions and can tolerate uncorrected errors, for example a voice network, then it is common practice to associate a threshold such that the maximum multiple access interference is tolerable to the speaker. This threshold would obviously be higher than for a system that requires the absolute minimum of errors.

In a system that supports retransmission of corrupt information, choosing a MAI threshold based on this same theory is pointless since, if we do not expect to require any information retransmission, then why implement information retransmission? In this case, it is more effective to base the channel threshold value on the maximum allowable *message error probability* (the message error probability has an inverse linear relationship with the message throughput), rather than some maximum allowable *bit error probability* [Bischl & Lutz, 1995]. The theory is as follows: When the threshold is increased, more users are allowed to transmit in the channel and hence more information is being transmitted across the channel. However, the increase in users is accompanied by an increase in MAI and hence the information has a higher probability of being in error. The optimum value of β is found when this difference between the throughput gained by allowing more information across the channel and the throughput reduction due to the increased MAI is maximised and in favour of the throughput advantage (obviously!). In other words, the optimum value of β occurs when the channel throughput is maximised. Since the throughput is directly related to the packet (or message) error probability, we can see that equation 3-1 is in fact valid if we consider $P_{E(\max)}^{data}$ as the maximum acceptable packet (or message) error probability.

A preliminary investigation into determining β for a given MAC protocol with a set of given network parameters revealed that it is not as trivial a task as it first appears. Although much research has been done on the topic of the MAI capacity and threshold determination for CDMA networks supporting voice traffic with no information retransmissions (e.g. the definitive works of [Viterbi & Viterbi, 1993] and [Gilhousen *et al*, 1991]), very little comprehensive or conclusive work could be found that considered CDMA capacity and user threshold issues for data networks with information retransmission. Since the focus of this thesis falls specifically on the MAC protocol, and the topic of channel capacity issues constitutes an entire thesis on its own, we choose to place a mathematical determination of β beyond the scope of this thesis, and rather obtain the optimum value of β through experimentation. For those interested in furthering the work done in this thesis, we propose this as a possible topic of further study.

The most obvious way to determine β through experimentation is to ignore the blocking threshold α and compute the system performance measurements for various settings of β. It should not take much thought to convince the reader that the simplest means of negating the effect of α is to set $\alpha = \beta$. Without the blocking component, the MAC protocol concedes its worst case throughput, but this allows us to isolate the effect that β has on the network performance.

After quantitative experimentation with β for the network parameters in table 3.1, it was found that a value of β=17 appeared to provide the optimum message throughput solution. In figures 3.6 and 3.7, we illustrate several message throughput versus offered data load curves for various values of β below and above (respectively) the apparent optimum value of 17. We also include the case β=∞, which is exactly equivalent to the standard slotted-SS/ALOHA case in which there are no channel restrictions (as discussed in section 2.5.1, although in this case the messages have variable length). We plot two graphs to avoid the confusion that sometimes occurs when there are too many curves that cross each other on the same plot. In most cases, analytical results are plotted as a solid line, while simulated results are plotted as single points (e.g. Δ,■,O,◆,●,□).

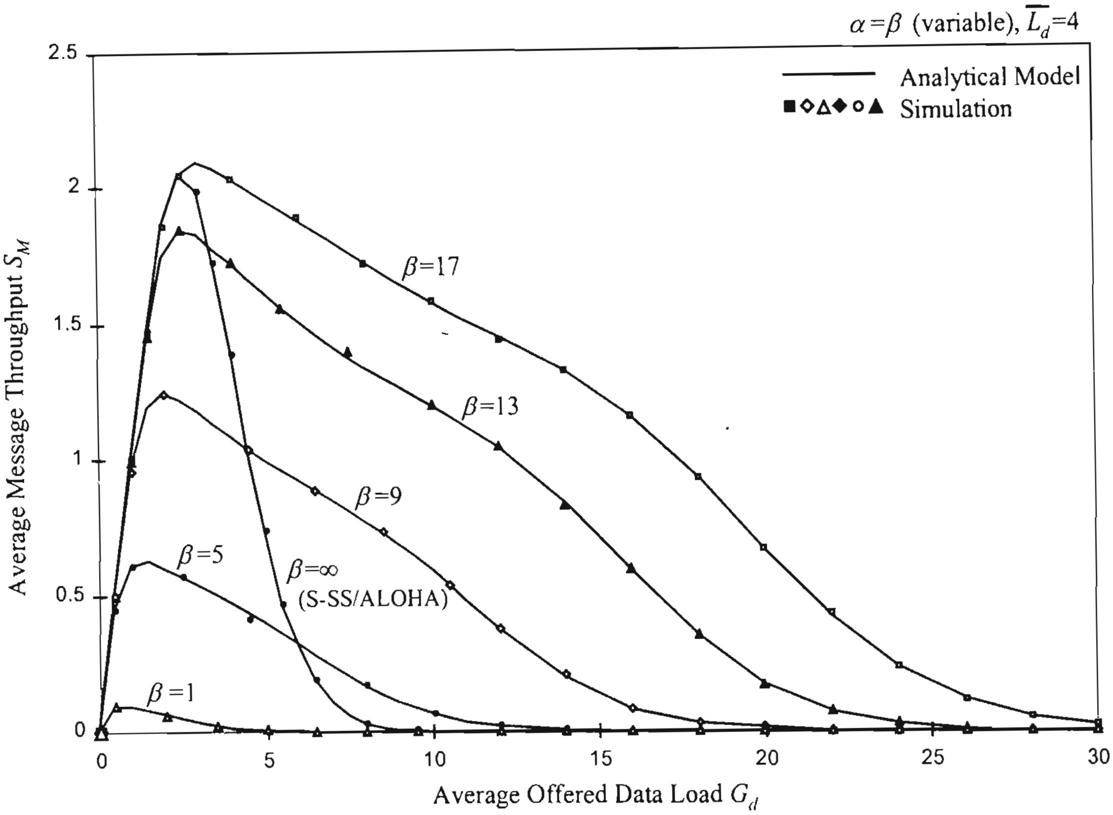


Figure 3.6 : Effect of β on the data message throughput (values of $\beta \leq 17$)

It is clear from figure 3.6 that for threshold values below β=17, the *maximum* throughput is always less than the S-SS/ALOHA case. This occurs because the collision threshold has been set to be too low, and full use is not being made of the channel capacity. At β=17, we see that the maximum throughput becomes equal to, if not slightly higher, than the S-SS/ALOHA model. At higher loads (G_d>5), the system performance improvement gained by employing the collision detection threshold β is very clear, even for values of β which are far below the apparent optimum value of 17. Later, in section 3.5.3, we will see that by employing the second message blocking threshold α which is less than β, we can improve further on these results.

In figure 3.7 we illustrate what happens to the system performance when the collision threshold is set to be too large (β>17). As $\beta \to \infty$, it can be seen that the message throughput curves start to approach that of the S-SS/ALOHA case. At high loads, the system performance is seen to increase again (the humps which occur for β=25 and β=40) before finally dropping off to zero. This can be explained by examining the purpose of the collision detection threshold.

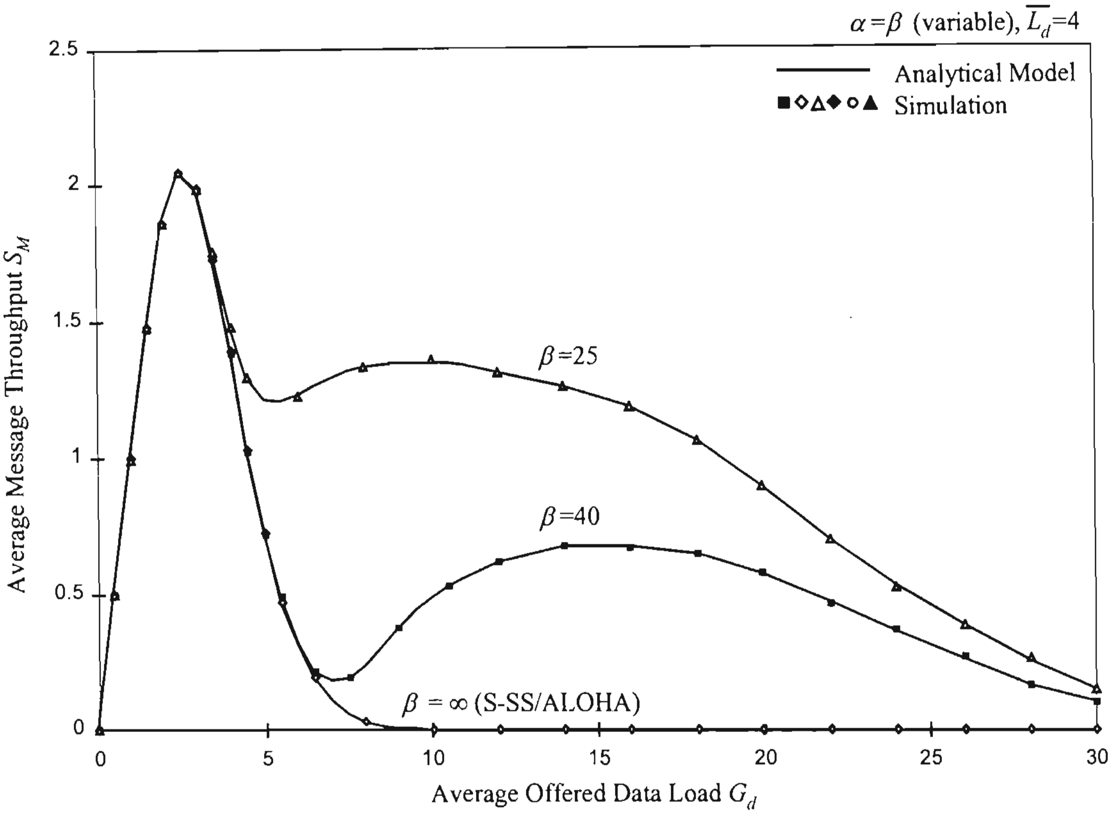


Figure 3.7 : Effect of β on the data message throughput (β>17)

Let us, for example, consider the case of β=40. In equation 3-29, we computed the average carried data load or the average number of transmitting users per slot, $\bar{x}$. When the channel access rate is low and there is no blocking or collisions then, obviously, we expect to admit, on average, G_d users per slot. If the average data message length is $\overline{L_d}$, then we expect the carried load to be approximately $\bar{x} \approx G_d.\overline{L_d}$. At very low loads (below G_d=2.5), we expect the carried load to be less than 10, where the average bit error rate is low enough to ensure that most messages are successful. This is explained by the linear region of the load/throughput curve for G_d<2.5. As G_d increases, $\bar{x}$

increases, the average bit error rate increases, and the fraction of messages that become corrupt starts to increase. At this point, the throughput starts to decrease. As G_d starts to approach 10 message arrivals per slot, the average carried load is expected to approach the threshold of β=40 users per slot. At this point, we expect collisions to start occurring and messages starting to abort. At each abortion, the carried load returns to zero and thereafter starts to increase again. For a short while after each collision we thus have a period of low carried load in which some messages are transmitted correctly with a low bit error rate. This explains why the throughput starts to increase again. At very high loads the frequency of collisions becomes too great and there is little chance of a message being able to complete transmission without experiencing a collision.

3.5.3 Effect of varying the data blocking threshold α

We now discuss the main crux of the proposed data protocol, namely the inclusion of a dynamic blocking threshold α that is lower than the user threshold β. In figure 3.8, we provide several packet throughput vs. offered load curves for different values of α, and

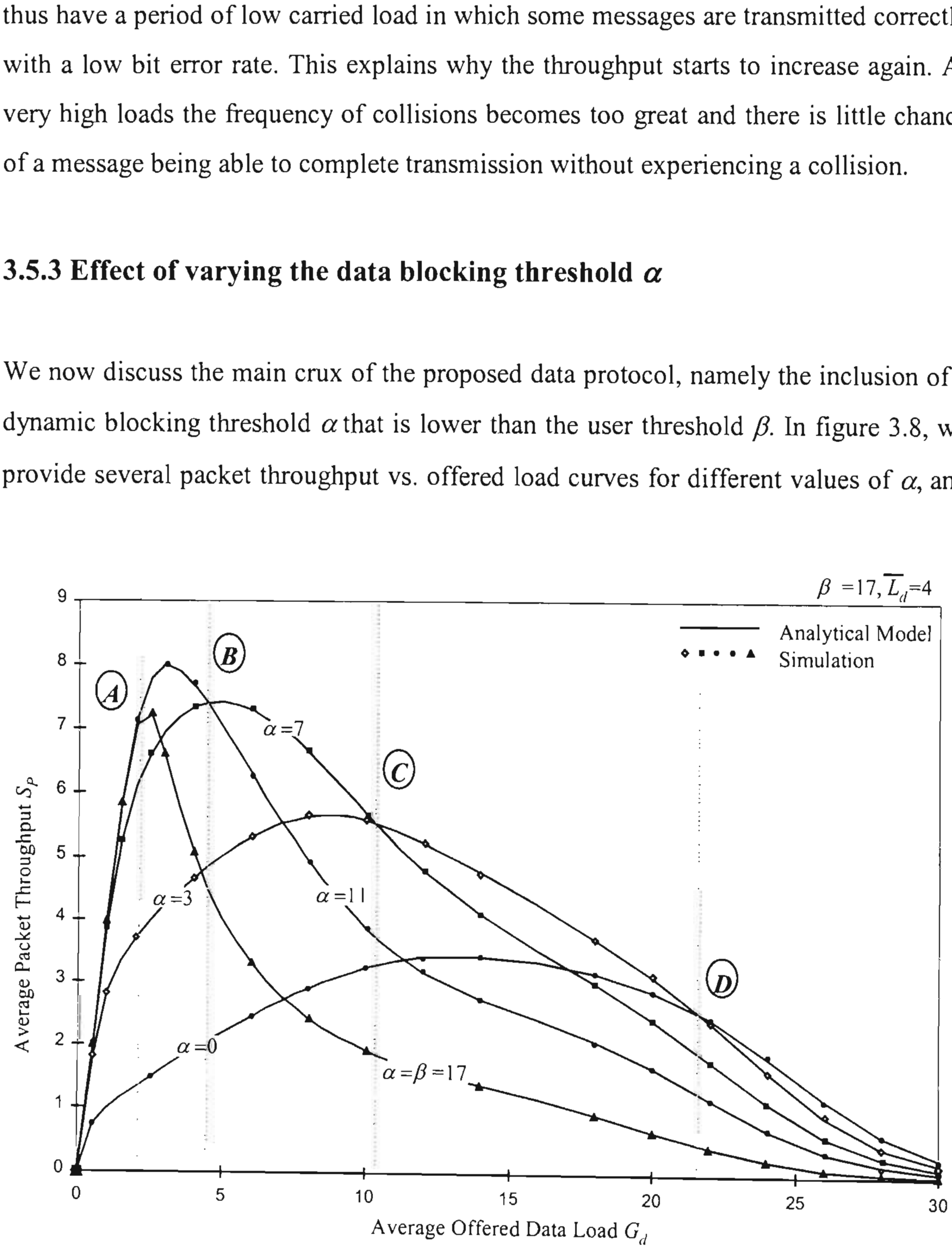


Figure 3.8 : Effect of varying α on the system throughput

a fixed collision threshold of β=17. For the sake of clarity, we plot curves for only a few (yet widely spread) values of α, namely α=0,3,7,11 and 17. We choose to present the *packet throughput* S_P rather than the *message throughput* S_M for various reasons. Firstly, the effect that α has on the system performance is more evidently seen for the packet throughput case. Our second justification will be presented later after analysing the exact reason why α improves the system performance.

Immediately, we see that the choice of α has a profound influence on the performance and secondly, that the effect that α has on the system performance is load dependent. At low loads, we see that high values of α provide better results than low values of α. The reverse is true at higher loads. For example, load (A) on figure 3.8 is the load at which it becomes more advantageous to have α=11 (if G_d is increasing). Similarly, load (B) represents the load at which α should drop to 7. Points (C) and (D) represent similar transition points. Obviously, there will be other intermediate transition points for the other values of α that are not plotted.

α	$\overline{\tau}$	P_{MB}	S_M	S_P	$\overline{D}$	P_{MS}
0	183.2372	0.912143	0.857736	3.415998	134.39	0.071478
1	93.61997	0.860605	1.127888	4.370999	101.237	0.093991
2	56.33509	0.840251	1.313534	4.930215	86.8887	0.109461
3	**36.39412**	**0.809898**	**1.451599**	**5.234913**	**79.4885**	**0.120967**
4*	24.68725	0.774959	1.556123	5.340626	76.401	0.129677
5	17.40312	0.735443	1.634416	5.282756	77.6026	0.136201
6	12.67326	0.690442	1.690949	5.092442	85.669	0.140912
7	9.497341	0.639007	1.728261	4.799647	111.009	0.144022
8	7.307893	0.580468	1.747326	4.433184	210.73	0.145611
9	5.768482	0.514742	1.748188	4.02051	1.01E+03	0.145682
10	4.672045	0.442705	1.731081	3.587894	2.08E+04	0.144257
11	3.887508	0.366639	1.697782	3.161051	1.74E+06	0.141482
12	3.329547	0.290509	1.652507	2.765531	5.67E+08	0.137709
13	2.938132	0.219185	1.601525	2.42329	*VLV*	0.13346
14	2.66472	0.156045	1.551249	2.144005	*VLV*	0.129271
15	2.467219	0.100761	1.506364	1.920509	*VLV*	0.12553
16	2.312717	0.049933	1.47001	1.737766	*VLV*	0.122501
17	**2.181285**	**0**	**1.445027**	**1.586623**	***VLV***	**0.120419**

Table 3.3 : Performance measurements for the full range of possible α values $(0 \le \alpha \le \beta)$ at a fixed offered load value (G_d=12 and β=17).

(*VLV* implies a very large value approaching infinity as α increases)

We now analyse the effect of α by considering a single offered load value and examining the effect that α has on the system performance for that particular load value. Table 3.3 provides various computed performance measures obtained from the analytical model for various different α values at a single, fixed load value. We choose G_d=12 since, as can be seen from figure 3.8, this load provides a wide spread of throughput values for different values of α.

We look firstly at the second column of table 3.3, namely the expected TP length $\bar{\tau}$. As discussed in section 3.4.3, a TP represents the number of slots between each collision. Clearly, we can see that by reducing the α threshold with respect to β, the probability of having a collision decreases and hence the average time between collisions increases. The reason for this is that the α threshold throttles message access in an attempt to avoid the load exceeding β. Once the load climbs above α, no more messages can be admitted until the load drops below α again. Figure 3.9 illustrates the distribution of TP lengths for values of α= 5,11 and 17, as measured in the simulator and found using equation 3-8. As can be seen, the mean TP length increases as α decreases. An obvious and unfortunate trade-off of decreasing α however, is that the average message blocking probability P_{MB} increases, as is shown in column three of table 3.3.

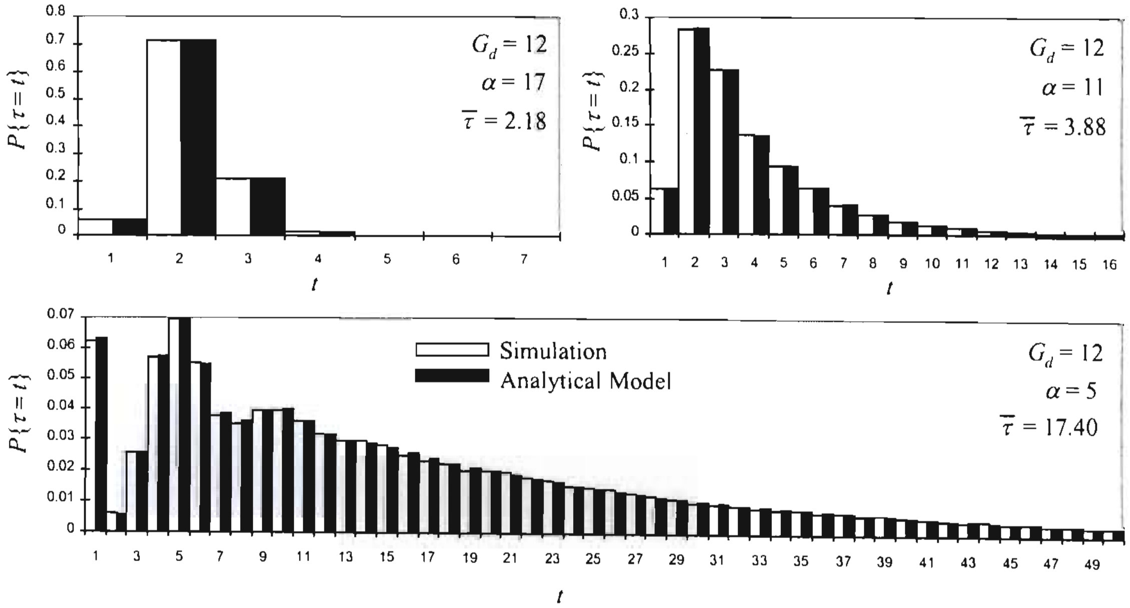


Figure 3.9 : Distribution of TP lengths for various values of α at G_d=12

For example, when $\alpha=\beta=17$, no blocking occurs but collisions occur with the highest frequency. In this scenario, the α threshold has no effect and the protocol basically degenerates into the collision-detection only scheme presented in [Resheff & Rubin, 1990]. The case $\alpha=0$ (i.e. blocking occurs as soon as one data user is admitted) provides the longest average time between collisions, but the high blocking probability causes the system performance to be very poor. Somewhere in the middle, we expect to find an optimum point where the trade-off is minimised. This middle point appears to occur at around $\alpha=4$ (marked in table 3.3 with a *) for this example load value of $G_d = 12$, where the average *packet throughput* S_P is largest and the average message delay $\overline{D}$ is lowest. We do note however, that the average *message throughput* S_M is **not** optimum at this α value.

At this point we justify our decision to base our measurements on the system's *packet throughput* rather than the system's *message throughput*. We highlight $\alpha=3$ and $\alpha=17$ in table 3.3 in order to clarify our justification, since we can see that the system yields almost equal average message throughputs for these two values of α, yet the average packet throughput for $\alpha=17$ is significantly poorer than for $\alpha=3$. For $\alpha=17$, we also observe an extremely large average message delay, whereas for $\alpha=3$ the average message delay is quite close to the optimum value that occurs for $\alpha=4$. An explanation for this phenomenon is made possible by examining the distribution of the probability of a message's success based on its length (expression $P_{MS}(L)$ in equation 3-24).

Figure 3.10 provides a graphical representation of the distribution of equation 3-24 for $\alpha=17$ and $\alpha=3$ at $G_d = 12$. For $\alpha=17$, we see that messages that are only one packet long have a significantly higher probability of being received correctly, whereas longer packets have little chance of making it. This is mostly due to the fact that the average TP length is only $\overline{\tau}=2.18$ slots long for $\alpha=17$ (table 3.3), and thus not many packets in a message get a chance to transmit completely. For the system with $\alpha=3$ however, we see that the distribution of message successes is more evenly spread, providing a significantly fairer opportunity for all messages to succeed. For $\alpha=3$, the mean TP length is significantly larger ($\overline{\tau}=36.39$).

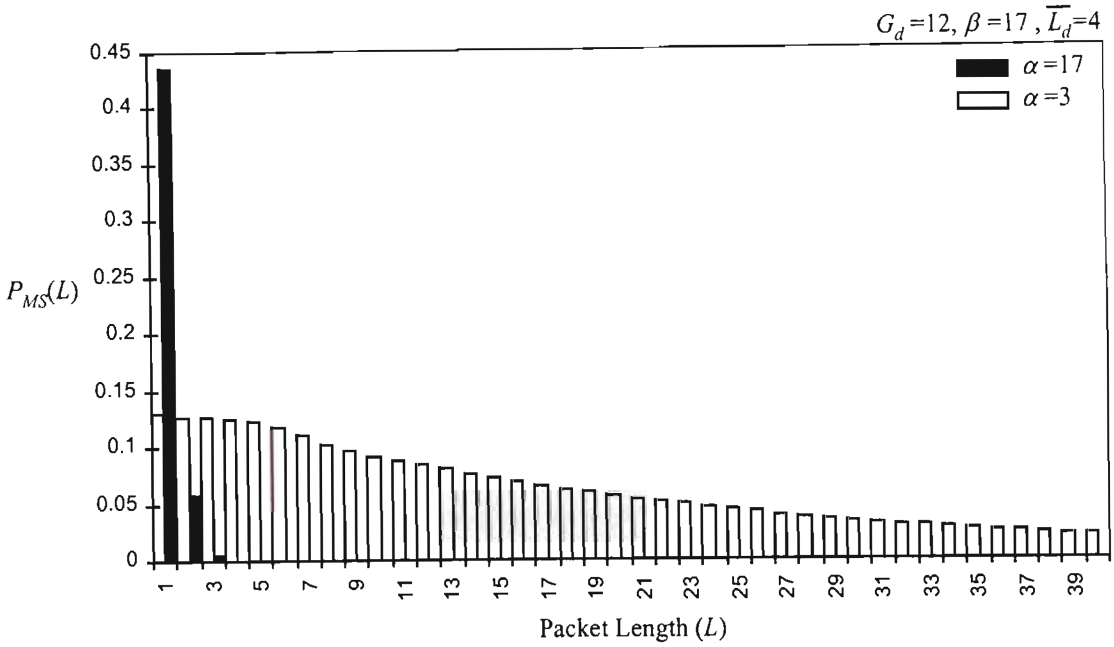


Figure 3.10 : Distribution of message success probabilities for α=17 and α=3 at G_d=12

When the distributions in figure 3.10 are averaged with respect to the number of packets per message using equation 3-28, a much larger packet throughput is obtained for α=3. When the distributions are averaged with respect to messages using equation 3-27 however, the same result for the expected probability of message success is obtained as observed in the last column of table 3.3. The reason why $P_{MS}(L=1)$ for α=3 is much smaller than the value obtained for α=17, as evident from figure 3.10, is due to the fact that there is message blocking for α=3 (P_{MB} =0.80 which implies that only 20% of messages are admitted).

From table 3.3 we can see that, based on the optimum packet throughput and message delay values, α=4 provides the best results for load value G_d=12. If we repeat this process for the entire range of load values, we expect to find that different α values will provide optimum results for different loads (as suggested in figure 3.8). If we associate each optimum α value with each corresponding load value, then we obtain the optimum packet throughput curve illustrated in figure 3.11. We also point out that the values of α which give optimum packet throughputs also give the optimum delay values, as seen in table 3.3 for G_d=12. This was seen to be true for all offered load values.

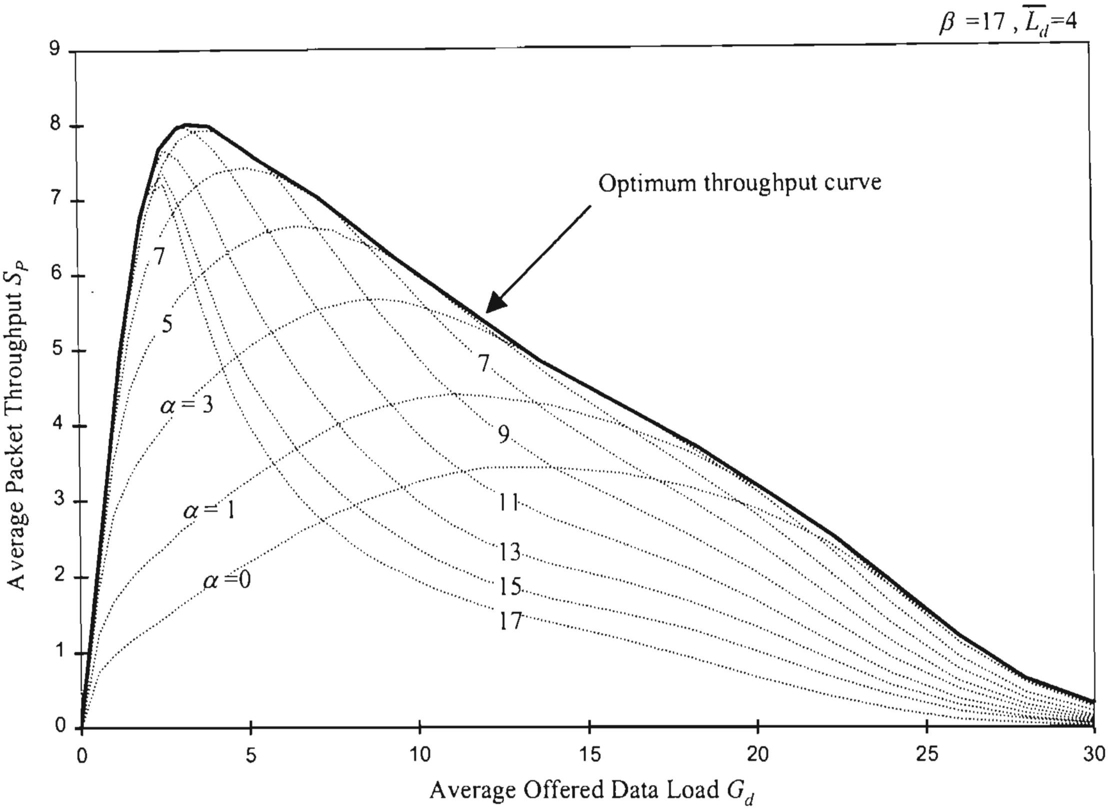


Figure 3.11 : Optimum packet throughput curve obtained by using a dynamic α threshold

A useful way to envisage the optimum throughput curve (the solid line in figure 3.11) is as an "envelope" which maps the peak throughput obtained after all possible α values (dotted lines) have been considered (we plot odd values of α for clarity purposes only).

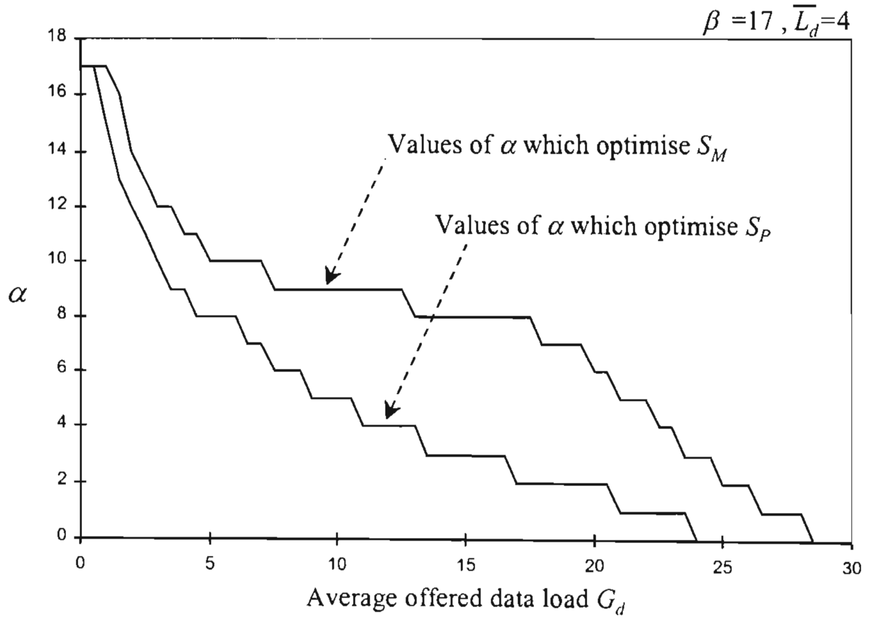


Figure 3.12 : Relationship between optimum α value and G_d for $\overline{L_d}$=4

Figure 3.12 provides the values of α which provide the optimum packet throughput for a given load value. Application of such a dynamic α threshold is not difficult to imagine. The values in figure 3.12 can be stored in a lookup table at the base station. Periodically, the base station can consider the average data load measured over a certain time period and adjust the value of α accordingly.

If, say, we decide to vary α in a manner which optimises the average *message throughput* instead of the packet throughput, then the optimum S_M curve in figure 3.13a is obtained. We also include the S_M curve which is obtained if we optimise the system packet throughput S_P. In both cases, the expected message throughput is better than for the case of no blocking threshold α(dotted line). The values of α which optimise S_M are also shown in figure 3.12. The stepwise nature of the curves are obviously due to the fact that α can only be an integer.

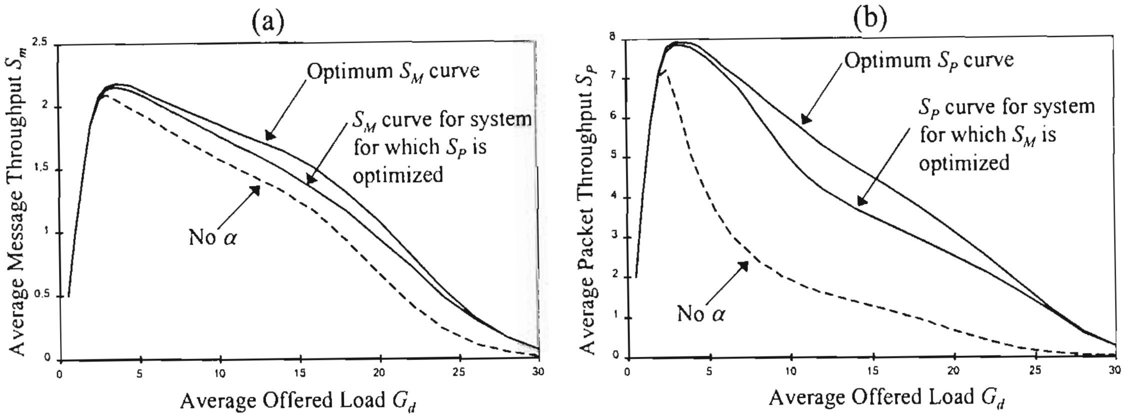


Figure 3.13 : Optimum message throughput curve obtained by using a dynamic α threshold

In figure 3.13b, we illustrate the drop in average packet throughput S_P that occurs if we choose to optimise the system for S_M rather than S_P. As mentioned previously, it is most beneficial to optimise the system based on the channel packet throughput, although the point is that, for all intents and purposes, α can be chosen to optimise any one particular performance measure. It is possible as well to choose α in a manner which optimises neither S_M nor S_P, but which lies somewhere between the two values. This will result in a curve which lies between both pairs of solid lines in figures 3.13a,b.

3.5.4 Effect of α on the average message delay

In equation 3-38 we compute the expected message delay. In the context of an infinite data population, it very difficult to obtain results for the average message delay from a simulator, due to the fact that an infinite table of timing information has to be kept in order to store the delay suffered by each user. Figure 3.14 provides the expected message delay versus the offered load G_d for various values of α. Unfortunately, these curves are a lot less representative of the system performance for the Poisson traffic model, since they can be adjusted to be anything by changing the average backoff delay ($\overline{B}$) without altering the other performance measures such as throughput and blocking probability. The throughput and delay results are thus completely independent of each other, which, of course, is not true in a real world nework. This is one of the shortfalls of the infinite population model.

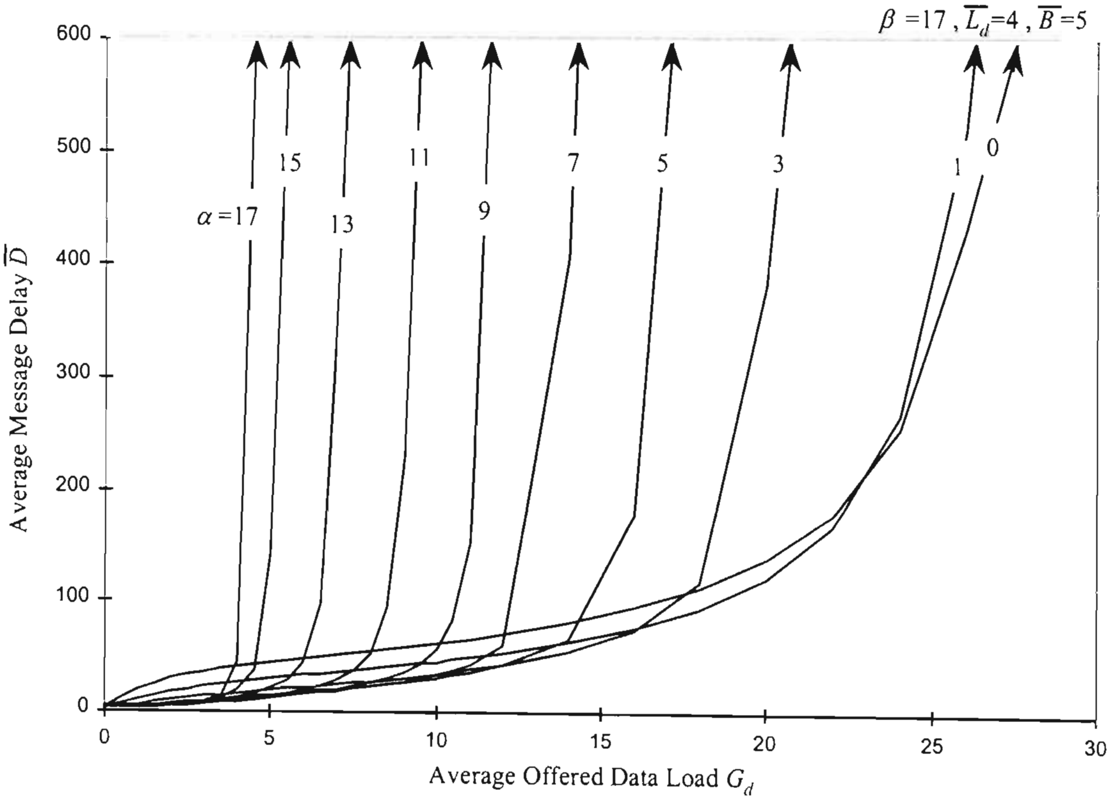


Figure 3.14 : Average message delay vs. offered load for various values of α

From figure 3.14, we can see that at low load values, higher values of α provide lower delays. This is in correspondence to the higher throughputs obtained at lower loads for higher values of α. At high loads, low values of α provide significantly lower delays. An example of this is given in table 3.3 for G_d=12. For α=17, we can see that the expected delay rapidly approaches infinity at very low loads, whereas for α=3 the convergence is much more gradual and occurs at far greater loads. Without presenting graphical results, we merely state that different values of $\overline{B}$ provided expected delay results which where similar in shape to the curves in figure 3.15, but scaled on an almost linear relationship. This is to be expected since, in equation 3-37, we see that the expected message delay is linearly related to $\overline{B}$.

3.5.5 Effect of varying the mean message length

We repeat the experimental procedure outlined in section 3.5.2 for various different mean message length values, namely $\overline{L}_d$=2, 7 and 10, and find the following thresholds to be optimum for each different mean message length:

$$\beta=18 \text{ for } \overline{L}_d=2$$

$$\beta=16 \text{ for } \overline{L}_d=7$$

$$\beta=16 \text{ for } \overline{L}_d=10$$

It is interesting to note that the optimum threshold value β is dependent on the *mean length of the data messages* (as hinted in section 3.5.2). Repeating the procedure in section 3.5.3 for each $\overline{L}_d$ yields the optimum packet throughput curves shown in figure 3.15. We include our previously computed case of $\overline{L}_d$=4 for comparison. This proves that, for a data system in which retransmissions are supported, the multiple access threshold of the network should be based on a value that provides the optimum system throughput. Since the mean message length has an influence on the system performance, it will become a factor in determining β. This is due mainly to the fact that the system is more efficient for short messages (figure 3.10), hence allowing a higher threshold. We also highlight the curve representing the network with no α threshold (dashed line), to show the improvements gained by employing a dynamic threshold. It is interesting to

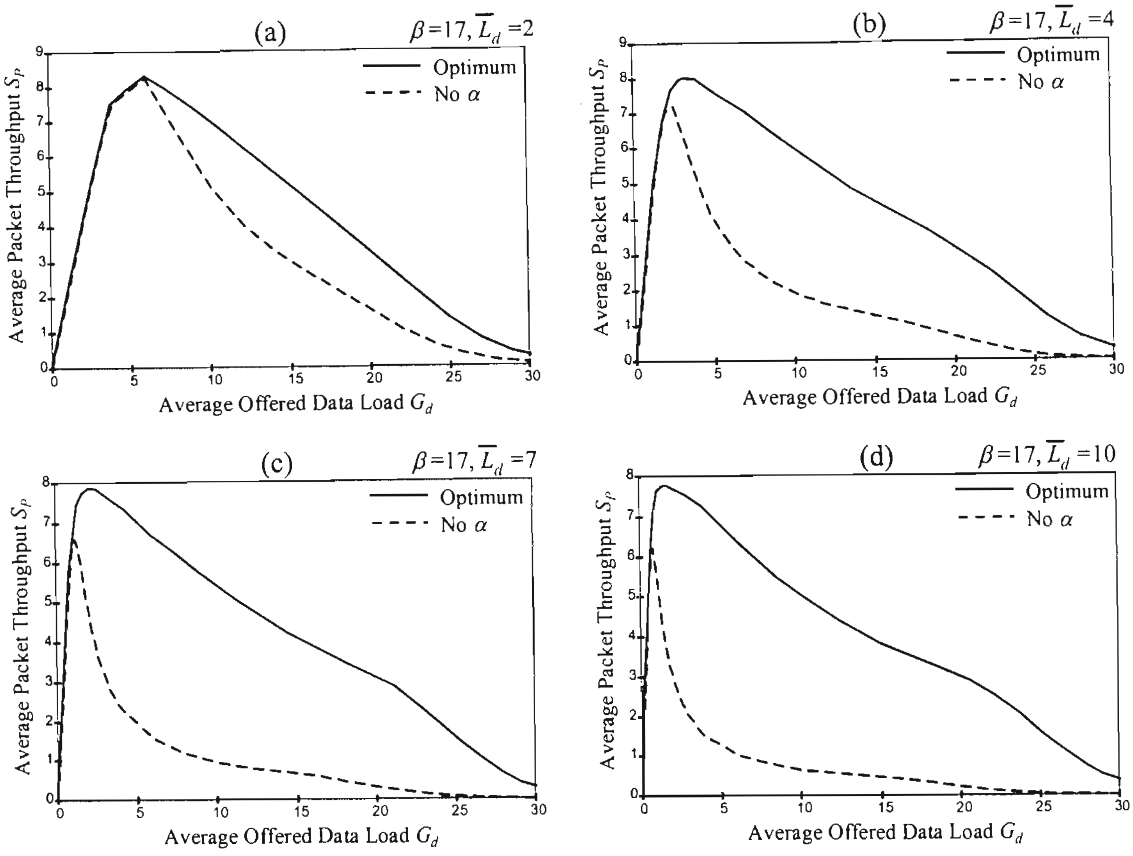


Figure 3.15 : Effect of different mean message lengths on the system performance

note that the degree of the throughput improvement increases as the mean message length increases. In figure 3.15a, we see that for $\overline{L}_d$=2, the optimum throughput curve is closest to the network with no α threshold, whereas figure 3.15d shows a much larger degree of improvement for $\overline{L}_d$=10. The reason for this is, again, best explained using figure 3.10. When the α threshold is applied at higher loads, the average probability of success of longer messages increases, thus giving all messages a more equal opportunity at success. When averaged over all message lengths, the longer messages contribute significantly to the average throughput. After applying the dynamic α threshold, we see that the optimum curves for all cases are very similar in shape and value, whereas for the case of no α, it is clear that the network with the shorter mean message length is more efficient. The latter point is to be expected since the longer the mean message length, the longer the message service time and hence, for a fixed message arrival rate G_d, the quicker it takes for the system to reach the collision threshold. The objective of the α threshold is to throttle the admission of messages into the network such that β is not breached, i.e. α does not discriminate against which message lengths it blocks. All messages are blocked with equal probability, hence regulating the channel load at a safe operating level below β and hence allowing those messages that have already been admitted a much greater probability of success.

3.5.6 Comparison of DT/CDMA with other random access CDMA models

In figure 3.16 we graphically compare the throughput of DT/CDMA, plotted as curve (D), with other common CDMA models such as:

(A) Conventional slotted-SS/ALOHA: In this scheme, there are no restrictions on the maximum number of transmitting users.

(B) Slotted-CDMA protocol that blocks new messages when the number of transmissions exceeds the maximum threshold β (e.g. [Toshimitsu *et al*, 1994] and [Abdelmonem & Saadawi, 1989]).

(C) Slotted-CDMA with Collision Detection (CD) that aborts all messages when the channel load exceeds the maximum threshold β (e.g. [Resheff & Rubin, 1990]).

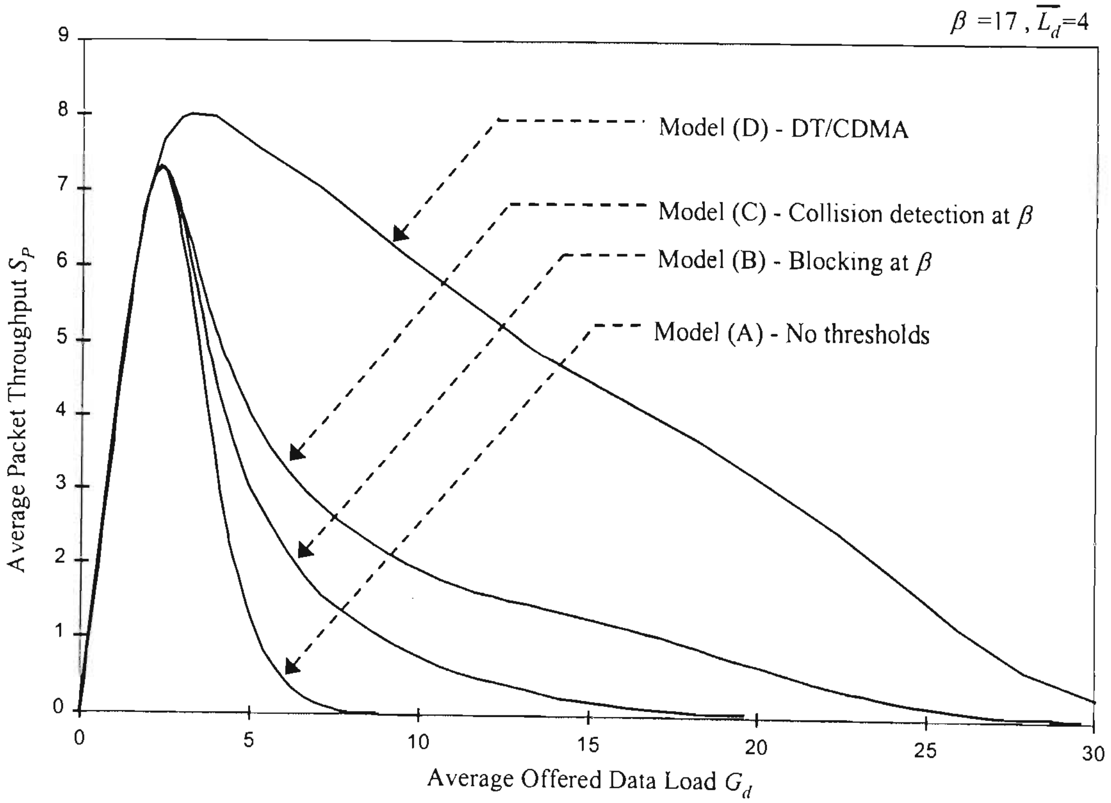


Figure 3.16 : Comparison of proposed data model with other common CDMA models

All curves in figure 3.16 are obtained for the same set of network parameters, as outlined in table 3.1. As can be seen, the conventional S-SS/ALOHA protocol (Curve A) with no channel load regulation is the most inefficient at high loads. If blocking is performed (Curve B) such that users may only transmit when the channel load is below β, then the system performance is improved at higher loads. If the collision detection scheme (Curve C) is employed instead of message blocking, then the system performance is further improved at higher loads. When a combination of both collision detection and blocking are employed, with a fixed collision threshold β and dynamic, load dependent blocking threshold α, then the optimum throughput curve obtained from our DT/CDMA protocol (Curve D) is seen to be far superior to the existing models.

3.5.7 Downfalls of the proposed data traffic model and DT/CDMA

In this section, we outline some of the downfalls associated with the proposed MAC policy and data traffic model used. The main shortcoming of the protocol is that it, as with other similar random access protocols proposed in the literature, does not treat all message lengths fairly. Figure 3.10 clearly represents the fact that shorter messages have a higher probability of success than longer messages. In the context of the retransmission policy and combined message arrival process, this has several implications that invalidate some of the assumptions made at the outset of the data model. Firstly, we assume that the mean retransmission delay (backoff time) is fixed for all load values. In an infinite population model, we assume that all retransmitted messages are absorbed into the Poisson arrival process. This is not strictly true however, since the fraction of terminals in the backlog state will have a significant impact on the arrival process in a real finite population. Secondly, we assume that the distribution of data messages is always geometric with a fixed mean length that is independent of the offered load.

In a real finite population, the computation of the distribution for the length of messages in the system is a multi-dimensional problem that requires the solution of a large number of simultaneous equations. The fact that entire messages are retransmitted, coupled with the fact that the probability of a message's success is inversely proportional to its length, means that longer messages will be backlogged with a higher probability than shorter messages. The distribution of retransmitted messages (from backlogged terminals) is thus expected to have a higher mean than the distribution of new first time messages (from idle terminals). The *combined* message length distribution from idle and backlogged users will thus have a higher mean than the distribution from idle users. Although the distribution for the length of first time messages will always be geometric with the same, non-varying mean, the distribution for the length of retransmitted messages is considerably more complex to model. For the purposes of this discussion, we define $P_{RML}(l)$ as the distribution of the length of retransmitted messages. In fact, it is not difficult to see that $P_{RML}(l)$ does not even closely resemble a geometric distribution (a memoryless Bernoulli trial in each slot). Instead, $P_{RML}(l)$ is more likely to be a more complex "peaked" distribution. The reasons for this, and why $P_{RML}(l)$ is more complex, is due to the fact that it is a function of many variables, including the distribution for new messages ($P_{ML}(l)$ given by equation 3-3), the distribution for the success of a message based on its length ($P_{MS}(l)$ given by equation 3-24), the number of retransmissions that a message of length l is expected to require ($1/P_{MS}(l)$), and the number of terminals that are in the backlogged state with a message of length l. Shorter messages occur with higher probability than longer messages, but are more successful. Very long messages are very rare, but are also almost never successful. Also, since longer messages generally require more retransmissions than shorter, they tend to remain in the backlog state for longer and hence increase the probability of retransmitting a message of that length. If we define $b(l)$ as the number of terminals in the backlogged state with a message of length l to retransmit, then we obviously expect to see an increase in $b(l)$ as l increases, although the fact that very long messages are very rare, we might expect to see the distribution tail off towards zero for very large l, resulting in a peak in the distribution for both $b(l)$ and $P_{RML}(l)$. The steady state distribution of $b(l)$ is thus an important parameter of $P_{RML}(l)$. As mentioned in section 2.4.2, population drift and system instability are important concepts of random access networks. The proposed protocol is no exception and is indeed significantly susceptible to instability due to the fact that the probability of message success is inversely proportional to the message length. We expect a "snowball effect" in which the backlogged population retransmits longer messages, these retransmissions interfere with new messages and hence decrease the overall throughput, and more users drift into the backlogged state. Inevitably, all data users end up in the backlogged state with long messages to transmit, the probability of success being too low to ensure that the system can return to the preferred stable state where most users are idle.

In the following chapter, we propose to improve the performance of the DT/CDMA protocol by implementing a *selective-repeat*, or *partial message* retransmission scheme. This modification, apart from improving the performance of DT/CDMA significantly, will also enable us to easily model the distribution for $P_{RML}(l)$ and hence allow for a more realistic data model with a finite population.

3.6 Summary

This chapter in its entirety dealt with our proposed MAC protocol for the reverse link of a CDMA data packet radio network, which we call "Dual-Threshold CDMA". We began by presenting the imagined network architecture and channel assumptions, followed by a description of the DT/CDMA MAC protocol.

The second part of this chapter was devoted to presenting the derivation of an accurate statistical analysis, using a Markov modelling approach, of the performance of DT/CDMA. In this model, we considered an infinite population of identical data terminals and modelled the arrival process from data users as a Poisson process. The length of data messages was assumed to conform to a geometric distribution. Although we modelled the performance of a single cell only, we can also assume that any handoff traffic is incorporated into the Poisson arrival process. In particular, we derived expressions to predict the expected data state distribution, the expected data message and packet throughputs and the expected message delay measurements. We also derived expressions for predicting the mean data message blocking probability.

The third part of this chapter presented results for the proposed model as obtained from the derived Markov model and a custom built simulator for an arbitrary set of network parameters. In all measurements, results from the Markov model and simulator corresponded well, thus validating our analytical approach. The reason for this high degree of correlation is not surprising, since we have basically presented an exact statistical analysis of the simulator algorithm. In the simulator, a Poisson random number generator was used to determine the number of arriving messages per slot. To determine whether or not a message was completed or not, a simple Bernoulli trial process was used. Since these two processes constitute the derivation of the transition matrix $\boldsymbol{\pi}$ in equation 3-6, and this matrix is the fundamental building block of the entire analysis, it is not difficult to see that if these two processes are matched in both simulator and analysis, an exact statistical analysis should result.

We firstly showed that significant improvements can be made over the standard S-SS/ALOHA protocol, in which no thresholds or access restrictions apply, by implementing a channel overload (collision) detection scheme that aborts messages when the user load exceeds the multiple access overload threshold β. We then showed that further performance improvements can be achieved by implementing a second dynamic message blocking threshold, α, which is related to the average message arrival rate. The α threshold is shown to throttle the admission of messages into the system and hence avoid frequent overload conditions that occur at high loads. We showed that our DT/CDMA MAC policy, which implements both an overload detection threshold (β), as well as another message blocking threshold ($\alpha<\beta$), is far superior to existing random access ALOHA based CDMA protocols that employ neither blocking nor collision detection, or either blocking only or collision detection only. In particular, it was shown that the selection of α and β should be made such that the overall system performance is optimised.

A downfall of the proposed DT/CDMA policy is that shorter messages have a greater probability of success than longer messages. This results in various assumptions that were made at the outset of the model (such as Poisson arrival rate and geometric distribution of message lengths with a mean that is independent of the offered load value) being strictly untrue. In Chapter 4, we modify the DT/CDMA access and retransmission policy such that it favours all message lengths more fairly. We also consider a more realistic traffic model that considers a finite population and allows us to examine the statistics of the traffic generated by idle terminals as well as the traffic generated by terminals that are in the backlogged state.

CHAPTER 4

SELECTIVE-REPEAT DUAL-THRESHOLD CDMA (SR-DT/CDMA) MAC PROTOCOL

4.1 Introduction

In this chapter, we propose an improvement to our original DT/CDMA MAC protocol presented in the previous chapter. The modification is based on applying a selective-repeat, or partial message retransmission scheme, as opposed to the retransmission scheme in the previous chapter in which the entire message was retransmitted regardless of the number of corrupt packets in the message. For the purposes of comparison, we will refer to this modified DT/CDMA protocol as “Selective-Repeat Dual-Threshold CDMA”, or SR-DT/CDMA. The advantages of applying a selective-repeat retransmission scheme are two-fold: (1) The performance of the MAC protocol is improved significantly, (2) The performance analysis for a finite population model becomes computationally tractable. The modelling of a finite population allows us to consider a more realistic traffic arrival process, and also allows us to accurately model and evaluate various aspects of the protocol performance that were not possible for the infinite population Poisson model, such as the expected message delay and stability issues of the data network.

4.2 Description of the SR-DT/CDMA MAC Protocol

The network architecture, channel assumptions and signalling assumptions are as described in section 3.2. The dual multiple access threshold model and base station broadcast tones, as described in sections 3.3.1 and 3.3.2, also apply in exactly the same manner. SR-DT/CDMA improves on the original DT/CDMA MAC protocol by modifying steps 3, 4 and 5 of the MAC procedure in section 3.3.3 such that only the corrupt packets in a data message are retransmitted (refer to figure 4.1 for an illustration of SR-DT/CDMA). The SR-DT/CDMA MAC procedure is then described as follows:

1. Exactly as described in (1) of section 3.3.3
2. Exactly as described in (2) of section 3.3.3
3. Step 3 is now different in that upon detecting the **DC** tone, ***only the contending users abort transmission*** and proceed to the backlogged routine (6). Messages that were *admitted* prior to the collision, continue transmission throughout and after the collision.
4. As a result of the change in (3), once a user becomes *admitted*, the channel is "secured" until message completion.
5. As described in section 3.3.3, terminals are notified whether a message was successfully received or not. Unlike the DT/CDMA model however, ***only the corrupt packets*** in the message are retransmitted. This retransmission scheme is known as *selective-repeat* or *selective-reject* [Raychaudhuri, 1987]. These packets which require retransmission are grouped together to form a new message. It is assumed that the ARQ feedback scheme (which, for selective repeat retransmission is often referred to as SRQ) reveals which packets require retransmission. It is also assumed that the base station is capable of reconstructing the original message, in order, once all packets have been received correctly.
6. The backoff delay/retransmission procedure is the same, as described in (6) of section 3.3.3.

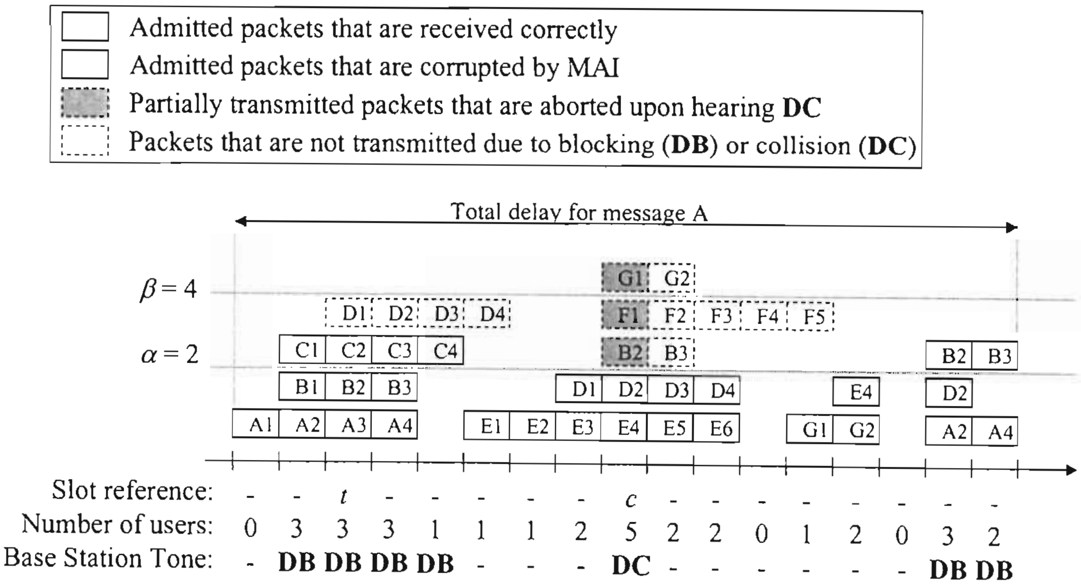

Admitted packets that are received correctly
Admitted packets that are corrupted by MAI
Partially transmitted packets that are aborted upon hearing DC
Packets that are not transmitted due to blocking (DB) or collision (DC)
Total delay for message A
$\beta = 4$
$\alpha = 2$
G1 G2
D1 D2 D3 D4
F1 F2 F3 F4 F5
C1 C2 C3 C4
B2 B3
B2 B3
B1 B2 B3
D1 D2 D3 D4
E4
D2
A1 A2 A3 A4
E1 E2 E3 E4 E5 E6
G1 G2
A2 A4
Slot reference: - - t - - - - - c - - - - - - - -
Number of users: 0 3 3 3 1 1 1 2 5 2 2 0 1 2 0 3 2
Base Station Tone: - DB DB DB DB - - - DC - - - - - - DB DB

Figure 4.1 : Illustration of the SR-DT/CDMA MAC policy

Figure 4.1 illustrates the admission and retransmission policy of SR-DT/CDMA. In this example, attempting message D is blocked because the state of the network in the previous slot exceeded α. Packets A2 and A4 of message A (a message containg four packets) are corrupt due to MAI and are retransmitted as a backlogged message (containing just the two packets requiring retransmission) sometime later where, at the second attempt, they are successful. The total message delay for message A is illustrated on figure 4.1, as defined in section 3.4.12.

Attempting messages B, F and G arrive and, together with previously admitted message E and the second attempt of message D, cause a collision. Only the three new messages (B, F and G) abort, while the other admitted messages E and D that were transmitting prior to the collision continue doing so. The reason for only the new messages aborting is based on the improved selective-repeat packet retransmission scheme. During the collision slot, it is assumed that all packets become corrupt with a high probability. If the attempting messages that arrive to cause the collision continue transmitting after the collision slot, they would continue to corrupt all the previously admitted users.

In DT/CDMA, we assumed that, since the packets involved in the collision were corrupt with a high probability, continued transmission of all collided messages was wasteful since the entire message would only have to be transmitted again. Since only the packets involved in the collision are now required to be retransmitted in SR-DT/CDMA, it makes sense to allow the admitted messages (those that were transmitting before the collision) to continue transmitting. After the collision, the attempting messages immediately abort and the state drops to below β to allow the remaining packets in the admitted messages to be successful.

4.3 Finite Population Traffic Model Issues and Assumptions

In this section we discuss the issues and assumptions we will be making regarding the traffic model we will use for the analytical performance modelling of SR-DT/CDMA.

4.3.1 Solution to the problem of finding $P_{RML}(l)$

In section 3.5.7 we discussed the problems associated with analysing a finite population model. In particular, the problem of finding a distribution for the length of retransmitted messages, $P_{RML}(l)$, was found to be very complicated and multi-dimensional. Fortunately, the modified selective-repeat retransmission scheme simplifies this problem significantly and allows us to make a very reasonable and accurate approximation for $P_{RML}(l)$, namely: ***that $P_{RML}(l)$ appears to closely resemble a geometric distribution***. The main reason for the significant change in the distribution of $P_{RML}(l)$ is as follows. Each time a message is unsuccessful, the fact that only the corrupt packets in the message are retransmitted means that unsuccessful messages are continually broken down into successively smaller "sub-messages". These sub-messages have, at each retransmission, a higher probability of success than the previous retransmission (unless of course all packets in the message are corrupt). The problem discussed in section 3.5.7 of having a complex distribution for $P_{RML}(l)$ which contains mainly longer messages is now negated by the new retransmission policy, since the long messages in the distribution are replaced by shorted sub-messages.

Before validating and justifying the approximation that $P_{RML}(l)$ is geometric, we present the following population model. We consider a data network that has a finite population of data users ($N_d << \infty$). As in DT/CDMA, we assume that all data terminals are identical and that each terminal operates completely independently of all other terminals. The finite population is divided into two sub-populations:

- ***Idle*** terminals: Each *idle* terminal generates a new message in a slot with probability ρ_{d0}, where the subscript "0" is used to denote new messages. All new messages that are generated from idle terminals terminate with probability μ_{d0} at the end of the current slot. Again, this implies that new message lengths are geometrically distributed with mean length $\overline{L}_{d0} = \mu_{d0}^{-1}$, as in the previous chapter. This geometric distribution for ***N***ew ***M***essages ***L***engths is defined as $P_{NML}(l)$. We assume that only idle terminals are capable of generating new messages.

- ***Busy*** terminals: A busy terminal is basically a non-idle terminal which can further be defined to be in one of two possible states: *transmitting* or *backlogged*. *Transmitting* terminals are terminals that are in the process of (re)transmitting a message, while *backlogged* terminals are terminals that are waiting to retransmit a message. As mentioned in step 6 of section 4.2, *backlogged* terminals retransmit their backlogged message in the current slot with probability ρ_{dR}, where subscript "*R*" denotes a retransmitted message. This implies that the backoff time is geometrically distributed with mean $\overline{B} = \rho_{dR}^{-1}$. *Busy* terminals cannot generate new messages.

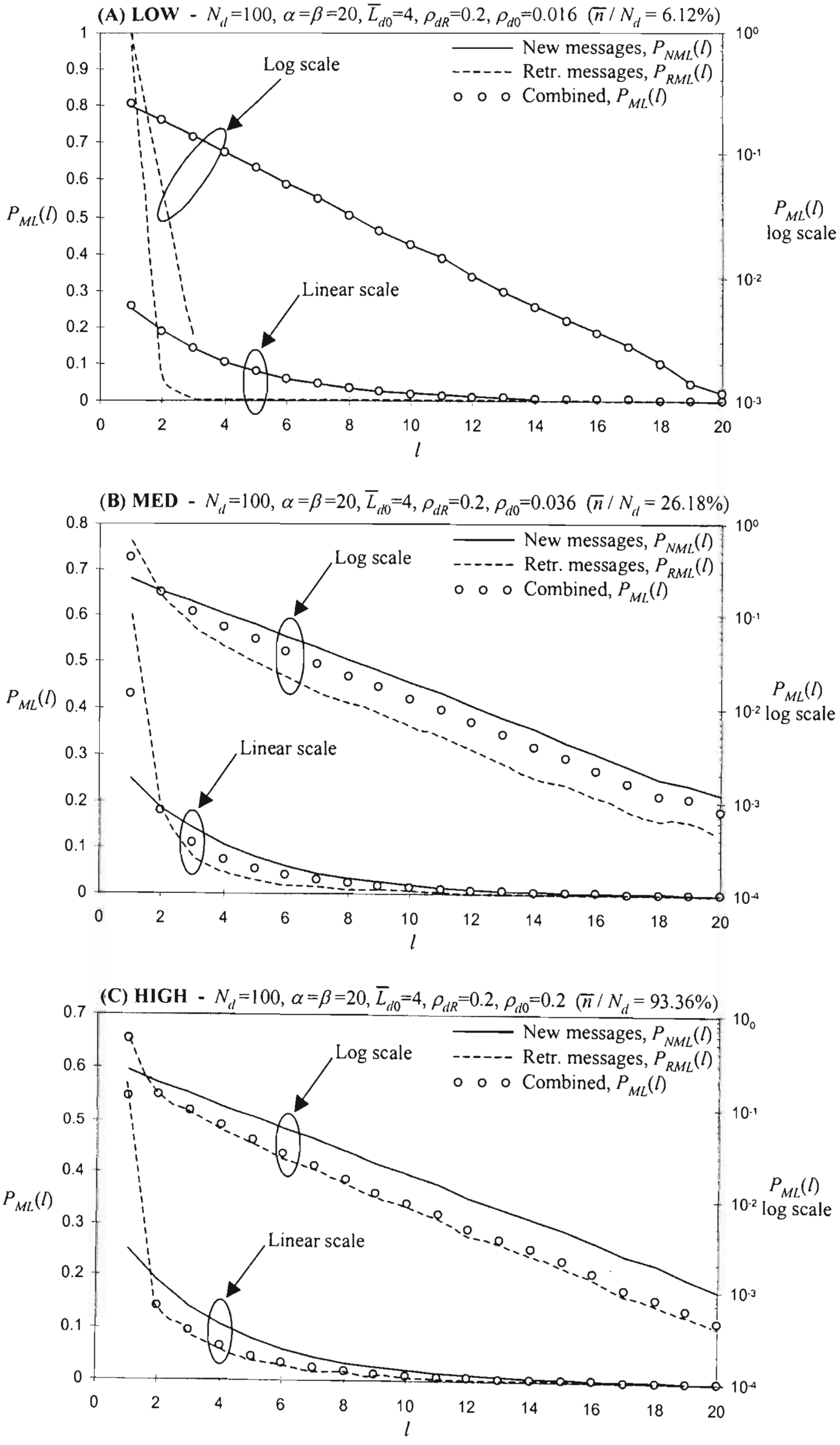
The following state variables are defined to represent the number of terminals in each sub-population:

x = instantaneous number of *transmitting* users
b = instantaneous number of *backlogged* users
n = instantaneous number of *busy* or non-idle users ($n = x + b$)
and hence $N_d - n$ = instantaneous number of *idle* users

We also define $\overline{x}$, $\overline{b}$ and $\overline{n}$ as the expected steady state sub-population sizes.

At this point, we formalise our assumption by stating that ***the distribution of retransmitted messages, $P_{RML}(l)$, is geometric, with mean length $\overline{L}_{dR}$***. This assumption has been verified and justified in many other works (e.g. [Raychaudhuri, 1987] and [Joseph & Raychaudhuri, 1993]). Rather than provide a mathematical justification for the assumption, we choose to justify it further through simulation results.

Figure 4.2 illustrates some simulation results, for a network with arbitrary parameters, of the distributions of new and retransmitted message lengths ($P_{NML}(l)$ and $P_{RML}(l)$ respectively) for the new SR-DT/CDMA policy proposed in section 4.2. We also include the overall message length distribution, $P_{ML}(l)$, which incorporates both new



(A) LOW - N_d=100, $\alpha=\beta$=20, $\overline{L}_{d0}$=4, ρ_{dR}=0.2, ρ_{d0}=0.016 ($\overline{n}$ / N_d = 6.12%)

(B) MED - N_d=100, $\alpha=\beta$=20, $\overline{L}_{d0}$=4, ρ_{dR}=0.2, ρ_{d0}=0.036 ($\overline{n}$ / N_d = 26.18%)

(C) HIGH - N_d=100, $\alpha=\beta$=20, $\overline{L}_{d0}$=4, ρ_{dR}=0.2, ρ_{d0}=0.2 ($\overline{n}$ / N_d = 93.36%)

Figure 4.2 : Simulation results of the SR-DT/CDMA protocol with arbitrary network parameters showing new, retransmitted and combined message length distributions.

and retransmitted messages. If the means of $P_{NML}(l)$ and $P_{RML}(l)$ are $\overline{L}_{d0} = \mu_{d0}^{-1}$ and $\overline{L}_{dR} = \mu_{dR}^{-1}$ respectively, then the mean of the *combined* distribution, defined as $\overline{L}_d$, lies somewhere between $\overline{L}_{d0}$ and $\overline{L}_{dR}$, i.e.

$$\overline{L}_{dR} \leq \overline{L}_d \leq \overline{L}_{d0} \qquad \textbf{(4-1)}$$

The reason for this is that the mean of $P_{RML}(l)$ is bounded in the following manner:

$$1 \leq \overline{L}_{dR} \leq \overline{L}_{d0} \qquad \textbf{(4-2)}$$

Firstly, it is impossible for any message length to be less than one packet and thus the mean can never be less than one. Since retransmitted messages consist of the subset of unsuccessful packets from previously corrupt messages and the subset of previous messages that were blocked, it is impossible for the mean length of retransmitted messages to be greater than that of new messages. In order to evaluate the geometric approximation effectively, we consider three different network scenarios in figure 4.2:

A. Figure 4.2a considers a very low offered data load, where the probability of a terminal generating a message is low ($\rho_{d0} = 0.016$). In this scenario, the average carried data load is very low and the packet success probability is almost 100 percent and, as a result, the proportion of terminals in the backlogged state is only 6.12 percent. Since most messages are thus being generated by idle terminals, we expect the overall distribution, $P_{ML}(l)$, to be almost exactly the same as $P_{NML}(l)$. In figure 4.2a, this is seen to be correct. Also, since the message success probability is very high at such a low load, we expect the mean length of retransmitted messages to be very low. Indeed, the curve of $P_{RML}(l)$ in figure 4.2a proves this to be true.

B. In figure 4.2b, we consider a medium offered load ($\rho_{d0} = 0.032$) which results in approximately 26 percent of the population being backlogged on average. In this example, it can be seen that the contribution of messages from the backlogged terminals is now comparably greater. Equations 4-1 and 4-2 are also justified, where it can be clearly seen that the mean of $P_{RML}(l)$ is less than that of $P_{NML}(l)$, and that the mean of $P_{ML}(l)$ lies between the two.

C. Figure 4.2c provides the opposite result to figure 4.2a. In this case, the offered load is set to be too high ($\rho_{d0} = 0.2$), resulting in a poor system throughput and the network becoming saturated with backlogged terminals (93.36 percent). As is expected, backlogged terminals are generating most of the messages.

On each graph, we also provide logarithmic plots of each distribution to show how close the distribution is to being geometric. A geometric distribution, when plotted on a logarithmic scale yields a straight line (a simple and convenient test). The gradient and offset of the line are indications of the mean of the distribution. As can be seen in all three scenarios, the distribution of $P_{NML}(l)$ is exactly geometric. This is to be expected since a geometric random number generator was used in the simulator for new message lengths. The shape and trend of the curves for $P_{RML}(l)$ and $P_{ML}(l)$ are seen to appear reasonably geometric in all three cases. When plotted on a log scale, the distribution yields a gentle curve for low l that becomes a straight line for larger l. The region of the curve where l is low reveals the part of the distribution that is most geometric in nature, while it is expected that the geometric assumption becomes more and more inaccurate with increasing l (the line for $P_{RML}(l)$ and $P_{NML}(l)$ should continue away from that of $P_{NML}(l)$ for a geometric distribution, instead of tending to parallel it as seen from the simulation). Nevertheless, the geometric assumption is still reasonable for the more frequent shorter message lengths.

4.3.2 Computation of the data message arrival process

It is obvious from figure 4.2 that the distributions $P_{RML}(l)$ and $P_{ML}(l)$ are related to the steady state number of terminals in the non-idle state, $\overline{n}$. In this thesis, we assume that the mean backoff time is always the same, regardless of the network state. As a result, whereas ρ_{d0} is varied to present different offered loads to the network, ρ_{dR} remains constant for all network conditions. That is, we do not consider any of the variable retransmission delay schemes discussed in section 2.4.2. We choose to do this since the performance enhancements gained through the use of such schemes are well known and well documented in the literature. A preliminary analysis revealed that our proposed protocol was no exception when it came to such improvements, and the results showed nothing ground-breaking enough to warrant a detailed analysis. Rather, we find it more interesting to maintain a constant backoff time, and analyse the performance enhancements gained through the use of our dual threshold model with variable blocking threshold, α.

We can generate an expression for the combined message arrival process (from both idle terminals and backlogged terminals) as follows. In knowing that there are exactly b backlogged terminals, x instantaneously transmitting terminals, and hence $N_d - b - x$ idle terminals, and that each idle terminal generates a message with probability ρ_{d0} and that each backlogged terminal retransmits its message with probability ρ_{dR}, we can define the combined message arrival process as

$$\begin{aligned} f_{N_d}^{b,x}(a) &= P\{Arrivals = a | backlog = b, transmitting = x\} \\ &= \sum_{k=0}^{(a,b)^-} b(k, b, \rho_{dR}) \cdot b(a-k, N_d - b - x, \rho_{d0}) \end{aligned} \qquad \textbf{(4-3)}$$

where $f_{N_d}^{b,x}(a)$ represents the probability of a messages arriving in a slot given the instantaneous population states b, x and $n = b + x$ during that slot (N_d is a constant).

4.3.3 Computation of the combined message length distribution $P_{ML}(l)$

The computation of the expected message length is slightly more difficult to compute. Since it is an expectation, we require the expected values of the population sizes, as well as the expected length of retransmitted messages. To summarise, we present in table 4.1 which variables are known and which are unknown at the outset of the analysis:

Known	Unknown
ρ_{d0} , μ_{d0}	
ρ_{dR}	μ_{dR}
$\overline{L}_{d0}$	$\overline{L}_{dR}$, $\overline{L}_{d}$
N_d	$\overline{n},\overline{x},\overline{b}$

Table 4.1: Table of known and unknown variables at the outset of the analysis

Since $\overline{L}_{dR} = \mu_{dR}^{-1}$ and we are required to compute $\overline{L}_d$ as a function of the remaining variables, the number of unknowns is reduced to four, namely: $\overline{L}_{dR}$, $\overline{n}$, $\overline{x}$ and $\overline{b}$. In order to simplify the impending analysis further, we make the following assumption: *The steady state expected number of busy terminals is almost equal to the steady state number of backlogged terminals*, i.e. $\overline{n} \approx \overline{b}$. This assumption is not so unrealistic when the population size N_d is large and the expected number of transmitting users, $\overline{x}$, is small. Recall that the MAC protocol attempts to limit $\overline{x}$ to α by implementing the blocking threshold and should thus be significantly smaller than the other network population sizes, which are bounded by N_d . The assumption will only be invalid at very low offered loads when the number of backlogged terminals is low. However, it will be seen later that this should not be too much of a problem. In making this assumption, the number of unknown variables has been reduced to two: namely, $\overline{L}_{dR}$ and $\overline{n}$.

Based on the pertinent known variables ($\overline{L}_{d0}, \rho_{d0}, \rho_{dR}$) and the two unknown variables ($\overline{L}_{dR}, \overline{n}$) we can now present an approximate expression for the expected length, $\overline{L}_d$, of a message in the network:

$$\overline{L}_d = \frac{1}{\mu_d} \approx \frac{\overline{L}_{d0}\rho_{d0}(N_d - n)}{\rho_{d0}(N_d - n) + \rho_{dR}n} + \frac{\overline{L}_{dR}\rho_{dR}n}{\rho_{d0}(N_d - n) + \rho_{dR}n}$$

$$= \frac{\overline{L}_{d0}G_{d0} + \overline{L}_{dR}G_{dR}}{G_d} \qquad \textbf{(4-4)}$$

where G_{d0} is the expected load generated by the idle population (number of new data messages per slot), G_{dR} is the expected load generated from the backlogged population (number of retransmitted messages per slot), and G_d is the overall expected offered message load:

$$G_{d0} = (N_d - \overline{n})\rho_{d0} \qquad \textbf{(4-5)}$$

$$G_{dR} = \overline{n}\rho_{dR} \qquad \textbf{(4-6)}$$

$$G_d = G_{d0} + G_{dR} \qquad \textbf{(4-7)}$$

Since we have assumed and proven that the distribution of $P_{ML}(l)$ is approximately geometric, the overall expected probability of a message terminating in the current slot is equal to $\mu_d = \overline{L}_d^{-1}$.

Having described an arrival process and message length distribution, we can now present an analysis for the network.

4.4 SR-DT/CDMA Protocol Performance Analysis

The analysis of the SR-DT/CDMA protocol for the proposed finite population is complicated by the fact that we have the two unknown variables $\overline{L}_{dR}$ and $\overline{n}$ at the outset of the analysis. Although we can obtain expressions for the system performance, they are functions of the two unknown variables, $\overline{L}_{dR}$ and $\overline{n}$. This presents a "circular reference" problem in that $\overline{L}_{dR}$ and $\overline{n}$ are required in order to compute the system performance, but are themselves a function of the system performance. Intuitively, (and also as shown in figure 4.2) we expect that $\overline{L}_{dR}$ and $\overline{n}$ are directly related to the system throughput. For low channel loads and high message/packet success probabilities, we expect both $\overline{L}_{dR}$ and $\overline{n}$ to be smaller than at higher loads where the probabilities of success are smaller.

4.4.1 Concept of the system information flow model

A common way of obtaining a solution to this "circular reference" problem is to apply the flow equilibrium model developed by Kleinrock & Lam [1975]. This model is based on developing flow equations for messages and packets into and out of the system for different values of $\overline{n}$ and $\overline{L}_{dR}$. When the network is at a steady state equilibrium point, it is a proven and well known fact that [Kleinrock & Lam, 1975], [Raychaudhuri, 1987]

> ***... the rate of flow of packets into the network equals the rate of flow of packets out of the network and, simultaneously, the rate of flow of messages into the network equals the rate of flow of messages out of the network.***

Application of this flow model results in what is commonly referred to as an "Equilibrium-Point Analysis" (EPA) of the network. In EPA, the assumption made is that, at steady state conditions, the system will stay near the states where the amount of information being generated will balance with the amount of information being successfully transmitted. It has been shown that this state of information flow equilibrium occurs only at very specific pair combinations of $\overline{n}$ and $\overline{L}_{dR}$ [Kleinrock & Lam, 1975], [Raychaudhuri, 1987].

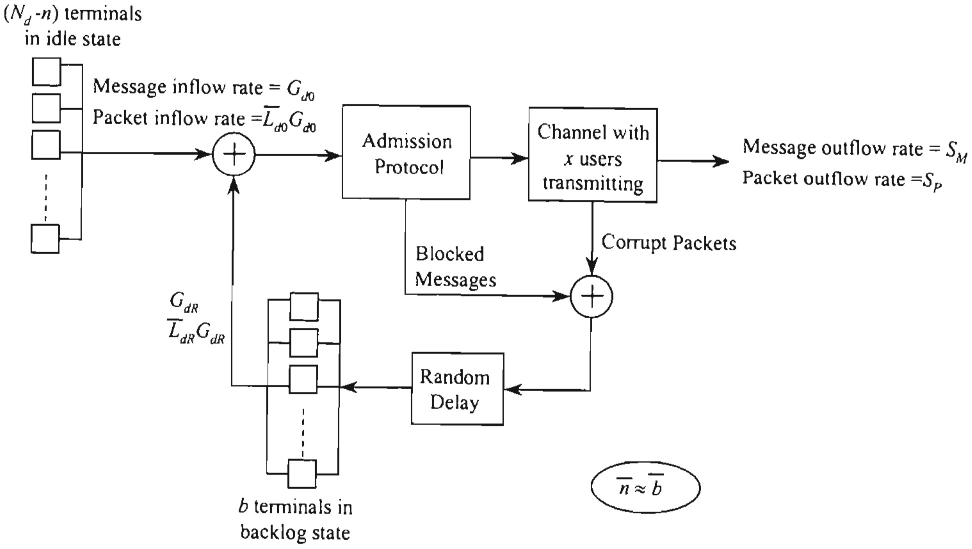

$(N_d\text{-}n)$ terminals in idle state
Message inflow rate = G_{d0}
Packet inflow rate =$\overline{L}_{d0}G_{d0}$
Admission Protocol
Channel with x users transmitting
Message outflow rate = S_M
Packet outflow rate =S_P
Corrupt Packets
Blocked Messages
G_{dR}
$\overline{L}_{dR}G_{dR}$
Random Delay
b terminals in backlog state
$\overline{n} \approx \overline{b}$

Figure 4.3 : Flow diagram for messages and packets in the network [Raychaudhuri, 1987]

Figure 4.3 illustrates the flow diagram for messages and packets into and out of the network, along with the different population subsets which give rise to the different flow components. Recall that S_M and S_P in figure 4.3 are the expected message and packet throughputs respectively.

Expressions for the input flow rates are simple to obtain. The message input flow rate is simply equal to the offered new message load, i.e.

$$\begin{aligned}\textbf{\textit{Message input flow rate}} &= G_{d0} \\ &= (N_d - \overline{n})\rho_{d0}\end{aligned} \qquad \textbf{(4-8)}$$

The packet input flow rate is simply equal to the message input flow rate multiplied by the expected number of packets per new message, i.e.

$$\begin{aligned}\textbf{\textit{Packet input flow rate}} &= G_{d0}\overline{L}_{d0} \\ &= (N_d - \overline{n})\rho_{d0}\overline{L}_{d0}\end{aligned} \qquad \textbf{(4-9)}$$

In order to obtain expressions for the output flow rates, we require expressions for the message and packet throughputs. The procedure for doing this is similar to that outlined in section 3.4.

4.4.2 Computation of the data system state vector

As in section 3.4.3, state variable x is defined as the number of transmitting data users, and $\hat{x}$ is the state as seen by the reference user (excluding itself). Because of the modified MAC policy, the state transition probability matrix for SR-DT/CDMA will be different. Apart from the obvious fact that the arrival process is now represented by $f_{N_d}^{b,x}(.)$ instead of $f_{\infty}(.)$, transition probability π_{ij} for SR-DT/CDMA differs from equation 3-6 (for DT/CDMA) in that when $i \le \alpha$ and $j \le i$ there also exists the possibility that $i - j$ users terminate normally with probability μ_d and that a large number (dummy variable $m > \beta - j$) of new users arrive, cause a collision, and are immediately dropped such that the load returns to j. The state transition probability π_{ij}, as defined in section 3.4.4, for SR-DT/CDMA is thus given by:

$$\pi_{ij} = \begin{cases} \sum_{k=0}^{i} f_{N_d}^{b,i}(j-i+k).b(k,i,\mu_d) & i \le \alpha, j > i \\ \sum_{k=i-j}^{i} f_{N_d}^{b,i}(j-i+k).b(k,i,\mu_d) & \\ \qquad + b(i-j,i,\mu_d) \cdot \sum_{m=\beta+1-j}^{N_d-j} f_{N_d}^{b,i}(m) & i \le \alpha, j \le i \\ b(i-j,i,\mu_d) & i > \alpha, j \le i \\ 0 & i > \alpha, j > i \end{cases} \qquad \textbf{(4-10)}$$

The new protocol does not abort all data messages upon detecting a collision and hence the data state does not return to a known value of x=0. Because of this, we abandon the concepts of a transmission period (TP) and renewal process that we used in the analysis of DT/CDMA. Computation of the state occupancy probability distribution for x, $\hat{X}(j) = P\{\hat{x} = j\}$ as defined in section 3.4.4, then becomes far simpler to obtain than the recursive procedure used in equations 3-13 to 3-16. Since x is no longer represented by a regenerative stochastic process, we can use steady state assumptions and compute the state occupancy probability vector $\hat{\mathbf{X}} = [\hat{X}(0), \hat{X}(1), \ldots.., \hat{X}(\beta)]$ by solving the following set of linear equations obtained from $\hat{\mathbf{X}} = \hat{\mathbf{X}}\boldsymbol{\pi}$

$$\hat{X}(j) = \sum_{i=0}^{\beta} \hat{X}(i)\pi_{ij} \qquad \textbf{(4-11)}$$

while maintaining the conservation relationship

$$\sum_{i=0}^{\beta} \hat{X}(i) = 1 \qquad \textbf{(4-12)}$$

4.4.3 Computation of the message blocking probability

Due to the fact that only contending users abort during a collision in SR-DT/CDMA, the probability of blocking messages is now slightly more complicated to compute. Again, we condition on the number of transmissions ($\hat{x} = j$) that the arriving reference message finds in its slot of arrival:

$$P_{MB}(j) = \begin{cases} \displaystyle\sum_{i=\alpha+1}^{\beta} \frac{\hat{X}(i) \cdot \pi_{ij}}{\hat{X}(j)} & j > \alpha \\ \displaystyle\sum_{i=\alpha+1}^{\beta} \frac{\hat{X}(i) \cdot \pi_{ij}}{\hat{X}(j)} + \sum_{i=j}^{\alpha} \left(\frac{\hat{X}(i) \cdot b(i-j, i, \mu_d)}{\hat{X}(j)} \cdot \sum_{m=\beta+1-j}^{N_d-j} f_{N_d}^{b,i}(m) \right) & j \leq \alpha \end{cases}$$

(4-13)

In equation 3-17, blocking only occurred when the state of the previous slot was larger than α. In SR-DT/CDMA, attempting messages are also considered to be blocked when a collision occurs, since only attempting messages abort.

If the number of users in the previous slot (dummy variable i) is larger than α, then all new messages are blocked by the **DB** tone, as in the DT/CDMA model.

When $i \leq \alpha$ and $j \leq i$, then we must consider the case that $i - j$ users terminate normally with probability μ_d and that a large number (dummy variable $m > \beta - j$) of new contending users arrive, cause a collision, and are immediately dropped such that the load returns to j. This probability is expressed in the second term of equation 4-13 for $j \leq \alpha$.

If $j > i$, then no blocking can occur since it implies an increase in the system state. The overall message blocking probability (unconditioned on j), P_{MB}, is found by substituting equation 4-13 into equation 3-18.

4.4.4 Computation of the message success probability

The procedure for computing the success of a reference message is the same as used in section 3.4.10. Again, $R_n(j)$ is the probability of success of the first n packets of a message, given that the n^{th} packet sees $\hat{x} = j$ other data users:

$$R_1(j) = \hat{X}(j).[1 - P_{MB}(j)].P_{PS}(j+1) \qquad j < \beta \qquad \textbf{(4-14)}$$

$$R_n(j) = \sum_{i=0}^{\beta-1} R_{n-1}(i) \cdot \hat{\pi}_{ij}^{PS} \qquad n > 1, j < \beta \qquad \textbf{(4-15)}$$

where $P_{PS}(j+1)$ is given in equation 3-19 (or equation 3-20 if FEC coding is used). As in equation 3-23, $\hat{\pi}_{ij}^{PS}$ represents the probability that the n^{th} reference packet is successfully received given that it sees the system state change from $\hat{x} = i$ to $\hat{x} = j$. Again, we incorporate the probability of the reference packet's success into $\hat{\boldsymbol{\pi}}^{PS}$ by including the probability of packet success term, $P_{PS}(\cdot)$:

$$\hat{\pi}_{ij}^{PS} = \begin{cases} \displaystyle\sum_{k=0}^{i} f_{N_d}^{b,i}(j-i+k).b(k,i,\mu_d).P_{PS}(j+1) & i \le \alpha - 1, j > i \\[2ex] \displaystyle\sum_{k=i-j}^{i} f_{N_d}^{b,i}(j-i+k).b(k,i,\mu_d).P_{PS}(j+1) & \\ \displaystyle\quad + b(i-j,i,\mu_d) \cdot \sum_{m=\beta-j}^{N_d-j-1} f_{N_d}^{b,i}(m).P_{PS}(j+m+1) & i \le \alpha - 1, j \le i \\[2ex] b(i-j,i,\mu_d).P_{PS}(j+1) & i > \alpha - 1, j \le i \\[2ex] 0 & i > \alpha - 1, j > i \end{cases} \qquad \textbf{(4-16)}$$

The second condition ($i \le \alpha - 1, j \le i$) includes the case when a collision occurs during the reference packet's transmission. In this case, the number of users is temporarily equal to $j + 1 + m$, where m is the number of users that cause the collision and are dropped. These users have to be accounted for when determining the success of the reference packet, since they cause a period of higher bit error rate before being aborted. The average number of *transmitting* users per slot, $\bar{x}$, or *mean carried data load*, is

$$\bar{x} = E[\mathbf{X}] = \sum_{k=1}^{\beta} k.X(k) \qquad \textbf{(4-17)}$$

4.4.5 Computation of the system throughput (output flow rates)

Since the analysis presented in equations 4-3 to 4-17 has been conditioned on the two unknown variables $\overline{L}_{dR}$ and $\bar{n}$, we present the *message throughput* or *message output flow rate*, as functions of the unknowns:

$$\begin{aligned} \textbf{\textit{Message output flow rate}} &= S_M(\bar{n}, \overline{L}_{dR}) \\ &= G_d.P_{MS} \end{aligned} \qquad \textbf{(4-18)}$$

where P_{MS} is obtained by substituting equations 4-14 and 4-15 into equation 3-24. In DT/CDMA, the success of a packet was conditioned on the success of the entire message. This is reflected in equation 3-28, where we see that the expected packet throughput is found by averaging over the distribution of message lengths with respect to the number of packets in each message and the probability of success of the message. In SR-DT/CDMA, some packets in a message may be successfully received, but the total message itself not. We instead use the following expression to compute the expected *packet throughput* or *packet output flow rate*, again as a function of $\bar{n}$ and $\overline{L}_{dR}$:

$$\begin{aligned} \textbf{\textit{Packet output flow rate}} &= S_P(\bar{n}, \overline{L}_{dR}) \\ &= \sum_{j=0}^{\beta} \sum_{i=0}^{\beta} j.X(i).\pi_{ij}^{PS} \end{aligned} \qquad \textbf{(4-19)}$$

where transition probability π_{ij}^{PS} is equivalent to $\hat{\pi}_{ij}^{PS}$, the only difference being that π_{ij}^{PS} represents the probability that the *base station* counts a reference packet as being successful given that it sees the system state change from $x = i$ to $x = j$.

$$\pi_{ij}^{PS} = \begin{cases} \sum\limits_{k=0}^{i} f_{N_d}^{b,i}(j-i+k).b(k,i,\mu_d).P_{PS}(j) & i \le \alpha, j > i \\ \sum\limits_{k=i-j}^{i} f_{N_d}^{b,i}(j-i+k).b(k,i,\mu_d).P_{PS}(j) \\ \quad + b(i-j,i,\mu_d)\cdot \sum\limits_{m=\beta+1-j}^{N_d-j} f_{N_d}^{b,i}(m).P_{PS}(j+m) & i \le \alpha, j \le i \\ b(i-j,i,\mu_d).P_{PS}(j) & i > \alpha, j \le i \\ 0 & i > \alpha, j > i \end{cases} \tag{4-20}$$

This transition probability is required in equation 4-19 because collisions have to be accounted for in the probability of packet success. Although the number of carried packets can never exceed *β*, collisions cause higher bit error rates during the period before the *m* contending packets abort (case $i \le \alpha, j \le i$).

4.4.6 Computation of the expected message delay

The average message delay for the finite population model is significantly simpler to express than it was in section 3.4.12. Derived from queuing theory, a version of Little's formula states that if the number of backlogged terminals ($\overline{n}$) and the expected message throughput ($S_M(\overline{n},\overline{L}_{dR})$) are known, then the expected delay is given by (conditioned on $\overline{n}$ and $\overline{L}_{dR}$) [Raychaudhuri, 1987]:

$$\overline{D}(\overline{n},\overline{L}_{dR}) = \frac{\overline{n}}{S_M(\overline{n},\overline{L}_{dR})} = \frac{\overline{n}}{G_{d0}} \tag{4-21}$$

In terms of queuing theory, Little's formula states that the expected time waiting in a queue is given by the expected number of people waiting in the queue divided by the rate at which people join the queue. In terms of our protocol, $\overline{n}$ represents the number of terminals waiting in the queue and G_{d0} represents the rate of flow of people into the queue (data message input flow rate). Since, at steady state, the rate of flow of messages into the system equals the message throughput, equation 4-21 effectively describes the message delay in terms of $\overline{n}$ and $S_M(\overline{n}, \overline{L}_{dR})$.

4.4.7 Equating the flow equations to obtain a solution

In the context of the analysis, the fact that we have two unknowns means that we require at least a pair of simultaneous equations in order to solve for the two unknowns. As in [Raychaudhuri, 1987] and [Joseph & Raychaudhuri, 1993], we can create a pair of simultaneous equations in $\overline{n}$ and $\overline{L}_{dR}$ by equating the message and packet input and output flow rates. In doing this, we define the following flow equations:

The "***message flow equation***" is derived by equating 4-8 and 4-18

$$\begin{gathered} \textit{Message inflow} = \textit{Message outflow} \\ (N_d - \overline{n}).\rho_{d0} = S_M(\overline{n}, \overline{L}_{dR}) \end{gathered} \qquad \textbf{(4-22)}$$

Similarly, the "***packet flow equation***" is derived by equation 4-9 and 4-19

$$\begin{gathered} \textit{Packet inflow} = \textit{Packet outflow} \\ (N_d - \overline{n}).\rho_{d0}.\overline{L}_{d0} = S_P(\overline{n}, \overline{L}_{dR}) \end{gathered} \qquad \textbf{(4-23)}$$

The solution of this pair of simultaneous equations reveals the solution to the two unknowns that satisfy the equilibrium state. These solutions, defined as $\overline{n}^*$ and $\overline{L}^*_{dR}$, which satisfy both equations 4-22 and 4-23 will yield the final system performance measures when substituted into equations 4-3 to 4-21. Since, at equilibrium, the packet input flow rate is equal to the message input flow rate multiplied by the mean length of new messages, the following is also true for the output flow rates, i.e.

$$S_P(\overline{n}^*, \overline{L}^*_{dR}) = S_M(\overline{n}^*, \overline{L}^*_{dR}).\overline{L}_{d0} \qquad \textbf{(4-24)}$$

A popular and common means of solving for $\overline{n}^*$ and $\overline{L}_{dR}^*$ is by plotting the message flow equation 4-22 and packet flow equation 4-23 as separate curves on the $(\overline{n}, \overline{L}_{dR})$ plane, as described in [Raychaudhuri, 1987]. The intersection(s) of these two curves yields the system's equilibrium point(s), $(\overline{n}^*, \overline{L}_{dR}^*)$. An obvious disadvantage of this method is that it requires a large number of computations of equations 4-3 to 4-20, at a high enough resolution (very small increments of $\overline{n}$ and $\overline{L}_{dR}$) such that an accurate intersection(s) point can be obtained.

4.4.8 System performance for an infinite population model

To conclude the analysis of SR-DT/CDMA, we modify the analysis in equations 4-10 to 4-20 to account for an infinite population model. These modifications are only superficial and are presented as follows:

- Replace $f_{N_d}^{b,x}(.)$ in equations 4-10, 4-13, 4-16 and 4-20 with the Poisson probability $f_\infty(.)$ from equation 3-2.
- μ_d as used in equations 4-10, 4-13, 4-16 and 4-20 is obtained as explained in section 3.4.1 (equation 3-3) rather than as obtained using equation 4-4.
- Use $G_d = \lambda t_s$ (section 3.4.1) for the total offered load instead of equation 4-7.
- Only one iteration of equations 4-10 to 4-20 is required, as in chapter 3, since there are no unknown network parameters other than the final throughput values.

Since the expected message delay is less representative of the actual system performance for the infinite population model (as discussed in section 3.5.4), we omit its derivation for this model. In fact, equations 3-30 to 3-38 should suffice given that the appropriate changes are made to $\hat{\pi}$ and $\hat{X}(i)$ to accommodate the SR-DT/CDMA policy.

4.5 Performance Results

In this section we evaluate the performance of the SR-DT/CDMA MAC protocol, as well as the validity and accuracy of our Equilibrium-Point/Markov analysis for the finite population traffic model. In order to compare the performance of SR-DT/CDMA with the DT/CDMA protocol of the previous chapter, we also present results for the infinite population (Poisson) model, using the modifications discussed in section 4.4.8.

4.5.1 Network parameters

All system parameters as outlined in section 3.5.1 and table 3.1 are used in the analysis and simulation of SR-DT/CDMA, with the following exception: For the finite population model, each user generates a new message in a slot with probability ρ_{d0} and each backlogged user retransmits its backlogged message with probability ρ_{dR}. The value of ρ_{d0} is varied to present different load values to the network (G_d is related to ρ_{d0} through equations 4-5 to 4-7). Table 4.2 lists the imagined network parameters.

Item (*units*)	**Symbol**	**Value**
Data population (***terminals***)	N_d	100, ∞
Data collision threshold (***transmissions***)	β	To be determined
Data blocking threshold (***transmissions***)	α	To be determined
Average new data message length (***slots***)	$\overline{L}_{d0}$	4
∴New message termination probability	μ_{d0}	0.25
New message transmission probability ($N_d = 100$)	ρ_{d0}	Variable
Offered data load ($N_d = \infty$)	G_d	Variable
Average data backoff delay (***slots***)	$\overline{B}$	5
∴Data message retransmission probability	ρ_{dR}	0.2

Table 4.2 : Example network parameters used in the analytical evaluation and simulation

4.5.2 Determination of the collision threshold β

For DT/CDMA, we used the quantitative experimentation procedure outlined in section 3.5.2 to determine a suitable value for the collision threshold, β. In this section, we repeat this procedure for SR-DT/CDMA, using exactly the same network parameters, in order to see if the MAC protocol has any effect on the optimum value of β. Experimentation with β reveals that a value of β=20 appears to provide the best overall system performance for SR-DT/CDMA with the given set of network parameters. In figure 4.4 we plot a few, yet widely spread, values of β and, as can be seen, a value of $\beta = 20$ appears to give the optimum throughput results. For values $\beta >> 20$, the curves tend towards the unrestricted slotted-SS/ALOHA curve ($\beta = \infty$) while for values $\beta < 20$ it can be seen that premature collision detection takes place and full usage is not made of the channel capacity.

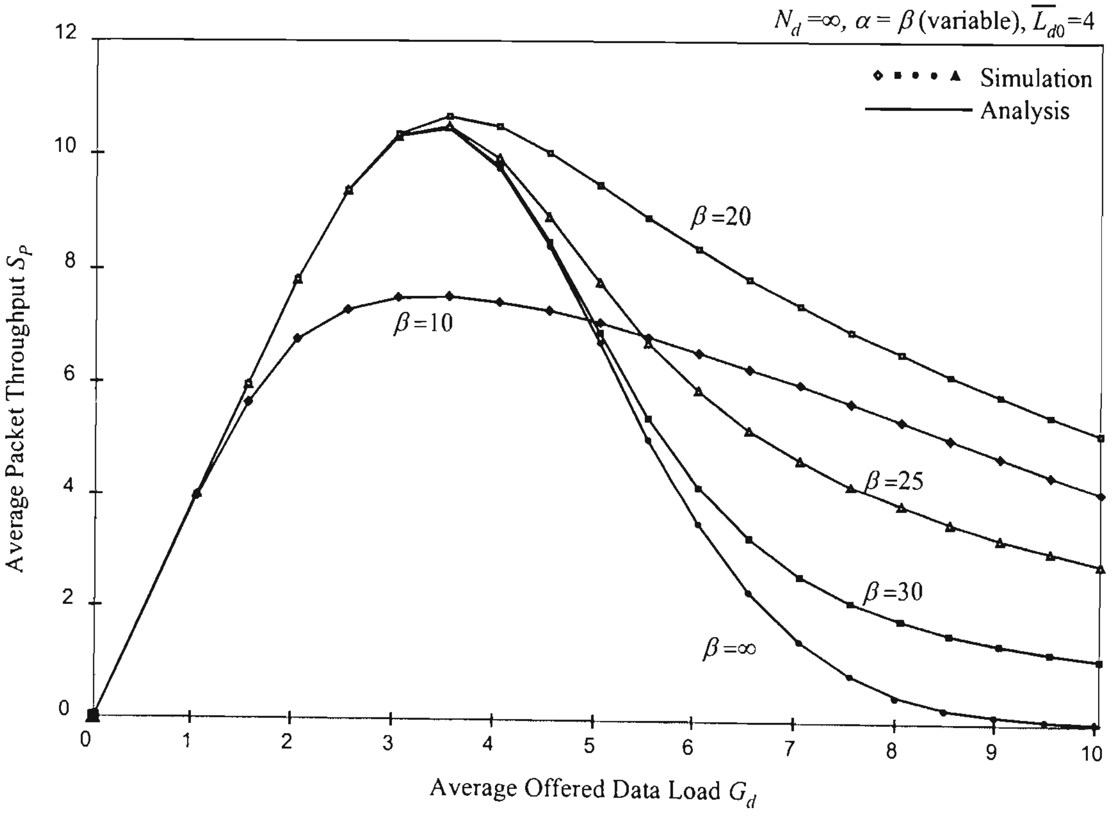


Figure 4.4 : Data system throughput plots for various values of β

In figure 4.5a (infinite population) and figure 4.5b (finite population, $N_d = 100$) we plot simulation results for values of β which are closer to the proposed optimum. As can be seen, the throughput curves are all very similar at and around $\beta = 20$. It is interesting to note that different values of β close to $\beta = 20$ provide slightly better throughput results than others at different loads, in much the same way as the variable α threshold, only to a much lesser extent. In particular, slightly lower thresholds than the optimum appear to provide slightly better results at higher loads, even though the peak of the curve is lower than that of the optimum. It may be argued therefore that perhaps β should also be a function of the channel load. A preliminary investigation into having a variable β as well as a variable α threshold was performed and the results yielded a negligible performance improvement over the case where β is kept constant and α is varied according to the channel load. The performance improvements gained by varying α completely swamped the advantages gained by varying β as well, and all outcomes were equal to the case of having a fixed β and a variable α.

The next point of interest is the fact that the optimum value of β for the DT/CDMA protocol in Chapter 3 is different to that for the SR-DT/CDMA protocol. This verifies our intuitions in section 3.5.2 that the MAC protocol and retransmission scheme have an effect on the maximum threshold. Although difficult to quantify the difference, we may at least offer explanations for the difference and for why the threshold is slightly higher for the SR-DT/CDMA scheme. In the DT/CDMA scheme, corruption of a single packet meant corruption of the entire message and hence the success of a packet is dependent on the success of the message in which it resides. We can effectively say then, that the *message* is the primary unit of interest in DT/CDMA since we require an entire message to be successful in order for all its packets to be successful. In the SR-DT/CDMA protocol, the probability of packet success is entirely independent of the success of any other packets in the message. This independence of packets means that we can consider the *packet* to be the primary unit of interest in ST-DT/CDMA. For a given bit error rate, we thus expect the probability of a message's success in DT/CDMA to be poorer than the probability of a packet's success in SR-DT/CDMA, purely because of the fact that a multi-packet message has more bits than a single packet. From a different viewpoint, we may say that: since SR-DT/CDMA is more efficient in

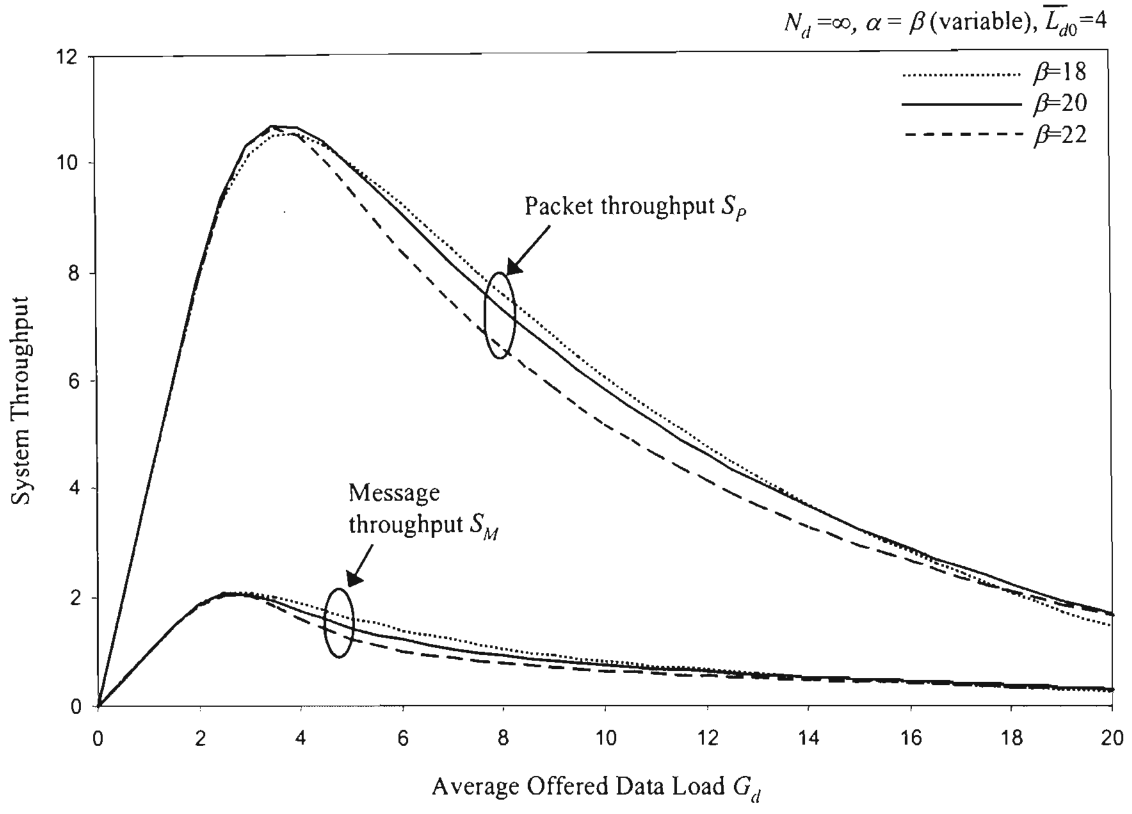


Figure 4.5a : Data system throughput plots for various values of β close to the optimum threshold (infinite population model)

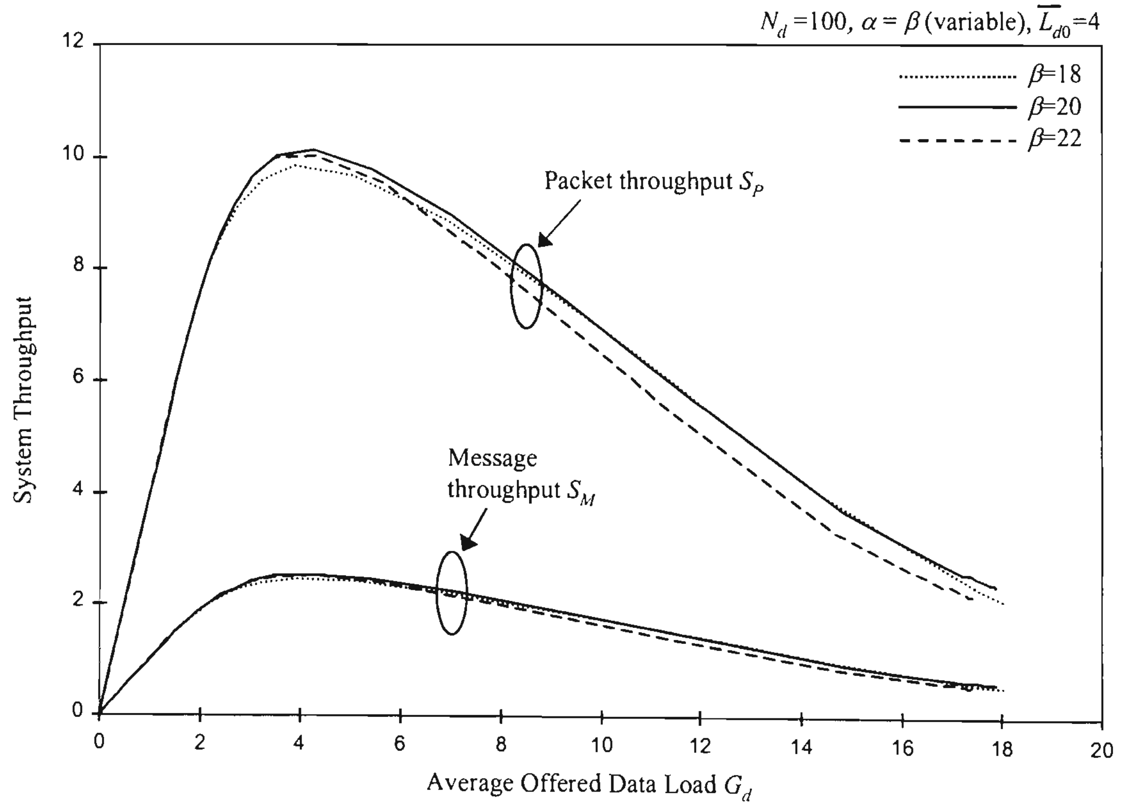


Figure 4.5b : Data system throughput plots for various values of β close to the optimum threshold (finite population model)

successfully transporting information across the channel than DT/CDMA, we can afford to increase the number of users allowed in the channel.

Finally, the last point of interest in figure 4.5b (finite population), is that the simulation results prove that the mathematical assumption made in equation 4-24 is true, i.e. for a finite population SR-DT/CDMA protocol, it does not matter whether we plot the *packet throughput* S_P or the *message throughput* S_M, since $S_P = \overline{L}_{d0} S_M$ at equilibrium. For our network parameters ($\overline{L}_{d0}$=4), we expect the packet throughput to be always four times the message throughput for all possible load values at steady state equilibrium conditions. The message throughput curves on figure 4.5b verify this. The same cannot be said for the infinite population model in figure 4.5a however, and the reason for this is simple. Retransmitted messages (and therefore all messages) in the infinite population model are assumed to have the same mean length as new messages. At each retransmission of a message, the message length is supposed to be reduced by the number of packets that were successfully transmitted and hence the retransmitted message success probability is supposed to increase at each retransmission. Instead, for the infinite population scheme we assume that all retransmissions have the same mean length and hence the same probability of success. Because of this, the message throughput will always be less than the packet throughput divided by the mean message length (i.e. $S_M < S_p / \overline{L}_{d0}$) for the infinite population model.

4.5.3 Effect of the α threshold on the system performance

We now evaluate the effect of the blocking threshold α on the performance of the SR-DT/CDMA data admission policy. In figure 4.6, we compare the throughput performance of the model for various values of α, with a fixed collision threshold β=20. A few α values are considered for clarity purposes only.

Again, we observe similar results to those obtained in section 3.5.3 for DT/CDMA. At low loads, high values of α are preferable, whereas at high load, lower values of α are preferable, for both throughput and delay. Whereas in DT/CDMA, α could be chosen to optimise either the message throughput or the packet throughput, in SR-DT/CDMA

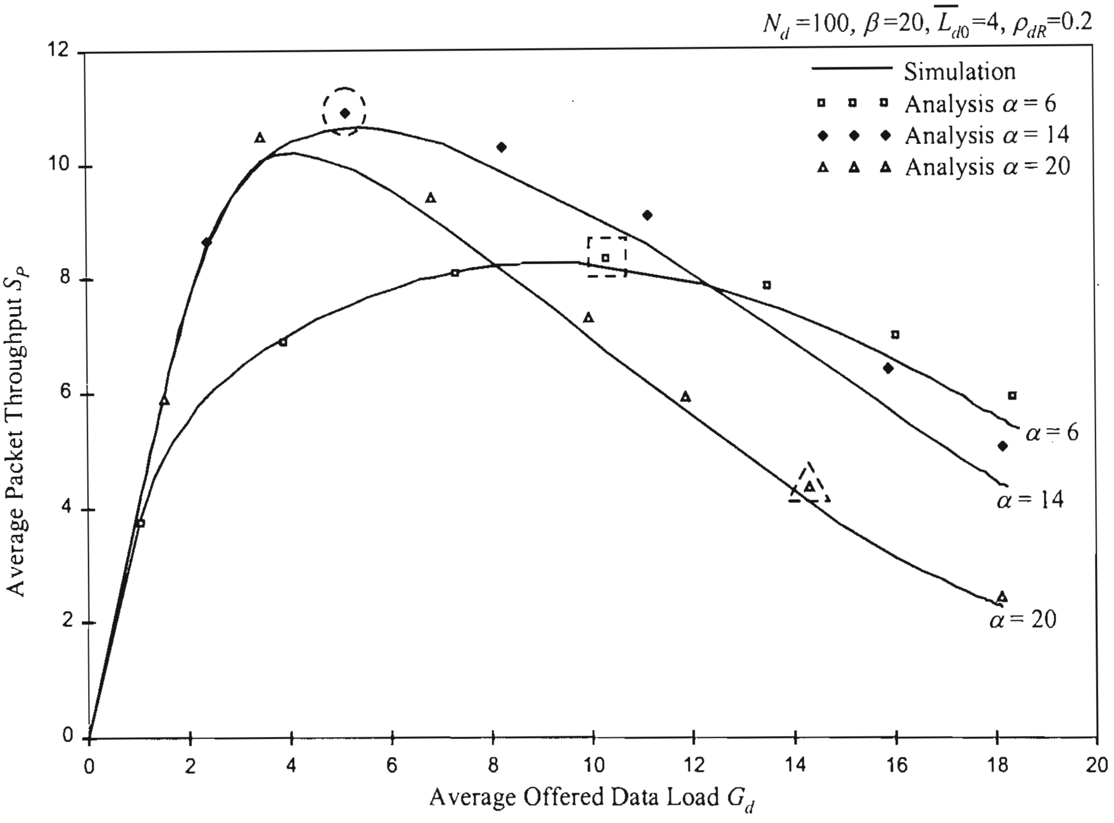


Figure 4.6a : Effect of the α threshold on the system throughput (finite population)

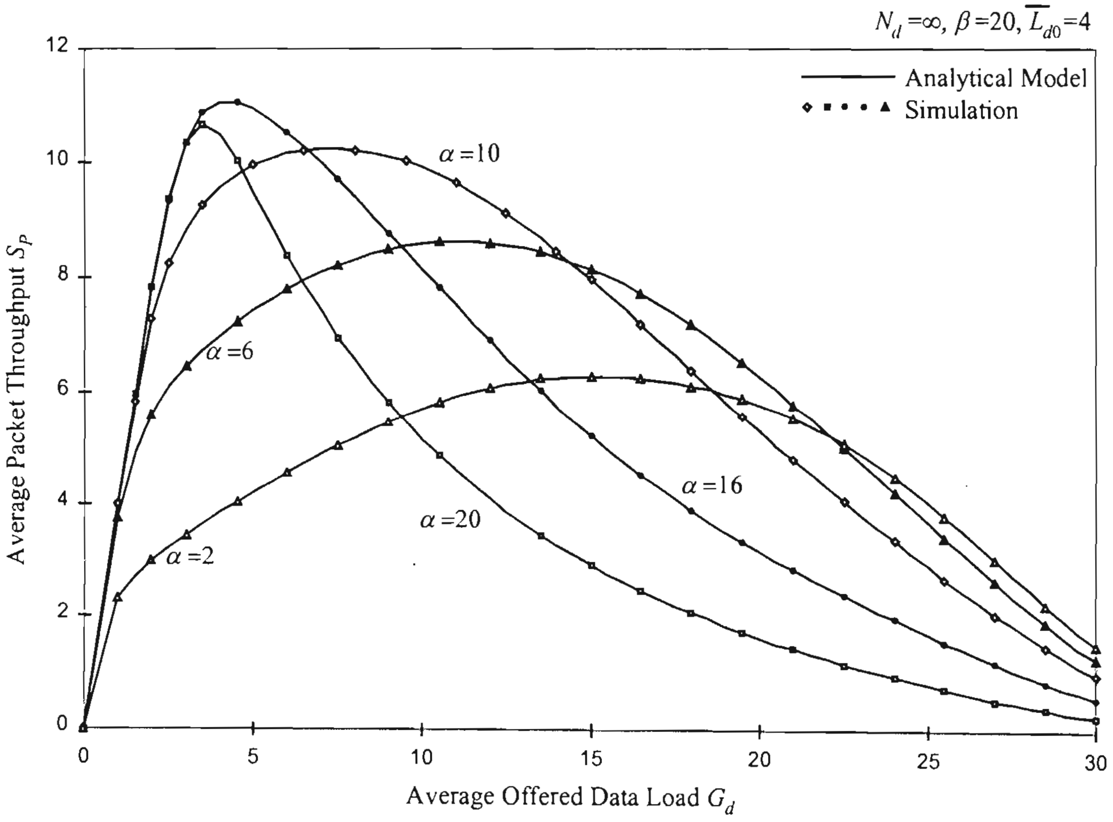


Figure 4.6b : Effect of the α threshold on the system throughput (infinite population)

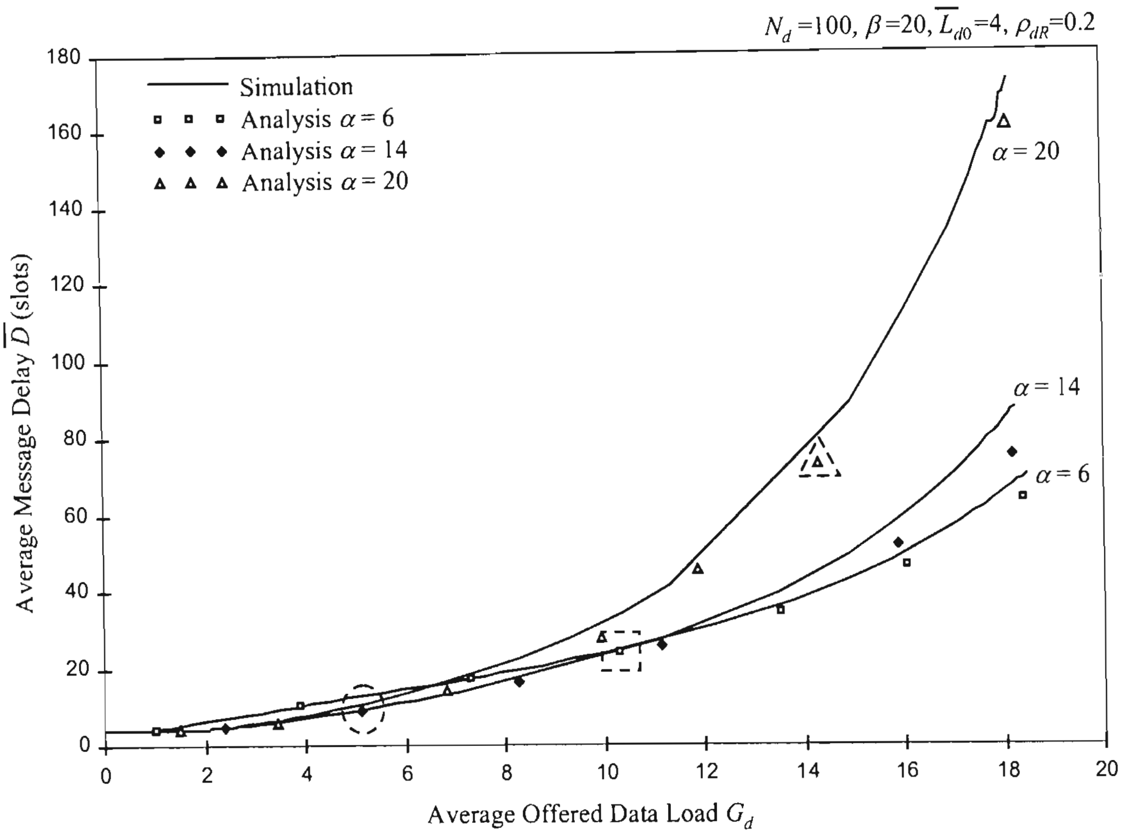


Figure 4.7 : Effect of the α threshold on the average message delay

both message and packet throughputs are simultaneously optimised due to the fact that they are theoretically linearly related (equation 4-24). In figure 4.7 we plot delay results for SR-DT/CDMA as computed using equation 4-21 for the finite population model. Again, the significant reduction in delay can be seen by decreasing α as the offered load increases.

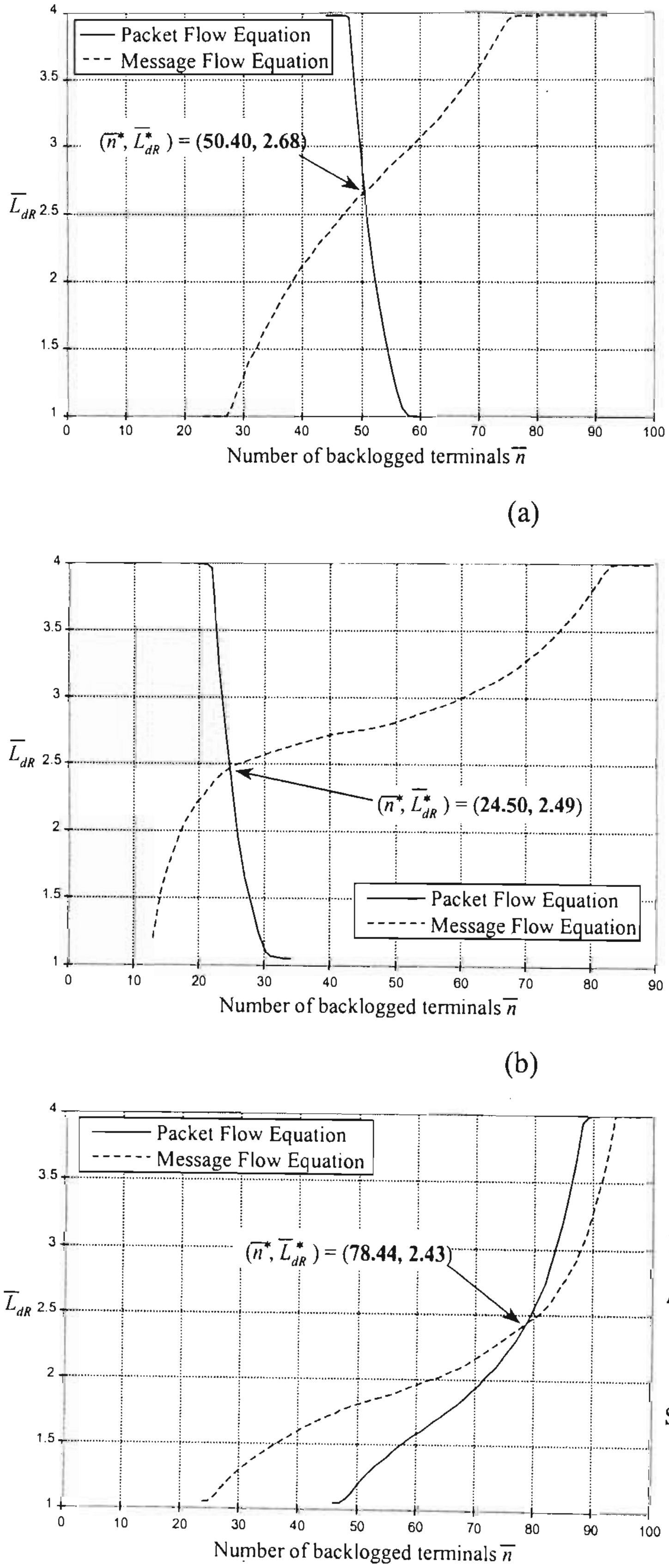
4.5.4 Evaluation of the finite population equilibrium-point analysis

In figures 4.6a and 4.7 for the finite population model, we plot simulation results as solid lines and analytical solutions as a few single points. The reason for this is due to the fact that the analytical model results are computationally intensive. For each point on the curve, the EPA model has to compute multiple flow equations for each point on the $(\overline{n}, \overline{L}_{dR})$ plane before being able to determine the equilibrium points. The resolution at which this is performed determines how accurate the final equilibrium point is. In fact, for a single load-throughput point the simulation is able to generate a decent steady-state result in a time period that is at least an order of magnitude smaller than the time period required for the EPA to converge to a result. For the infinite population model in which only one iteration of the Markov analysis is required, the reverse is seen to be true, (i.e. the analytical model is at least an order of magnitude faster than the simulation).

Figures 4.8a, 4.8b and 4.8c give example $(\overline{n}, \overline{L}_{dR})$ flow equation plots for some of the analytical points on figures 4.6a and 4.7 (namely, ⊡, ⊙ and △ respectively). The exact procedure for determining each equilibrium point is as follows:

1. the value of $\overline{n}$ was varied between 0 and N_d=100 in increments of 1.0
2. for each $\overline{n}$ value, $\overline{L}_{dR}$ was varied between 1.0 and $\overline{L}_{d0}$=4.0 in increments of 0.02
3. all values of $(\overline{n}, \overline{L}_R)$ which produced equal message input and output flow values were plotted on the $(\overline{n}, \overline{L}_R)$ plane as a dotted line.
4. all values of $(\overline{n}, \overline{L}_R)$ which produced equal packet input and output flow values were plotted on the $(\overline{n}, \overline{L}_R)$ plane as a solid line.
5. The point at which the two lines intersect yielded an approximate $(\overline{n}^{/}, \overline{L}_R^{/})$ solution.
6. Once an approximate solution was found, the value of n was varied between $\overline{n}^{/} - 5$ and $\overline{n}^{/} + 5$ in increments of 0.01, while the value of $\overline{L}_{dR}$ was varied between $\overline{L}_{dR}^{/} - 0.2$ and $\overline{L}_{dR}^{/}$+0.2 in increments of 0.01.
7. The improved resolution in step 6 yielded the final approximate equilibrium point $(\overline{n}^{*}, \overline{L}_R^{*})$

Another possibility for obtaining a solution is to use a mathematical optimisation package to simultaneously solve equations 4-22 and 4-23. We prefer the above scheme as it provides some graphical insight into the nature of the flow equations, and the crossing points themselves. As can be seen, for each pair of curves there exists only one crossing point. In the context of stability, this implies that the system is unconditionally stable for these network parameters, as discussed in section 2.4.2 and the literature [Raychaudhuri, 1987], [Jenq, 1980], [Carleial & Hellman, 1975]. By stable, we refer to the fact that if all idle terminals stop transmitting ($\rho_{d0} = 0$) then the system throughput



Point ⊡

Analysis : $(\overline{n}^*, \overline{L}^*_{dR}) = (\mathbf{50.40, 2.68})$
$G_d = \mathbf{10.30,}$
$S_P = \mathbf{8.33}$, $\overline{D} = \mathbf{24.17}$

Simulation : $(\overline{n}, \overline{L}_{dR}) = (\mathbf{51.24, 2.85})$
$G_d = \mathbf{10.44,}$
$S_P = \mathbf{8.18}$, $\overline{D} = \mathbf{25.01}$

(a)

Point ⊙

Analysis : $(\overline{n}^*, \overline{L}^*_{dR}) = (\mathbf{24.50, 2.49})$
$G_d = \mathbf{5.14,}$
$S_P = \mathbf{10.87}$, $\overline{D} = \mathbf{9.035}$

Simulation : $(\overline{n}, \overline{L}_{dR}) = (\mathbf{26.14, 2.21})$
$G_d = \mathbf{5.41,}$
$S_P = \mathbf{10.62}$, $\overline{D} = \mathbf{9.82}$

(b)

Point △

Analysis : $(\overline{n}^*, \overline{L}^*_{dR}) = (\mathbf{78.44, 2.43})$
$G_d = \mathbf{14.30,}$
$S_P = \mathbf{4.322,}$ $\overline{D} = \mathbf{73.02}$

Simulation : $(\overline{n}, \overline{L}_{dR}) = (\mathbf{81.62, 2.79})$
$G_d = \mathbf{14.87,}$
$S_P = \mathbf{4.01}$, $\overline{D} = \mathbf{84.44}$

(c)

Figure 4.8: Example packet and message flow equation contours for some of the equilibrium solutions plotted on figures 4.6a and 4.7.

is high enough such that all backlogged terminals will eventually be successful and the population will drift back to the scenario in which all terminals are idle. In an unstable saturated network, practically all terminals are backlogged. The system throughput is practically zero due to the high rate of retransmissions from backlogged terminals and, as a result, the network remains saturated forever (or until all terminals completely abandon retransmissions and return to the idle state, or increase the backoff delay according to a suitable CRA).

In table 4.3, we compare the analytical (AN) and simulation (SIM) results for the curves plotted in figures 4.6a and 4.7 for α=6. The highlighted row (ρ_{d0}=0.042) corresponds to point ⊡ in figures 4.6a and 4.7, as found using figure 4.8a. Similarly, tables 4.4 and 4.5 compare analytical and simulation results for the points on the curves α=14 and α=β=20 of figures 4.6a and 4.7 respectively. Highlighted values in tables 4.4 and 4.5 correspond to the points ⊙ and ◬ as found using figures 4.8b and 4.8c respectively.

When comparing the results from figures 4.6a, 4.7 and 4.8, and tables 4.3, 4.4 and 4.5, we can see that the EPA model provides a reasonably good approximation of the network performance when compared with results obtained from the simulation. With respect to the estimated average message length of retransmitted messages, $\overline{L}_{dR}$, we can see that large discrepancies occur. At low channel loads (low values of ρ_{d0}), $\overline{L}_{dR}$ is grossly overestimated. This is not so serious however, since at low loads we expect the fraction of retransmitted messages to be very low. At higher loads, the analytical model starts to underestimate $\overline{L}_{dR}$. These discrepancies are almost certainly attributed to the fact that the two simplifying assumptions we made at the outset of the EPA (in section 4.3) are not entirely valid: i.e.

- the distribution for the length of retransmitted messages is not exactly geometric
- and, perhaps more influential, the fact that we assumed in section 4.3.3 that the number of backlogged terminals ($\overline{b}$) is approximately equal to the total number of non-idle terminals ($\overline{n}$).

ρ_{d0}	$\overline{n}$		$\overline{L}_{dR}$		G		S_P		S_M		$\overline{D}$	
	SIM	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**
0.01	4.19	4.05	3.93	3.95	1.03	1.02	3.80	3.73	0.95	0.96	4.38	4.23
0.021	17.49	17.75	3.85	3.90	3.84	3.89	6.93	6.90	1.73	1.72	10.07	10.30
0.03	32.87	34.48	3.41	3.16	6.92	7.30	8.02	8.07	2.01	2.02	16.34	17.22
0.042	**51.24**	**50.40**	**2.85**	**2.68**	**10.44**	**10.30**	**8,18**	**8.33**	**2.04**	**2.08**	**25.01**	**24.17**
0.06	69.00	67.35	2.54	2.37	13.78	13.50	7.43	7.83	1.86	1.96	37.03	34.41
0.09	82.25	80.60	2.47	2.26	16.31	16.03	6.37	6.98	1.59	1.46	51.46	46.14
0.2	93.36	92.66	2.51	2.22	18.44	18.36	5.33	5.87	1.33	1.74	69.92	63.25

Table 4.3 : Comparison between analytical model (AN) results and simulation (SIM) results for the points on the α=6 curve of figures 4.6a and 4.7.

ρ_{d0}	$\overline{n}$		$\overline{L}_{dR}$		G		S_P		S_M		$\overline{D}$	
	SIM	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**
0.024	10.21	9.97	1.67	3.45	2.42	2.39	8.62	8.645	2.15	2.16	4.74	4.61
0.036	**26.14**	**24.50**	**2.21**	**2.49**	**5.41**	**5.14**	**10.62**	**10.87**	**2.66**	**2.71**	**9.82**	**9.03**
0.044	44.55	41.60	2.32	2.21	8.75	8.27	9.73	10.28	2.43	2.56	18.26	16.21
0.053	61.46	57.32	2.41	2.12	11.87	11.16	8.15	9.05	2.03	2.27	30.10	25.27
0.09	84.95	82.64	2.56	2.18	16.31	15.90	5.43	6.38	1.36	1.59	62.55	51.73
0.2	94.57	93.72	2.65	2.24	18.21	18.16	4.31	5.02	1.08	1.25	87.10	74.91

Table 4.4 : Comparison between analytical model (AN) results and simulation (SIM) results for the points on the α=14 curve of figures 4.6a and 4.7.

ρ_{d0}	$\overline{n}$		$\overline{L}_{dR}$		G		S_P		S_M		$\overline{D}$	
	SIM	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**	**SIM**	**AN**
0.016	6.12	6.00	1.02	2.92	0.96	1.52	5.99	5.90	1.50	1.50	4.08	3.98
0.031	17.99	15.60	1.72	2.85	3.78	3.46	10.17	10.46	2.54	2.61	7.05	5.96
0.036	37.11	33.18	2.36	2.42	7.01	6.84	8.91	9.40	2.20	2.37	16.88	14.00
0.039	56.92	49.86	2.61	2.19	10.60	9.95	6.35	7.29	1.67	1.82	34.01	27.40
0.043	71.36	66.10	2.69	2.29	12.92	11.90	4.88	5.88	1.22	1.46	57.99	45.22
0.05	**81.62**	**78.44**	**2.79**	**2.43**	**14.87**	**14.30**	**4.01**	**4.32**	**1.00**	**1.07**	**84.44**	**73.02**
0.2	97.17	96.96	2.97	2.71	18.12	18.11	2.26	2.43	0.56	0.60	172.2	160.7

Table 4.5 : Comparison between analytical model (AN) results and simulation (SIM) results for the points on the α=β=20 curve of figures 4.6a and 4.7.

In incorrectly estimating $\overline{L}_{dR}$, we naturally expect the EPA to estimate an incorrect value for $\overline{n}$. In most cases, the model appears to slightly underestimate the number of busy users $\overline{n}$. When $\overline{n}$ is large, it invariably means that the number of backlogged users ($\overline{b}$) is also large, since the number of transmitting users ($\overline{x}$) can never be larger than β=20. For small values of $\overline{n}$, we expect the backlog $\overline{b}$ to be almost zero, with the bulk of $\overline{n}$ consisting of transmitting users. Thus when underestimating $\overline{n}$, we expect the combined offered load, G_d, to be underestimated as well. In equation 4-5, the load from idle users, G_{d0}, will be overestimated. In equation 4-6, the expected load from the backlogged users, G_{dR}, will be underestimated. However, since ρ_{d0} is significantly less than ρ_{dR} in most cases (ρ_{dR}=0.2), we expect G_{dR} to be underestimated to a higher degree than G_{d0} and hence the total load G=$G_{d0}+G_{dR}$ to be underestimated. This is a possible explanation for why the EPA model overestimates the system throughput measures S_M and S_P. As a result of underestimating the number of busy users $\overline{n}$ and overestimating the system throughput, we expect the average delay $\overline{D}$ (equation 4-21) to be underestimated. This is indeed confirmed in the tabulated results and plots in figures 4.6a and 4.7. All in all, the correspondence between the simulation and analytical model results are reasonably good. In fact, the EPA model does a satisfactory job in predicting the system performance, despite the significant simplifying assumptions that were made in section 4.3 regarding the traffic model.

4.5.5 Comparison of DT/CDMA and SR-DT/CDMA

In this section, we compare the relative performance of the two proposed MAC protocols DT/CDMA and SR-DT/CDMA of Chapters 3 and 4. Figure 4.9 compares the optimum packet throughput curves, as obtained if we imagine the envelopes of the curves from figures 4.6a, 4.6b and 3.11.

We can clearly see that the partial message retransmission scheme in SR-DT/CDMA provides far better throughput results than the entire message retransmission scheme in DT/CDMA. The two delay curves in figures 3.14 and 4.7 reveal that SR-DT/CDMA is far superior to DT/CDMA with respect to the average message delay and overall system

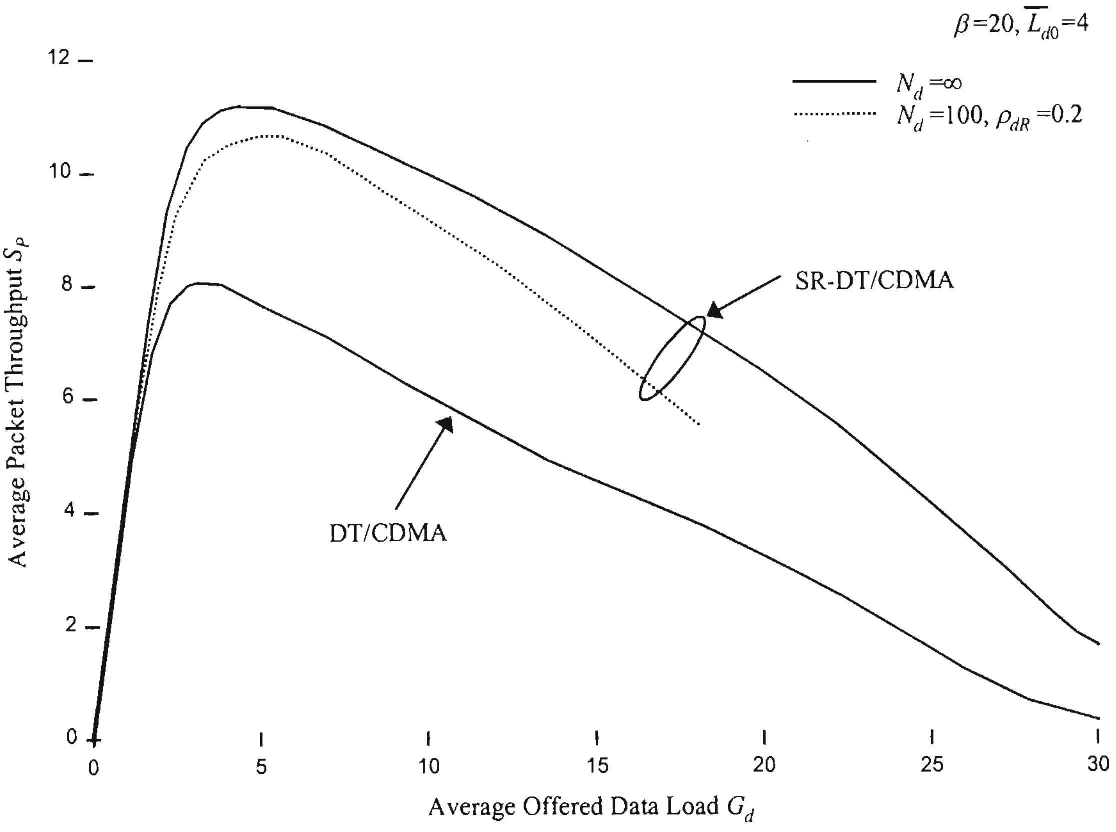


Figure 4.9 : Comparison between DT/CDMA and SR-DT/CDMA

stability. At this point, it is worth mentioning that the stability of the data network can be directly controlled by the base station through the use of the dynamic blocking threshold α. Although not explicitly addressed in this thesis, the issue of network stability can be considered when analysing the throughput and delay performance improvements. In order for a random access network, such as the one proposed for the data network in this thesis, to become more stable, two things must occur: firstly, the message throughput has to increase to ensure that a larger proportion of the backlogged population can successfully retransmit their backlogged messages and return to the idle state; and secondly, the average message delay must decrease to ensure that backlogged terminals remain backlogged for as short a period as possible. By dynamically adjusting α with regards to the channel load, it is possible to increase the system throughput and decrease the average message delay and thus improve the network stability with respect to the message and packet flow rates and backlog vs. idle population ratio. This novel method differs from the conventional schemes which adjust the average backoff delay $\overline{B}$ by varying the retransmission probability p_{dR}, as discussed in section 2.4.2 (stabilising CRA's). In our model, we fix p_{dR} to be quite large (p_{dR}=0.2 which results in a small average backoff delay of only $\overline{B}=5$ slots) and find it more interesting to control the performance via the dynamic α threshold. We choose a low average backoff delay to ensure that the overall message delay is small. The choice of ρ_{dR} also determines the maximum possible offered load to the network. If all N_d terminals are backlogged, then the maximum possible offered load to the system is $G_{d(\max)} = N_d \rho_{dR} = 100 * 0.2 = 20$. Invariably, there will always be a handful of terminals transmitting and hence the backlog can never be exactly N_d. Simulation and analysis results for the given network parameters with $\rho_{dR} = 0.2$ yielded that the system backlog $\overline{b}$ generally converged to about 90 users for large values of ρ_{d0} (with $\overline{n}$ converging to almost N_d=100). This yields a maximum offered load of approximately 90*0.2=18 messages per slot. Figures 4.6a and 4.7 confirm this result, where $G_{d(\max)} \approx 18$. Even at this high load, we can see that the optimum throughput is a non-zero value that is high enough that, should all idle terminals defer from attempting to transmit, the system should drift back to the stable state in which all terminals are idle.

4.6 Summary

In this chapter we proposed and analysed an improved data MAC policy, called "Selective-Repeat Dual-Threshold CDMA" for the data network. The new policy, which employs retransmission of only those packets in a message which are corrupt, is shown to be far superior to the DT/CDMA MAC protocol proposed in Chapter 3, which retransmits the entire message regardless of the number of corrupt packets in the message.

The modification to the MAC protocol, together with several simplifying assumptions, allow us to use a more realistic finite population traffic model that is able to distinguish between new and retransmitted messages. This traffic model is discussed in section 4.3. As a result of the finite population traffic model, the analytical modelling of the MAC protocol is complicated by the introduction of several unknown network parameters. To solve for these unknowns, an Equilibrium-Point Analysis approach is adopted. In section 4.4, the Markov analysis derived in Chapter 3 is modified to accommodate the improved MAC policy and selective-repeat message retransmission aspect, and a combined EPA/Markov analysis is derived to obtain expressions for the performance measurements of the SR-DT/CDMA MAC protocol for the proposed finite population traffic model. In order to compare SR-DT/CDMA with our original DT/CDMA protocol, we also derive a Markov model for the performance of SR-DT/CDMA for the infinite population Poisson model.

In the results section 4.5, we again show that the system performance can be greatly enhanced through the use of a dual collision detection (β) and channel load throttling system with a dynamic message blocking threshold ($\alpha<\beta$). As for DT/CDMA, we show that for each channel load value, there exists an optimum setting for the blocking threshold α which optimises the system throughput and delay, whatever the operating conditions may be. Furthermore, we propose that the dynamic α threshold scheme is a novel means by which the network stability can be improved, as opposed to conventional schemes that employ changing the message retransmission probability p_{dR}.

The proposed SR-DT/CDMA MAC protocol and accompanying statistical analysis, as derived in this chapter, was presented at the IEEE ISSSTA'98 conference [Judge & Takawira, 1998a] and has been accepted for publication in the telecommunications journal "Wireless Networks" (ACM/Baltzer Science Publishers) in 2000/2001 [Judge & Takawira, 1998b].

CHAPTER 5

PROTOCOLS FOR INTEGRATING VOICE AND DATA

5.1 Introduction

Previous mobile and packet radio systems supporting multiple services have done so by implementing separate RF systems for each service. This independent architecture requires separate RF channels and dedicated equipment for each service. With the evolution of telecommunications towards global integrated service networks, much research has been done towards finding optimal ways in which to integrate different service types onto a single channel.

The most common problem addressed in the literature is the integration of voice and data onto a common channel. The particular combination of these two service types addresses a fundamental problem in integrated service networks: i.e. integrating a service (voice) which requires a *long term channel commitment*, real-time delivery and which is best transmitted over a circuit switched network, with a characteristically different service (data) which can be buffered and queued and is best transmitted over a packet switched network with a *short term commitment channel.* In addition, data services require a low bit error rate and a high probability of correct reception of the entire message, whereas humans are more tolerant of the audible effects of channel noise on speech.

In this chapter, we review several schemes that have been proposed for integrating voice and data (and sometimes other multimedia traffic types as well) onto a common channel. The first half of this chapter deals with narrowband TDMA/reservation based MAC protocols, while the second half deals with CDMA MAC protocols. We avoid presenting analyses and results for these schemes due to the nature of this brief review, and instead focus on the principles of each scheme.

5.2 Exploitation of the Silent Periods in Human Speech

A common method of integrating voice and data is to exploit the silent periods inherent in human speech to transmit data. This multiplexing technique is commonly referred to as packetised data, voice dedicated (PDVD) burst switching and various implementations of it have been proposed in the literature. It has been shown [Lee & Un, 1986], [Gruber, 1982], [Stern & Sobol, 1995] that two different types of silent periods can be categorised within a typical voice conversation :

1. **Talkspurt gaps**: human speech can be modelled as short bursts of vocal energy (talkspurts), separated by short periods of silence. These silent periods occur between sentences, phrases, words and even syllables.
2. **Exchange gaps**: these silent periods occur during exchanges in communication when a speaker finishes speaking and pauses while listening.

The *voice activity factor* of a voice channel is defined as the percentage usage of the channel for the actual transmission of vocal energy during a typical two-party conversation. Many studies have been done on the patterns of vocal energy in a conversation (e.g. [Lee & Un, 1986] and [Gruber, 1982]), taking into account the above two categories. Both Gruber and Lee & Un present measurements done on databases of recorded conversations. Lee & Un present results for only one person in the conversation, while Gruber analyses the total voice activity factor for both parties. Their results agree when showing that both the talkspurt durations and silent period durations (both talkspurt gaps and exchange gaps) can be approximated by a negative exponential distribution. For the one-way channel usage case, Lee & Un show that talkspurts are generated at a rate of about 72 talkspurts per minute, with a mean talkspurt duration of 227ms and a mean silent period duration of approximately 596ms. This implies a one-way voice activity factor of approximately 27 percent. They also show that about 40 percent of the talkspurt durations are less than 50ms long, and that the average intersyllabic gap duration is about 200ms, with 80 percent of the gap lengths being less than 200ms. In a similar fashion, Gruber measures an average talkspurt duration of 170ms and an average silent period duration of 124ms for the two-way case. This reveals a combined voice activity factor of 57 percent (approximately twice the result obtained by Lee & Un).

A common technique implemented in these systems is hangover [Gruber, 1982]. Hangover is used to bridge the very short intersyllabic gaps (less than 200ms) between talkspurts. Figure 5.1 illustrates this principle.

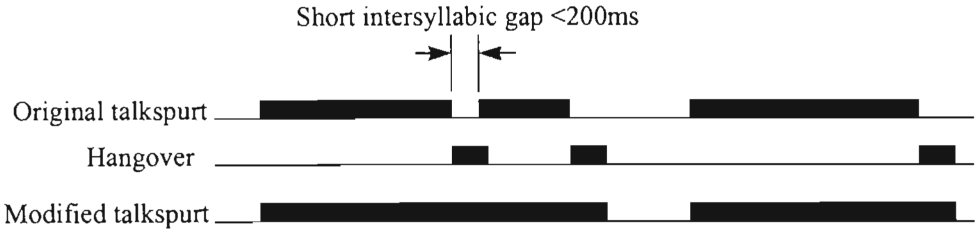


Figure 5.1: Illustration of the effects of hangover on talkspurt lengths

This process decreases the talkspurt rate and increases the mean talkspurt duration. It also increases the mean silent period duration by removing the larger fraction of short intersyllabic gaps, and leaving the longer inter-word and inter-sentence gaps. Other benefits of hangover include the minimising of back-end clipping of the talkspurts, a process which often occurs when the voice activity detector threshold fails to detect the lower power energy that occurs when a talkspurt trails off. Hangover ensures a more continuous flow of speech, and reduces the signalling burden on the multiplexing system by reducing the talkspurt rate.

The main disadvantage of hangover is the increase in the voice activity factor. Both Gruber and Lee & Un provide measurements done in a system which implemented an optimal hangover time of approximately 200ms. For one-way speech, the talkspurt rate decreases from 72 to 16 talkspurts per minute, while the average talkspurt and silent period durations increase to 1.3s and 2.2s respectively. This is unfortunately accompanied by an increase in the voice activity factor of 10 percent (to about 37 percent). A trade-off thus occurs between the conversation flow, the signalling burden and the voice activity factor.

Exploitation of silence gaps is not a novel concept. It was first proposed in 1938 for the Time Assignment Speech Interpolation (TASI) system that was implemented to improve the capacity of the voice-only transatlantic telephone communications cable system by statistically multiplexing more than one voice call onto a cable [Bullington & Fraser, 1962] (cited in [Stern & Sobol, 1995]). The advent of digital systems produced an improved digital form of TASI known as Digital Speech Interpolation (DSI) [Lyghounis et al, 1974].

Various protocols for voice/data integration have been proposed that can be implemented on the existing 25/30kHz TDMA/FDMA land mobile systems in the 450MHz and 800MHz bands (e.g. [Stern & Sobol, 1995] and the references cited therein). These systems propose to utilise the silent periods during speech to transmit data for mobile facsimile, automatic vehicle location and low bit rate mobile data communications. Figure 5.2 illustrates the concept of data insertion into a voice stream [Stern & Sobol, 1995]. Data messages might first be packetised such that insertion into smaller silence gaps is possible. This allows a single message to be transmitted during several non-continuous silence gaps (e.g. msg #3).

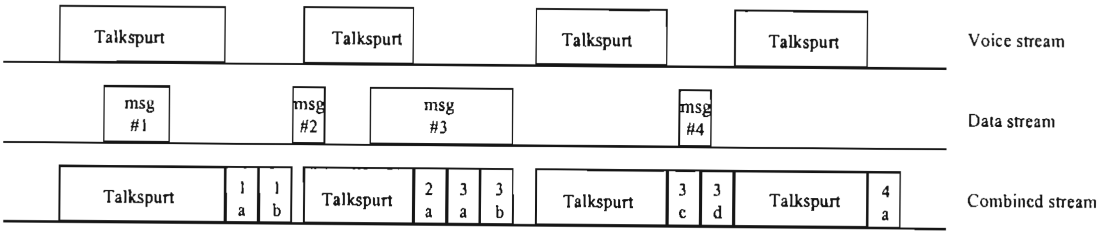


Figure 5.2 : Transmission of data messages during silent periods in speech.

PDVD lends itself well to random access protocols and DAMA techniques since the traffic profile of voice is made to appear like bursty data traffic. Intuitively however, these service multiplexing systems are not efficient in large area radio networks or large footprint satellite systems (multipoint-to-point random access uplink only) with long propagation delays. A node may be able to exploit the silent periods of its own voice stream, or those of a nearby node, but will not be able to sense and exploit the silent gaps in nodes further away due to the propagation delays involved.

5.3 Wideband TDMA Frame Protocols for Voice/Data

The most common schemes for voice/data integration in application today are those based on the demand assignment and reservation MAC protocols discussed in section 2.6. In these protocols, time is invariably divided into fixed-length slots and slots are grouped into TDM frames. The exact nature of the integration process can then be classified according to three main frame utilisation schemes.

5.3.1 Packet Reservation Multiple Access (PRMA)

PRMA is a form of R-ALOHA that has been designed such that voice and data can share the TDMA frame structure [Goodman *et al*, 1989]. Information in the network is categorised as being either *periodic* (such as speech or long data messages) or *random* (short data messages, signalling or control information). In PRMA, the silence gaps of speech are exploited and each talkspurt is treated as an independent periodic message. Sources with messages to transmit contend for available slots using the S-ALOHA protocol, exactly as in R-ALOHA (refer to figure 2.3).

Upon successfully transmitting in a slot, a terminal with a periodic message reserves that slot for uncontested use in all subsequent frames until message completion. After message completion, the terminal releases the slot for available use. Each talkspurt in the speech stream is required to contend for available slots and is thus susceptible to waiting delays. PRMA handles this by employing a maximum allowable waiting delay, after which time talkspurt packets that have been buffered for too long are discarded from the system. Terminals with random messages are not able to reserve slots in the frame. Random messages that are several packets long must contend for slots for each packet in the message. Comprehensive analyses of PRMA appear in [Goodman *et al*, 1989], [Goodman & Wei, 1991] and [Nanda *et al*, 1991].

Several extensions to the classical PRMA protocol have been proposed recently. Examples are the Dynamic-PRMA (DPRMA) scheme and the Centralised-PRMA (CPRMA) scheme (reviewed in [Akyildiz *et al*, 1999]). These extended protocols all follow basically the same TDMA frame structure and reservation request procedure of PRMA, but differ in the manner in which requests are made and resources are allocated. In DPRMA for example, time slots are allocated and granted according to the amount of bandwidth required by each user. A user submits a request for a specific bit rate using the standard PRMA reservation request scheme. Real-time traffic (voice) users' requests have higher priority than the non-real time (data) users'. After the contentions have been resolved, the base station allocates as much of the user's requested bandwidth as possible, and alerts each user of the slots in which they are to transmit. Since the time-dependent real-time services have the higher priority, non-real time data users are placed in a queue where they await service until the bandwidth becomes available. CPRMA is similar to DPRMA, although requests for bandwidth are granted based on the urgency of each user to transmit. CPRMA maintains a request schedule, which is ordered and managed, and to which the resource is allocated, in a manner that attempts to maintain a certain quality of service according to the delay, packet loss rate and bandwidth requirements of each user.

5.3.2 Derivatives of Robert's reservation scheme for voice/data integration

As with PRMA and R-ALOHA, the concept behind Robert's Reservation can also be extended to voice/data integration. Instead of contending for available slots directly using S-ALOHA (as in the PRMA schemes), contention and reservation requests are performed in the reservation subchannel (figure 2.4). Alternatively, the PRMA concept of periodic and random messages can also be applied, where periodic messages contend in the reservation subchannel minislots whereas random messages contend directly for available data subchannel slots using S-ALOHA. The latter alternative is more efficient in that random messages do not burden the load on the reservation subchannel.

In PRMA and Robert's Reservation voice/data integration schemes, all information slots are shared with equal priority between both periodic and random message types. It is thus possible for all slots in the frame to be reserved by periodic users, with no available slots for random message types. Numerous protocols for voice/data integration have been proposed for LANs, wireless LANs and common-bus networks based on hybrid PRMA/CSMA/ALOHA/TDMA schemes with and without PDVD. Some examples are: [Chen & Li, 1989], [Jangi & Merakos, 1994] and, more recently, [Jeon & Jeong, 1998].

5.3.3 Movable frame boundary techniques

In movable boundary techniques, a fraction of the slots in the frame are allocated to voice traffic only, and the remainder are allocated to data traffic only. The two service types are thus time division multiplexed in a frame environment, with each part of the frame being further allocated among users. The boundary that divides the two traffic types may be fixed or moving, depending on the protocol. Two separate multiple access protocols can then be applied in each subframe. Figure 5.3 illustrates the basic principle of a divided frame system with suitable multiple access schemes for each service.

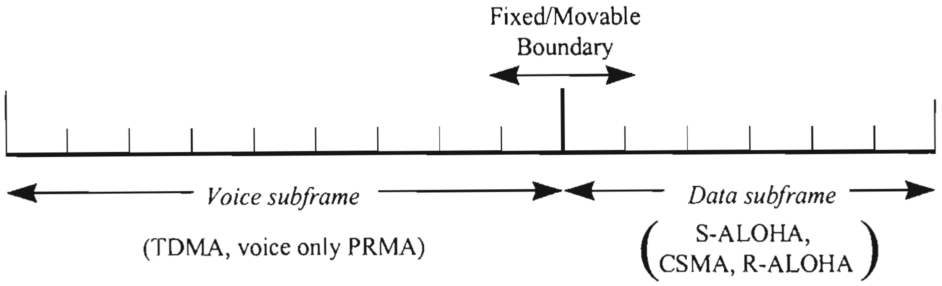


Figure 5.3: Divided frame structure for voice/data integration

For the *voice subframe* (*periodic* sources), the most effective approach would be a TDM based protocol such as TDMA or a DAMA/reservation based protocol such as PRMA, since these protocols best approximate a circuit switched system, i.e. they offer a constant delay and guarantee contentionless access and transmission for the entire call duration. Again, demand assignment schemes are preferred over fixed assignment TDMA as they are more flexible with regards to population changes and load variations. Pure random access protocols are obviously not suitable for the voice subframe, due to the synchronous transmission requirements of voice.

For the *data subframe* (*random* sources with short messages), a random access protocol such as S-ALOHA or *p*-persistent CSMA might be the most applicable, although as discussed in section 2.4, these protocols offer poor system efficiency and are inherently unstable. The DAMA protocols give improved performance for data, but only when the average message size is large enough (i.e. messages become periodic) to exploit the advantages that channel reservation offers [Lam, 1980].

A movable boundary scheme allows for a more flexible system with regards to channel division. Priority is usually always given to voice slots due to the transmission requirements of voice, with data users contending for the remaining slots in the frame. Figure 5.4 illustrates the concept of a movable boundary for a frame containing *L* slots. When the level of voice traffic (*V*) is low, the boundary shifts to allow more slots for data (*L*-*V*-*R*). When a new voice call arrives, the data users have to surrender a slot from the shared section of slots ($L-V-D-R$). A fraction of the bandwidth (*D*) is always reserved exclusively for data users, although voice users get unconditional priority over the first *L*-*D*-*R* slots. If a PRMA type protocol is used, there is no reservation subframe (*R*=0). If a reservation subchannel is used, then *R*>0, typically *R*=1 (with minislots, as shown in figure 5.4).

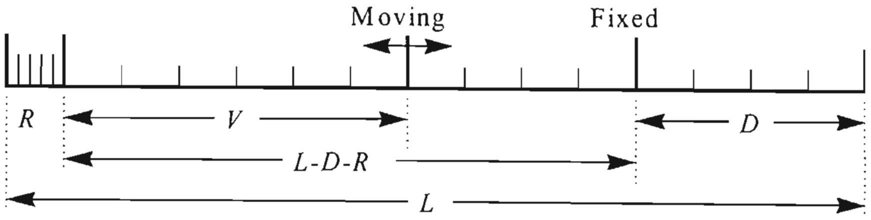


Figure 5.4: Typical moving boundary frame structure

Movable boundary schemes that are combined with demand assignment/slot reservation techniques have been shown to be a very effective means of integrating voice and data and are commonly used in multi-service satellite systems. This is because the large transmission delay (about 270ms for a geostationary satellite system) can be made comparable to the length of a frame, thus negating the negative effect of the delay. Since reservations are invariably made for the following frame, a frame length equal to the round trip delay allows all terminals to be notified of reservation results by the time the requested slot arrives [Wieselthier & Ephremides, 1995]. For a geostationary satellite, a round trip notification delay of 540ms is negligible when compared to the average call holding times of voice conversations (3 minutes). For data users however, constant failure to reserve bandwidth results in end-to-end message delays which far exceed the aggregate data packet length. For this reason, a mixture of demand assignment for voice and S-ALOHA for data (in a movable boundary scheme such as the one in figure 5.4) is usually used for satellite systems with large delays.

Wieselthier & Ephremides [1995] also show that for integrating synchronous traffic such as voice or fixed rate video, the frame length L is best kept fixed to ensure a constant delay with no message jitter. This requirement can be relaxed however through the use of buffering and the inclusing of timing information such that the signal can be accurately reconstructed at playback.

5.3.4 Narrowband wireless ATM MAC protocols

In wireless ATM, each traffic type is distinguished as belonging to one of five possible classes of traffic, as suggested by the ATM forum:

CBR : Constant bit rate - voice and video
RT-VBR : Real-time variable bit rate - compressed voice or compressed video
NRT-VBR : Non-real time variable bit rate - data
ABR : Available bit rate – non time-critical data
UBR : Unspecified bit rate – file transfers, E-mail etc...

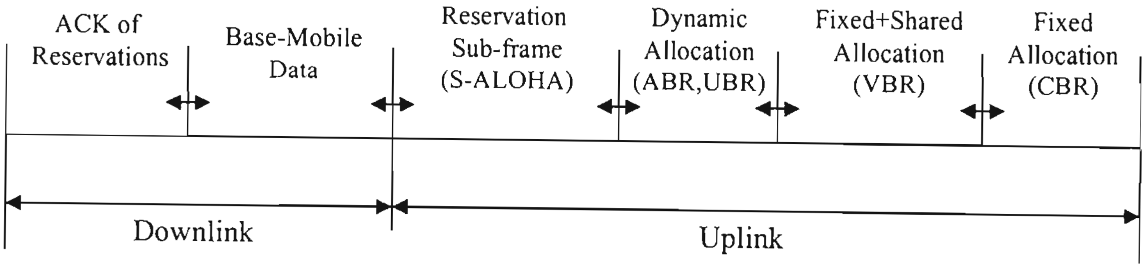

ACK of Reservations
Base-Mobile Data
Reservation Sub-frame (S-ALOHA)
Dynamic Allocation (ABR,UBR)
Fixed+Shared Allocation (VBR)
Fixed Allocation (CBR)
Downlink
Uplink

Figure 5.5 : Frame structure for the DTDMA/TDD Wireless ATM MAC protocol

Several protocols have been proposed that attempt to implement ATM (which was originally designed for a high bandwidth wired network environment) in a wireless environment. An example of one such protocol we will review is the DTDMA/TDD protocol [Raychaudhuri *et al*, 1997]. DTDMA/TDD is based a TDMA frame protocol used in conjunction with a TDD (Time Division Duplex) channel. The basic frame structure for DTDMA/TDD is illustrated in figure 5.5.The entire frame is subdivided into a fixed amount of 8 byte mini-slots, which represent the smallest unit of time. The sizes of the sub-frames that make up the total frame, are variable, as illustrated in figure 5.5. Users request bandwidth by sending request mini-packets in the reservation sub-frame on the uplink frame by using S-ALOHA. The requests are processed at the base station and a schedule table is constructed, based on the QoS requirements of each traffic type. The acknowledgement and scheduling data is broadcast in the downlink frame.

The CBR traffic only requires a single allocation since, once allocated, the user has exclusive use of that portion of the frame until completion of the message. If the CBR subframe is full, then further arriving CBR calls are blocked. The VBR subframe slots are allocated according to a Usage Parameter Control algorithm [Raychaudhuri *et al*, 1997], which is a complex statistical multiplexing scheme that is beyond the scope of this review. Like CBR, the VBR slot allocations are fixed, but the unused slots in the VBR subframe may be shared with other traffic types. Arriving VBR calls are also blocked when the VBR subframe is full. The allocation of CBR and VBR traffic can be viewed as being circuit switched in nature, since once slots have been allocated, users have exclusive use of them until call completion. The ABR and UBR slots are allocated according to a dynamic reservation scheme, and are inserted into the remaining slots of the uplink subframe, wherever possible. ABR and UBR traffic can be considered to be dynamic (packet) switched in nature. This DTDMA/TDD scheme, although fairly complex, has been shown to be a promising MAC protocol for wireless ATM, and is discussed in detail in [Raychaudhuri *et al*, 1997].

Other examples of similar frame structures for wireless ATM, which we will not review in this literature survey, include MASCARA, DQRUMA, DSA++ and PRMA/ATDD. For an excellent summary and comparison of these protocols, we refer the reader to [Sanchez *et al*, 1997] and [Akyildiz *et al*, 1999].

5.4 Multimedia and Voice/Data CDMA MAC Protocols

Since CDMA has only recently become a popular form of multiple access, significantly less papers have addressed integrating voice and data users in CDMA systems than for narrowband TDMA/PRMA based systems. CDMA and spread-spectrum offer several advantages over narrowband systems that make it appear a very viable solution for multi-service radio networks. Firstly, CDMA allows uncoordinated access to the channel with only a slight degradation in signal performance during multiple transmissions. Secondly, as a result of this immediate transmission, end-to-end delays are minimised. Thirdly, several service types can effectively be thought of as separate entities, each appearing almost transparent to the other due to the low-cross correlation of the individual signals and each only affecting the other during periods of high MAI.

In CDMA systems supporting different traffic types, the major limiting factor is that this total number of users allowed is limited to the traffic type with the most stringent bit error specifications. In other words, if a certain data packet demands a worst case bit error rate of say 10^{-6}, then the maximum number of other transmissions allowed during transmission of that packet (regardless of what traffic types they may be) is limited to that which guarantees a BER less than 10^{-6}.

We will begin by reviewing some of the early work on CDMA protocols for integrating voice and data, and end off with a discussion of some of the latest CDMA MAC proposals in consideration for implementation in third generation IMT-2000 and/or UMTS.

5.4.1 The pure spread-spectrum ALOHA (SS/ALOHA) approach

The simplest type of CDMA voice/data integration protocol is discussed in [Jamalipour *et al*, 1995]. Both voice and data use contentionless SS/ALOHA and transmit immediately when they have a ready message, and for as long as they need to until the message is complete. For example, in [Jamalipour *et al*, 1995], all information is transmitted as fixed length packets, although the system is not time slotted. Data traffic is transmitted as bursts of information at the maximum CDMA channel rate. The length of a data burst is unconstrained, and any packets in the burst which are corrupted are retransmitted repeatedly until success using a standard ARQ with time-out scheme. Low bit rate voice traffic is transmitted as a periodic sequence of packets at the peak CDMA channel rate. In each voice sequence, all bits accumulated since the last sequence are transmitted, and the duty cycle is adjusted to match the real-time constant bit-rate requirements of voice traffic. Voice packets that are unsuccessful are not retransmitted and are discarded from the system.

This type of scheme is efficient at low channel load values but suffers at high loads due to MAI. No provision is made to ensure that the number of transmissions remains below some the multiple access threshold, $K_{\max}$, to ensure a certain BER. For the voice network, no blocking of voice calls can occur but the high probability of packet dropout at high levels of MAI causes the system to become inefficient. For the data network, the system performance can be analysed using a model similar to the SS/ALOHA model reviewed in section 2.5.1, although messages are of variable length and the combined traffic arrival process is not Poisson due to the periodic nature of the voice transmissions.

5.4.2 Dual threshold voice/data MAC model with movable "code" boundary

Geraniotis, Yang & Soroushnejad produced a series of papers [Yang & Geraniotis, 1994], [Geraniotis, Soroushnejad & Yang, 1995], [Soroushnejad & Geraniotis, 1995] on access policies based on a two-service CDMA multiple access threshold model with a movable boundary scheme in the code domain for voice/data integration. They improve on the unconstrained SS/ALOHA scheme mentioned in section 5.4.1 by regulating channel access such that the MAI capacity of the system is not exceeded.

We present details of the policy proposed in [Soroushnejad & Geraniotis, 1995] as it is most applicable to the work this thesis presents. The other two aforementioned papers by this group of authors consider a forward link protocol only and a hybrid satellite/terrestrial scheme.

As in the uncoordinated pure SS/ALOHA case, voice and data users share a common spread-spectrum channel for simultaneous transmission. The system is time slotted and both voice and data traffic are assumed to use the same bit rate and spreading technique. Voice call lengths are assumed to be geometrically distributed and are transmitted over many contiguous time slots, whereas all data messages are assumed to be of a fixed size equal to exactly one slot. Two CDMA multiple access thresholds, $K_{\max}^{v}$ and $K_{\max}^{d}$, are defined as the maximum number of allowed simultaneously transmitting voice and data users, respectively, such that the expected packet error probabilities are below pre-specified values,

$$\begin{aligned} P_{Err}^{pkt}(k) &\le P_{\max}^{v} & \forall k &\le K_{\max}^{v} \\ P_{Err}^{pkt}(k) &\le P_{\max}^{d} & \forall k &\le K_{\max}^{d} \end{aligned} \qquad \textbf{(5-1)}$$

where $P_{\max}^{v}$ and $P_{\max}^{d}$ are the maximum acceptable packet error probabilities for voice and data respectively, and k is the number of transmitting voice and data users. As is common practice in telecommunications systems, $P_{\max}^{v} > P_{\max}^{d}$ due to the nature of the transmission requirements for each service. We thus expect $K_{\max}^{v} > K_{\max}^{d}$. With respect to these MAI thresholds, the following traffic admission policies are proposed by Soroushnejad & Geraniotis [1995].

The population of voice users is assumed to be finite, and each voice user uses a unique and pre-specified PN code. All new voice calls are accepted if the level of voice calls (defined as v) is below $K_{\max}^{v}$. As soon as v exceeds $K_{\max}^{v}$, all new voice calls are blocked. The voice admission policy is completely independent of the number of transmitting data users. Voice traffic thus has unconditional priority over data traffic with regards to channel access, due to the real-time requirements and relative length of voice calls. Once a voice user has been admitted, it secures the channel until completion of the entire voice call (a random number of slots which is geometrically distributed). Voice packets which are corrupt due to MAI are not retransmitted and are effectively discarded. Besides the necessary set-up delay, no other delay is suffered by voice users.

As with the voice population, the population of data users is assumed to be finite and each data user uses a unique PN code. The data population consists of two sub-populations: those data terminals which are idle and those which are backlogged. All idle data terminals are assumed to generate a new data message in a system slot with probability ρ_d, i.e. the channel access protocol for data users is the slotted-SS/ALOHA protocol with Delayed First Transmission (DFT). Terminals which are in the backlogged state (i.e. they have a data packet to transmit) are assumed to be incapable of generating new data messages. No buffering of packets takes place at data nodes. Under this protocol, all idle terminals with newly generated packets do not immediately transmit, but rather immediately join the backlog pool before their first transmission is attempted. Data packets which are corrupted by MAI are retransmitted according to a suitable CRA. In [Soroushnejad & Geraniotis, 1995], a backlog dependent retransmission CRA with backlog dependent retransmission probability, ρ_d, is assumed.

As mentioned previously, voice has unconditional priority over data traffic. In assuming the threshold definitions in equation 5-1, we see that the remaining capacity available to data traffic is dependent on the level of voice traffic in the system (v). The threshold value $K_{\max}^d$ for data is effectively the maximum number of allowable data users when the level of voice traffic v is zero. If v>0 voice users are transmitting, then the maximum number of allowed data users, β, is

$$\beta = \max(0, K_{\max}^d - v) \tag{5-2}$$

Obviously when v exceeds $K_{\max}^d$, there will be no system capacity available for data traffic. Since v is a time varying process (as voice users enter and leave the channel), we expect β to be a time varying process as well. In order to regulate the level of data access such that the data threshold β is not exceeded and that the number of data terminals in the backlogged state (defined as b) remains in a preferable stable state, the packet (re)transmission probability ρ_d is adjusted according to the following relationship: when β is high (data capacity is large) or b is small (small backlog population), then ρ_d is large to allow a higher channel access rate. When β is low or b is high (large backlogged population), then ρ_d is small to ensure a low channel access rate. At the end of slot t, each data node adjusts ρ_d according to global feedback information about the number of transmitting voice users (v), and the number of data users in the backlogged state (b). We do not present further details on the exact formula for ρ_d or the feedback schemes, but instead refer the interested reader to [Soroushnejad & Geraniotis, 1995] for a full system analysis and results.

This scheme is almost analogous to the movable boundary scheme in the time domain (section 5.3.3) although, in this CDMA scheme, capacity boundaries are defined in the code domain. The multiple access capacity $K^v_{\max}$ is analogous to the number of time slots in a system time frame. Of the $K^v_{\max}$ time slots, a portion of them ($K^d_{\max}$) are shared by both voice and data users (although voice users get unconditional priority for slot usage) while the remainder ($K^v_{\max}$ - $K^d_{\max}$) are available for voice only. Whereas a voice user will use the same slot number over several contiguous time frames, a data user only requires the use of a slot for one time frame.

5.4.3 Wideband-CDMA (W-CDMA)

The WCDMA MAC protocol is a recent protocol that has been proposed by ETSI (European Telecommunications Standards Institute) as a viable candidate for use in the IMT-2000 system (see [Akyildiz *et al*, 1999]). WCDMA is based on a complex set of physical layer transmission formats, and allows for the possibility of transmissions of different bit rates, out of a discrete set of allowed bit rates which are set multiples of some predetermined rate. The MAC protocol essentially consists of a shared channel controlled by a base station, where channel access is controlled via a reservation request scheme. In WCDMA however, there is no time frame element involved.

In WCDMA, a mobile terminal with a short data packet to transmit does so immediately by transmitting on a shared random access channel (RACH) using a SS/ALOHA scheme. If it has a longer data message to transmit, then a request is made, also via the

RACH channel for what is known as a DCH (Dedicated Channel). Each DCH is associated with a dedicated spreading code and specific transmission format. The set of transmission formats contains information on the type of coding technique and modulation scheme to be employed by the user on its DCH. Upon evaluation of each user's request, the network notifies the user of the DCH code to be used, bit rate to be used and transmission format to be used. The data user then begins transmitting contiguously on that particular DCH using the specified transmission format and bit rate.

During the transmission, the network is able to change the user's bit rate and transmission format, depending on the load on the network. For example, under high load conditions, the network may drop the user's bit rate. The rate control scheme and choice of transmission format are chosen in a manner which attempts to support the required BER (which is effectively the main measure of QoS in any pure CDMA system) requirements of each traffic type in the channel.

For the data users, once they have completed transmitting on a DCH they maintain the link for a specified amount of time. If they generate a new message within this time, they may transmit immediately without having to make another reservation. If no message is generated within this time, the link is terminated. This has been done under the assumption that the time between data messages is usually either very long (e.g. between server requests) or very short (short time durations between exchanges in information, handshaking etc...). In maintaining the link, several data messages that follow each other closely do not have to repeatedly make reservations. For real-time users (voice, video), the request scheme is similar to the long data message case, although the terminal does not maintain the DCH upon completion of the data message.

The exact reservation, allocation scheme and set of transmission formats are complex and, as of yet, have not been completely defined yet [Akyildiz *et al*, 1999].

5.4.4 WISPER

The Wireless Access Control Protocol with Bit Error Rate Scheduling (WISPER) is a TDMA/CDMA protocol that is designed, using the power control characteristics of the IS-95 protocol standard, to ensure that the channel operates below a predefined BER. The idea behind employing both TDMA and CDMA is to exploit the capacity of TDMA's ability to support high bit rates while, at the same time, exploiting CDMA's ability to allow a smooth coexistence of different multimedia traffic types. In WISPER, a TDMA frame structure is implemented, where each frame is divided into slots. One of the slots in each frame is allocated as a reservation slot, while the other slots are allocated as transmission slots. Each slot can carry any type of multimedia traffic type, and several traffic types can transmit simultaneously using CDMA. The aim of the reservation slot and associated resource allocation algorithm is to ensure that the BER that occurs in any slot is suitable for all those traffic types transmitting in that slot. In other words, the traffic type with the most stringent BER requirements determines the BER allowed in that slot. As in most CDMA systems, the primary QoS measurement is the BER.

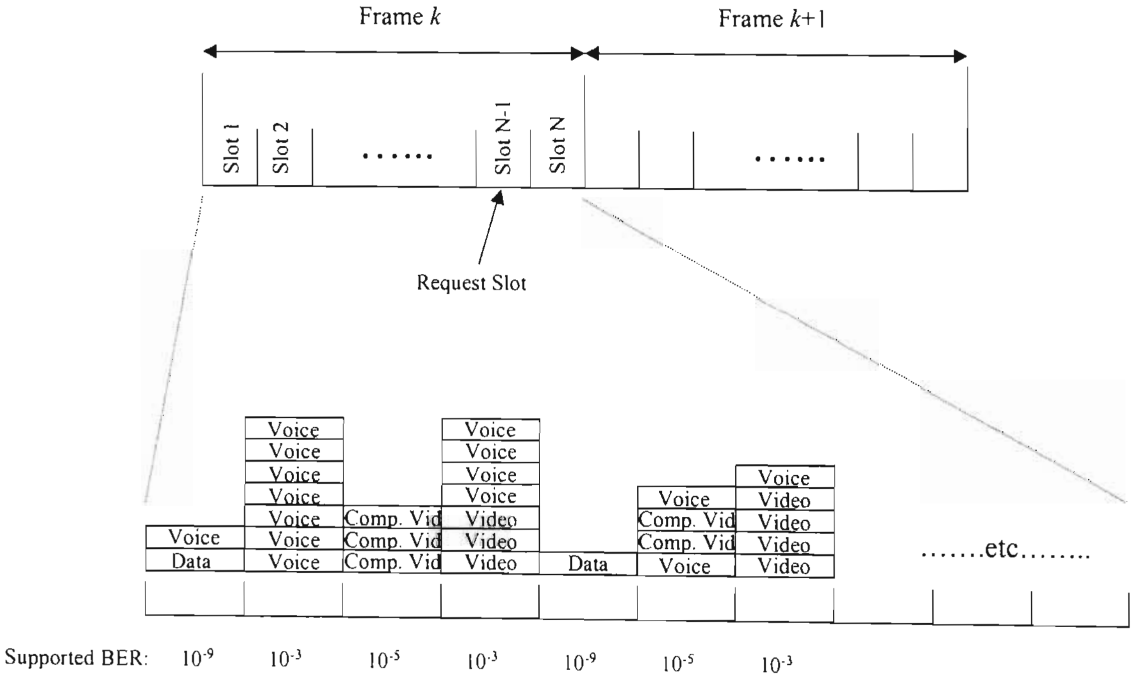


Figure 5.6: Example of the uplink frame structure used in WISPER (obtained from [Akyildiz *et al*, 1999])

Figure 5.6 illustrates the WISPER frame and traffic allocation concept for four traffic types: data (which requires a BER of 10^{-9}), video (which requires a BER of 10^{-3}), voice (which requires a BER of 10^{-3}) and compressed video (which requires a BER of 10^{-5}). Slots with less stringent BER requirements will be able to support more transmissions.

Other noteworthy hybrid TDMA/CDMA multimedia MAC protocols that have been proposed recently for application over the UMTS are [Brand & Aghvami, 1998] and [Prasad *et al*, 1997]. These papers propose a hybrid PRMA/CDMA and a hybrid TDMA/CDMA scheme respectively. In [Azad *et al*, 1999], a multimedia pure CDMA MAC strategy is proposed that considers a multi-rate system.

In addition to the CDMA protocols discussed above, there are several other schemes that are currently under the process of being standardised. An example is the CDMA-2000 protocol standard which is based on the evolution of the IS-95B and IS-95C standards. CDMA-2000, like W-CDMA, is also a candidate for the IMT-2000 system.

5.5 Summary

In conclusion, we have discussed several methods in which voice and data can be integrated onto a common channel. We discussed a method in which data is statistically multiplexed into the silent periods that occur during natural human speech. We then reviewed the class of voice/data integration protocols based on demand assignment and channel reservation techniques, as well as similar divided frame and movable boundary techniques in the time domain. We concluded by reviewing some of the more recent protocols based on multimedia support in a CDMA environment.

CHAPTER 6

VOICE AND DATA INTEGRATION IN DT/CDMA

6.1 Introduction

In this chapter we extend the proposed data-only network protocol of Chapter 4 to an integrated voice/data network. The proposed system is derived from the integrated traffic concept first presented in the series of papers by Geraniotis, Soroushnejad and Yang ([Yang & Geraniotis, 1994], [Geraniotis, Soroushnejad & Yang, 1995] and [Soroushnejad & Geraniotis, 1995]). The fundamental differences between our scheme and those presented in the aforementioned papers are: (1) We implement our centralised SR-DT/CDMA MAC protocol as opposed to their S-SS/ALOHA based protocol, and (2) We exploit and model the voice activity factor of a voice call as opposed to just the call initiation/termination statistics.

The layout of this chapter is as follows: In section 6.2 we outline the integrated services network architecture, traffic models, and MAC policies for each traffic type. As a result of the extension to a mixed traffic scenario, the analytical model of the network performance doubles in complexity, and requires extensive modifications. In section 6.3, we derive a Markov model to analyse the performance of the voice network performance. In section 6.4, we derive a joint-state Markov model to analyse the performance of our SR-DT/CDMA data MAC protocol in the presence of voice traffic. In section 6.5, we present the results of our Markov models along with simulation results, and discuss the efficacy of our integrated traffic scheme.

6.1.1 Justification for the proposed model

We choose to study CDMA voice/integration schemes as we strongly feel that CDMA and spread-spectrum offer many advantageous benefits when considered for cellular packet radio networks (section 2.2). In addition, at the time of commencement of this project, a literature review revealed that, although much work had appeared on CDMA protocols for single service voice only or data only networks, less work had appeared on multi-service integration schemes of the nature proposed in this thesis. Those that had, assumed idealised conditions, simple traffic models or relatively inefficient access protocols. Since commencement of this project, CDMA has become a popular choice for third generation telecommunications networks, and many multi-service protocols have been presented in the last two years, as reviewed in Chapter 5.

6.2 System Model

In this section, we present the complete proposed system model. The network architecture is imagined, as are the population models, traffic models and multiple access policies.

6.2.1 Network architecture

The network architecture under consideration remains unchanged from that outlined in section 3.2, apart from the fact that there is now a homogeneous population of voice and data terminals. Both voice and data traffic are assumed to operate in the same frequency band and have exactly the same signal properties. This includes the use of the same fixed bit rate, spreading method, modulation scheme, processing gain (*N*), code allocation scheme and information packetisation scheme. This latter point means that all voice calls are also packetised with the same fixed length packet size such that exactly one packet containing exactly *Q* bits is transmitted per slot. Voice and data packets thus appear exactly alike during transmission. The only differences are found in the actual information content of the packets, which may include different packet headers and different error correction/detection schemes.

The CDMA threshold model and centralised channel state broadcast scheme are also as described in section 3.3. The only main and obvious difference is that there are now two traffic types, each with different service requirements and each with different traffic admission policies.

6.2.2 MAC protocol for the voice traffic

The proposed voice admission policy is similar to the "direct admission policy" proposed in [Yang & Geraniotis, 1994]. Their work considers a policy suitable for the forward link however whereas, in our protocol, we are interested in modifying it to suit the reverse link. As was done in section 3.3.2 for the data policy, we impose a voice threshold on the maximum transmission capacity available for voice users, or the maximum number of allowed transmitting voice users. This channel threshold for voice traffic is defined as K, and is imposed such that the following channel conditions are satisfied for voice traffic:

$$\begin{aligned} P_E^{voice}(v) &\le P_{E(\max)}^{voice} && for \quad v \le K \\ P_E^{voice}(v) &> P_{E(\max)}^{voice} && for \quad v > K \end{aligned} \qquad \textbf{(6-1)}$$

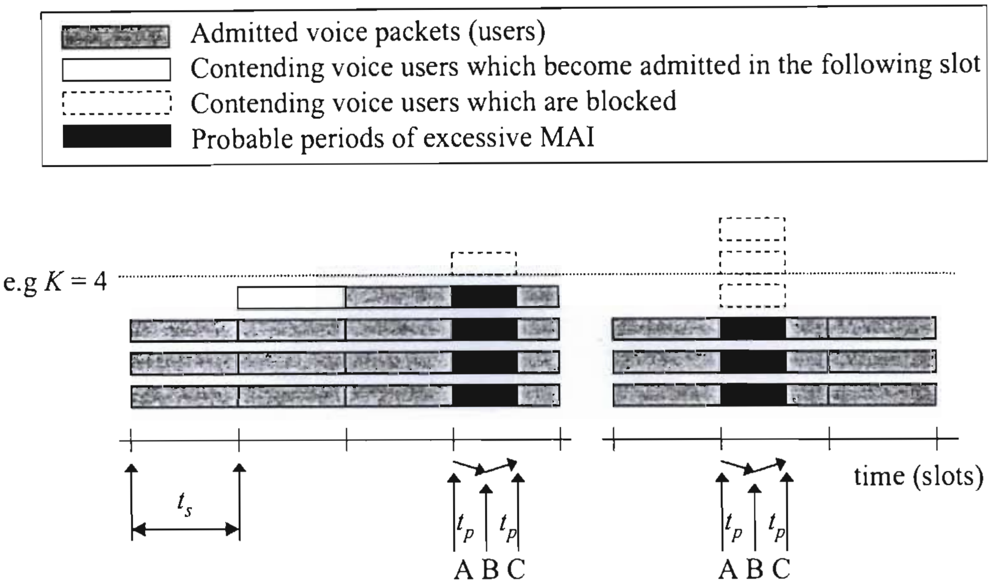
where $P_{E(\max)}^{voice}$ is the maximum acceptable error rate for voice traffic, and $P_E^{voice}(v)$ is the error rate given v simultaneous voice transmissions. As with the data policy, we associate a broadcast tone with respect to K.

- The base station broadcasts a "*Voice Collision*" (**VC**) signal (tone) if the number of voice users exceeds K.

The multiple access admission policy for voice traffic is as follows (refer to figure 6.1 for a graphical description, as well as all timing information regarding tones):

1. An idle voice terminal with a call ready to initiate does so immediately at the start of the current slot, and is known as an *contending voice user*.
2. All voice users that were admitted *before* the current slot, and are in the process of transmitting, are known as *admitted voice users*.
3. All contending voice users in the current slot are admitted if and only if the total number of voice users (i.e. *contending voice users* + *admitted voice users*) does not exceed K. Upon detecting more than K calls in the current slot, the base station emits the **VC** tone.

4. Upon detecting the **VC** tone, only the *contending voice users* abort transmission and are considered to be blocked (having only transmitted for a fraction of a slot). If **VC** is not detected, contending voice users become admitted in the following slot.
5. Once a voice user has been admitted, it effectively secures that "channel" for the entire duration of its call and can never be dropped
6. Voice packets that are corrupted due to MAI are ***not*** retransmitted. There is no information retransmission in the voice protocol.
7. The admission of voice users is dependent only on the level of active voice users in the network. The level of data users does not affect the admission of voice users.


Admitted voice packets (users)
Contending voice users which become admitted in the following slot
Contending voice users which are blocked
Probable periods of excessive MAI
e.g $K = 4$
t_s
t_p
t_p
A B C
t_p
t_p
A B C
time (slots)

A: base station measures load to exceed K and emits **VC** tone
B: terminals detect **VC** tone after delay t_p, and abort transmission
C: base station detects that new calls have aborted after delay 2 t_p, **VC** tone goes off

Figure 6.1 : Illustration of the voice call admission policy

The voice call admission policy is almost exactly the same as the data admission policy for SR-DT/CDMA. The only differences are that there is no secondary blocking threshold, and there is no retransmission of corrupt packets. Justification for these differences, as well as the fact that the voice traffic is given unconditional priority over data to the channel are related to the nature of the traffic arrival processes and the service requirements of each traffic type. We discuss these requirements and differences after first presenting the voice traffic model and data admission policy.

6.2.3 Voice traffic model

Due to the fact that there is no information retransmission in the voice protocol, the analysis for a finite voice population is simple. Hence, the voice traffic in the network is assumed to be generated by a finite population of N_v identical voice terminals. Each terminal operates completely independently of any other voice terminal, and without any knowledge of the state of other terminals in the network. In this thesis, we consider a model in which the *voice activity factor* (discussed in section 5.2) of human speech is exploited to improve the capacity of the shared channel. For this purpose, we define that each voice terminal can exist in one of three possible states: ***idle***, ***silent*** or ***talkspurt***. A voice terminal is in the idle state when it is not being used, i.e. when there is no voice call in progress. The silent and talkspurt states occur when a terminal is being used for a voice call, and a terminal is defined to be active if in either one of these two states (i.e. a terminal is active if it is non-idle). The silent state refers to the fact that a call is in progress, but no speech is being generated by the user. The terminal is in the talkspurt state when the user generates a speech talkspurt. In this thesis, we model a voice terminal's activity using a three state Markov model with voice states and associated state transition probabilities as illustrated in figure 6.2.

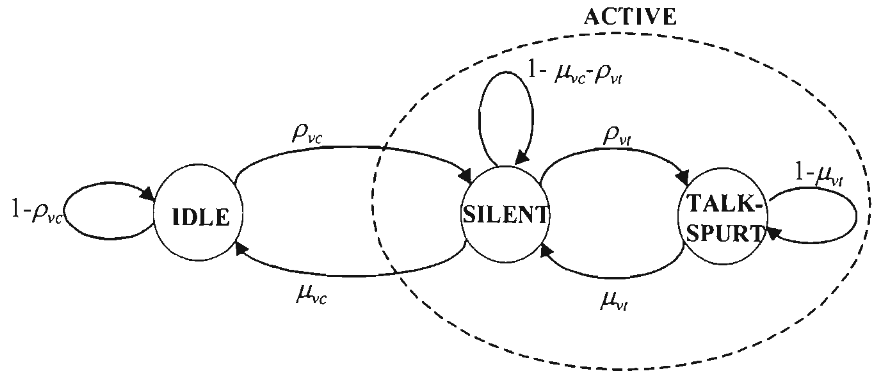


Figure 6.2 : Voice model showing voice terminal states and state transition probabilities

When in the idle state, a terminal attempts to initiate a voice call in the current slot with probability ρ_{vc}, the subscript "vc" denoting a voice call. The probability of an active user terminating a call (becoming idle) in a slot is defined as μ_{vc}. We assume that a terminal can only terminate a call while in the silent state, and that an idle terminal first enters the silent mode before beginning to talk. Since the periods of silence are extremely small on average, this assumption should not be at all unrealistic in the context of a telephony call. When a voice terminal is in the active state, the probability that a silent speaker starts to talk (initiating a talkspurt) in the current slot is defined as ρ_{vt}, where "*vt*" denotes a voice talkspurt. The probability of terminating the current talkspurt is defined as μ_{vt}.

If the voice state variable v is defined as the number of active voice calls in progress (number of non-idle voice terminals), and t is defined as the number of active talkspurts in progress, then the expected value of t/v, which also represents the fraction of time that each voice terminal is in the talkspurt state, or *voice activity factor* (VAF), is given as

$$\begin{aligned} \mathrm{VAF} &= \frac{\bar{t}}{\bar{v}} \\ &= \frac{\rho_{vt}}{\rho_{vt} + \mu_{vt}} = \frac{\mu_{vt}^{-1}}{\mu_{vt}^{-1} + \rho_{vt}^{-1}} \end{aligned} \qquad \textbf{(6-2)}$$

Obviously, the number of talkspurts can never exceed the number of calls in progress, thus $t \leq v$. With respect to this model, we assume the following with regards to the physical signal transmitted. Firstly, the terminal's transmitter is only active during the period when the terminal is ACTIVE and in the TALKSPURT state. During the IDLE or SILENT periods, it is assumed that the transmitted signal is severely suppressed (or perhaps even switched off entirely). The most obvious advantage of doing this is that voice terminals contribute negligibly to the channel MAI during the silent periods in which they have nothing useful to transmit. Exploitation of the voice activity factor to improve the capacity of CDMA networks is nothing new, and has been considered in many works (e.g. [Viterbi & Viterbi, 1993] and [Gilhousen *et al*, 1991]).

In this work, we assume that the base station is always aware of the number of terminals in the active state (v), regardless of the level of speech activity. Thus, even when a terminal is in the silent state, the base station knows that there is a voice call in progress from that terminal. Upon call initiation and call termination, the level of v is recomputed. The opposite is true of the measure of t, which will be fluctuating due to the on-off nature of each terminal's transmitter.

With respect to the call admission policy, we assume that new voice calls that are blocked are assumed not to re-occur as re-attempts at a call initiation. A blocked call is treated as a normal voice call, but with a zero call holding time. This assumption allows us to model the voice call arrival process as a memoryless process that is independent of the previous history of the network.

6.2.4 Data admission policy

The data terminals use the SR-DT/CDMA protocol, as discussed in Chapter 4. The only difference is that there are now voice transmissions in the channel which alter the data thresholds β and α. Whereas data traffic has no effect on the admission of voice users due to the fact that voice calls have unconditional priority to the channels, the admission of data messages depends on the number of transmitting voice users. Data users have to contend for the remaining "channels" left over after voice traffic has been admitted. We recall that the maximum number of transmitting data users allowed was defined as β. We now modify the definition of β slightly, to account for the integration of the voice traffic. Threshold β now represents the maximum allowable *total number of transmitting users* - both voice and data - that is allowed in order to satisfy the packet error criteria for the data traffic:

$$\begin{array}{ll} P_E^{data}(x+t) \le P_{E(\max)}^{data} & \text{for} \quad (x+t) \le \beta \\ P_E^{data}(x+t) > P_{E(\max)}^{data} & \text{for} \quad (x+t) > \beta \end{array} \qquad \textbf{(6-3)}$$

where $P_{E(\max)}^{data}$ is the maximum acceptable error probability for data traffic, x is the number of transmitting data users and t is the number of transmitting voice users (voice users in the talkspurt state). Alternatively, we can define a new "modified" data threshold, β', which represents the maximum allowed number of data users given that t voice users are transmitting:

$$\beta' = \max(0, \beta - t) \qquad \textbf{(6-4)}$$

Obviously, the value of β' depends on the difference between K and β . If K is larger than β , then there is the case that, when $t \geq \beta$, there are no channels available for data traffic. If $K< \beta$, then there will always be channels available for data, but there will be cases where the combined voice and data load (x+t) exceeds the voice threshold K. Ideally, we would want K=β. In section 6.5.4, after having presented the network performance, we will revisit this threshold problem.

We recall that, for a fixed β , the blocking threshold, α, is a function of the offered load. Since β' now varies over time, the blocking threshold must also change with time in order to induce its advantage. To accomplish this, we propose a modified blocking threshold $0 \leq \alpha' \leq \beta'$ which should always be a fixed fraction of β' for a given load. We thus define a threshold ratio ϕ, based on the original thresholds employed when there are no voice calls, such that

$$\phi = \frac{\alpha}{\beta} \qquad \textbf{(6-5)}$$

For any data collision threshold β' , an associated blocking threshold α' can be assigned to provide the optimum throughput at a given offered load. The channel information broadcast scheme (as defined in section 3.3.2) is now modified as follows, to account for the change in data thresholds:

- The **DB** tone is emitted if the data load x exceeds α' (or alternatively, the total load $x + t$ exceeds $\alpha' + t$)
- The **DC** tone is emitted if the data load x exceeds β' (or alternatively, the total load $x + t$ exceeds β)

The six protocol steps in sections 3.3.3 or 4.2 can then be applied to the above broadcast/admission scheme, but using α' and β' instead of α and β. Throughout the remainder of this chapter, we will focus on the SR-DT/CDMA protocol due to the fact that it is more efficient than the DT/CDMA protocol.

6.2.5 Data traffic model

We also choose to focus on the infinite population data traffic model rather than the finite population model, with the following justifications: Firstly, results are easier to obtain for the infinite population model due to the fact that we don't have to solve many iterations of the analysis in order to obtain the steady state solution. Secondly, and more importantly, the infinite population analysis is exact and as a result yields results which are far closer to the simulation results than the finite population analysis could provide. This makes it easier to judge the accuracy of our analysis when we integrate voice and data traffic in the same channel. The data traffic model used in the remaining analysis is thus the Poisson arrival process with geometrically distributed message lengths, as summarised in section 3.4.1.

6.2.6 Voice has unconditional priority over data

After reviewing the voice and data traffic admission policies, we are now in a stronger position to justify the decision to grant unconditional priority to voice traffic. The fact that voice traffic requires real-time delivery while data traffic can tolerate delays and non-continuous transmission provides one justification. Perhaps more important though is the relative lengths of the two messages types. It seems intuitively unfair to block a three minute voice call based on a few slots (a fraction of a second) in which there may be a temporary peak of data traffic, when instead, the voice call can be admitted and the data traffic be rescheduled for retransmission after a short delay.

In our model, the quality of the voice signal depends entirely on the number of users that are transmitting at that time. If the number of users is low, then the bit error rate is low and the listener observes a higher degree of signal quality. Poorer quality is experienced during periods of high channel load. The poorest signal quality occurs at points where the K threshold is exceeded. This however, only occurs for a fraction of a slot at a time and, as illustrated in figure 6.1, only a fraction of the admitted packet (darker region) is subjected to a higher bit error rate. In the event of such a voice collision, it seems fruitless to abort all voice calls when the only noticeable adversity may be, for example, a short degradation in signal quality, an audible click or short loss of signal.

As discussed in section 1.1.1, voice information does not require retransmission due to the fact that even a substantial proportion of bits being corrupt does not significantly affect the listener's capability to interpret the information. As such, we choose to follow the practice of existing wireless cellular protocols, and avoid retransmitting corrupt voice packets.

The second major difference between the voice admission scheme and the data admission scheme of SR-DT/CDMA is the omission of the blocking threshold, α, in the voice protocol. Justification for this omission is based on the relative arrival rates and relative message lengths of the two traffic types. The purpose of the α threshold is to decrease the frequency of collisions. Since the data protocol is designed to operate in an environment in which terminals generate short messages at a high rate, the expected number of messages attempting to access the channel per slot is very high.

In the voice network however, although the number of simultaneous voice transmissions may be equal to that of the data network, the number of voice call attempts per slot is very low. This is due to the fact that the mean length of voice calls is several orders of magnitude larger than that of data messages, as is the mean duration of the idle times between calls (for example, the data message retransmission delay used in Chapter 4 was only 5 slots). Since voice collisions (K crossings) are expected to be very few and far between, the function of a blocking threshold which is less than K will have almost negligible effect.

Before beginning the analysis of the integrated traffic model, we first provide a reference list of all pertinent network parameters (other parameters are as listed in section 3.4.2):

N_v : population of voice users in the network

N_d : population of data users in the network

v : the number of voice calls in progress

t : the number of talking voice users

s : the number of silent voice users (active but not talking), $s = v - t$

x : the number of transmitting data users (state of the data network)

K : voice multiple access (collision) threshold

β : maximum data collision threshold (no voice case)

α : data blocking threshold for the no-voice case $(0 \leq \alpha \leq \beta)$

ϕ : the ratio of α / β for a given load value

β' : number of channels remaining for data $\beta' = \max(0, \beta - t)$

α' : modified data blocking threshold for the voice case $(\alpha' = \phi\beta')$

λ : expected offered data load per second

G_d : expected offered data load per slot

ρ_{vc} : probability of an idle voice terminal initiating a call in the current slot

μ_{vc} : probability of a voice call terminating a call in the current slot

ρ_{vt} : probability of a silent voice user starting a talkspurt in the current slot

μ_{vt} : probability of a voice user terminating a talkspurt in the current slot

μ_d : probability of a data user completing a message in the current slot

Since voice calls have unconditional priority over data messages, the presence and admission of data messages has no effect on the admission of voice messages or the state (number of transmitting or active voice terminals) of the voice network. The only effect that the data network has on the voice network is that transmitting data terminals contribute to the level of MAI experienced by voice users and thus reduce the quality of the voice transmissions. Because the admission policy and state of the voice network are completely independent of the level of data traffic, we can analyse the voice network by itself.

6.3 Voice Protocol Performance Analysis

For the given voice admission policy, the main performance measurement of interest is the expected voice call blocking probability. Since voice talkspurts are transmitted as soon as they are generated, and are transmitted as a continuous stream of information, there are no delay or information "jitter" issues for the given protocol.

6.3.1 Definition and computation of the voice network state statistics

We begin by defining the states t and s as the current numbers of talkspurt and silent terminals in the network, respectively. Furthermore, we define state $v = t + s$ as the number of voice calls in progress. We also define the state occupancy probabilities $P\{t = j\}$ and $P\{s = l\}$ as the steady state probabilities of being in states $t = j$ and $s = l$ respectively. The joint probability $P\{t = j, s = l\}$ then represents the probability of having $v = j + l$ voice calls in progress. Since we can never have more than K calls in progress due to the admission/blocking scheme, we can never have the case $t + s > K$ and thus the number of possible joint states is equal to $1 + 2 + 3 + \ldots + (K + 1)$. The joint state transition probabilities for the joint system state make up a four-dimensional matrix, where element $P\{t_{n+1} = j, s_{n+1} = l | t_n = i, s_n = k\}$ represents the probability of having i talking users together with k silent users in the previous slot n, and j talking users together with l silent users in the current slot n+1. If all $P\{t_{n+1} = j, s_{n+1} = l | t_n = i, s_n = k\}$ are known, then the set of joint state probabilities can be found by solving the set of linear equations in

$$P\{t = j, s = l\} = \sum_{i=0}^{K} \sum_{k=K-i}^{K} P\{t = i, s = k\} . P\{t_{n+l} = j, s_{n+1} = l | t_n = i, s_n = k\} \qquad \textbf{(6-6)}$$

while maintaining the Markovian conservation property

$$\sum_{j=0}^{K}\sum_{l=K-j}^{K} P\{t=j,s=l\}=1 \qquad \textbf{(6-7)}$$

In order to simplify the analysis notation, we will define $\Pi_{ik,jl}=P\{t_{n+1}=j,s_{n+1}=l|t_n=i,s_n=k\}$. To avoid confusion, we use a capital Π for use with the voice protocol, and continue using a small π for the data protocol. The value of the joint state transition probability $\Pi_{ik,jl}$ can be easily found by considering the flow of the number of users into and out of each state.

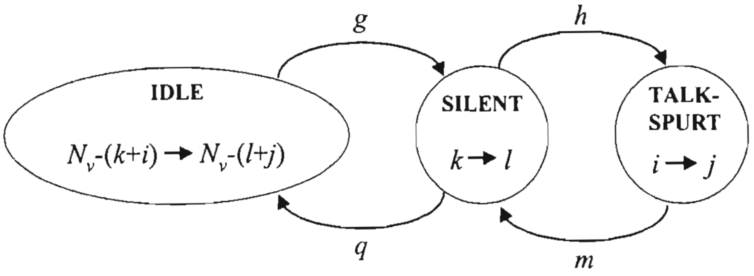


Figure 6.3 : Illustration of flow of users into and out of the different voice states

In figure 6.3, we illustrate the three terminal states and the number of users flowing into each state from the other states. In figure 6.3,

i = the number of talking users in the previous slot
h = the number of silent users beginning a talkspurt
m = the number of terminating voice talkspurts
j = the number of talking users in the current slot = $i-m+h$
k = the number of silent users in the previous slot
q = the number of terminating voice calls
g = the number of new voice calls
l = the number of silent users in the current slot = $k+m+g-h-q$

The number of voice calls in the previous slot is given by (k+i) and the number of voice calls in the current slot is given by (l+j). When computing the value of element $\Pi_{ik,jl}$, we must take cognisance of the admission policy and blocking scheme. The simplest case occurs when the number of voice calls in the current slot is larger than that in the previous slot, i.e. $(i+k)<(j+l)$. Obviously, no blocking can have occurred since the system does not block a fraction of new calls. *All* new calls are admitted if channels are available, or *all* are blocked if channels are unavailable.

For the cases when $(i+k)<(j+l)$ and $(i+k)\leq K$ and $(j+l)\leq K$, the elements $\Pi_{ik,jl}$ can be shown to be equal to

$$\Pi_{ik,jl}=\sum_{q=0}^{k}\sum_{m=i-j}^{i}b(q,k,\mu_{vc}).b(m,i,\mu_{vl}).b(h,k-q,\rho_{vl}).b(g,N_v-k-i,\rho_{vc})$$

$$\text{for all } (i+k)<(j+l) \text{ and } (i+k)\leq K \text{ and } (j+l)\leq K \qquad \textbf{(6-8a)}$$

In equation 6-8a, for any given q and m, the values of g and h are given by $g=l-k+q+j-i$ and $h=j-i+m$. Any illegal probabilities are avoided in equation 6-8a in that the binomial function defined in equation 3-4 precludes the use of any negative parameters.

If the number of voice calls in the current slot is less than or equal to that in the previous slot, then there is a non-zero probability that a collision and subsequent blocking occurred. This collision only occurs when the number of voice calls that terminate in the current slot is exactly $(k+i)-(l+j)$, and the number of new voice calls exceeds $K-(l+j)$. For the cases when $(i+k)\geq(j+l)$ and $(i+k)\leq K$ and $(j+l)\leq K$, the elements $\Pi_{ik,jl}$ can be shown to be equal to

$$\Pi_{ik,jl}=\sum_{q=(k+i)-(j+l)}^{k}\;\sum_{m=i-j}^{i}b(q,k,\mu_{vc})b(m,i,\mu_{vl})b(h,k-q,\rho_{vl})b(g,N_v-k-i,\rho_{vc})$$
$$+\sum_{g=K+1-(j+l)}^{N_v-k-i}\;\sum_{m=i-j}^{i}b(k+i-j-l,k,\mu_{vc})b(m,i,\mu_{vl})b(h,j+l-i,\rho_{vl})b(g,N_v-k-i,\rho_{vc})$$

$$\text{for all } (i+k)\geq(j+l) \text{ and } (i+k)\leq K \text{ and } (j+l)\leq K \qquad \textbf{(6-8b)}$$

Again, $h = j - i + m$ in 6-8b. For the "illegal" values ($i + k > K$) or ($j + l > K$), we must have:

$$\Pi_{ik,lj} = 0 \qquad \text{for all } (i+k) > K \text{ or } (j+l) > K \qquad \textbf{(6-8c)}$$

Once the joint state occupancy probabilities $P\{t = j, s = l\}$ are known, all state occupancy probabilities for the voice network can be found. The stationary state occupancy probabilities for the number of talking users, $P\{t = j\}$, are given by

$$P\{t = j\} = \sum_{l=0}^{K-j} P\{t = j, s = l\} \qquad \textbf{(6-9)}$$

Similarly, the state occupancy probabilities for the number of silent speakers are

$$P\{s = l\} = \sum_{j=0}^{K-l} P\{t = j, s = l\} \qquad \textbf{(6-10)}$$

The probability of having *v* voice calls in progress is simply equal to the probability of having *j* talking users and *v-j* silent users

$$P\{v = n\} = \sum_{j=0}^{n} P\{t = j, s = n - j\} \qquad \textbf{(6-11)}$$

6.3.2 Computation of the expected offered and carried voice loads

In knowing all state occupancy probabilities for the voice network, we can compute the number of users expected to be found in each state. The expected number of active voice calls per slot, or *average carried voice call load* $\bar{v}$, is simply

$$\bar{v} = E\{v\} = \sum_{n=0}^{K} n.P\{v = n\} \qquad \textbf{(6-12)}$$

The *average offered voice load*, G_v, is defined as the expected number of voice call attempts per slot from all idle voice terminals, and is given by:

$$G_v = (N_v - \bar{v}) \cdot \rho_{vc} \qquad \textbf{(6-13)}$$

where the expected idle population size is obviously $N_v - \bar{v}$. The expected number of talking users, or more importantly, the expected number of *transmitting voice users*, is given by

$$\bar{t} = E\{t\} = \sum_{n=0}^{K} n.P\{t = n\} \qquad \textbf{(6-14)}$$

6.3.3 Computation of the expected voice call blocking probability

The probability of voice call blocking is conditioned on the probability that more voice calls arrive than there are channels to support them. Blocking can only have occurred if the current call state $v_{n+1} = j$ is less than or equal to the previous call state $v_n = i$. If blocking occurs then the number of admitted calls that terminate is $i - j$, and the number of channels available for new calls is defined as $C = K - j$. The expected voice call blocking probability P_{VB} is then given by

$$P_{VB} = \sum_{i=0}^{K}\sum_{j=0}^{i} P\{v = i\}.b(i - j, i, \mu_{vc}).\left(\frac{\sum_{g=C+1}^{N_v - i} g.b(g, N_v - i, \rho_{vc})}{\sum_{g=1}^{N_v - i} g.b(g, N_v - i, \rho_{vc})} \right) \qquad \textbf{(6-15)}$$

where dummy variable g represents the number of arriving calls from the idle population $N_v - i$. A simpler way to obtain the average voice call blocking probability is to compute the ratio between the number of successfully admitted calls per slot and the number of call attempts per slot. The number of call attempts per slot is simply the mean offered voice load G_v. We define G_a as the average number of admitted calls per slot. This value is given simply by the expected carried call load, $\bar{v}$, divided by the mean voice call holding time (defined as $\bar{L}_{vc}$):

$$G_a = \frac{\bar{v}}{\bar{L}_{vc}} \qquad \textbf{(6-16)}$$

Another expression for the expected voice call blocking probability is then

$$P_{VB}^{*} = 1 - \frac{G_a}{G_v} \qquad \textbf{(6-17)}$$

where computational results confirm that $P_{VB}^{*} = P_{VB}$. Since a call can only terminate with probability μ_{vc} while in the silent state, and the fraction of time each terminal spends in the silent state is $1 - \mathrm{VAF}$, where VAF is given by equation 6-2, the expected length of a voice call is given by

$$\bar{L}_{vc} = \frac{1}{\mu_{vc}(1 - \mathrm{VAF})} \qquad \textbf{(6-18)}$$

Since the call termination probability μ_{vc} is orders of magnitude smaller than the talkspurt transition probabilities μ_{vt} and ρ_{vt}, the distribution of the length of voice calls can practically be considered to be geometric. The value of μ_{vc} is typically chosen to provide a mean call length of approximately 3 minutes, whereas μ_{vt} and ρ_{vt} are chosen to provide mean talkspurt and silence periods of approximately 200ms and 600ms respectively [Lee & Un, 1986].

6.4 Integrated Data/Voice Protocol Performance Analysis

The effect that voice traffic has on the data admission protocol is twofold. Firstly, the data collision threshold or number of channels remaining for data, β', is now time varying due to the fact that it is determined by the level of voice transmissions. Secondly, the transmitting voice users will contribute to the level of MAI and will thus degrade the probability of packet success. The Markov model derived in Chapter 4 is modified to account for these two factors.

6.4.1 Computation of the statistics for the modified data threshold β'

The statistics for β' can be found from the talkspurt state occupancy probabilities $P\{t = j\}$ and the four-dimensional joint state transition probability matrix $\mathbf{\Pi}$, the elements of which are computed in equation 6-8. From $\mathbf{\Pi}$, we can extract the two-dimensional talkspurt state transition probability matrix, defined as $\mathbf{\Pi}'$, where element $\Pi'_{ij} = P\{t_{n+1} = j | t_n = i\}$ is the probability of having i talking voice users in slot n and j talking users in slot n+1. The value of Π'_{ij} can be shown to be

$$\Pi'_{ij} = \frac{\sum_{k=0}^{K}\sum_{l=0}^{K}\Pi_{ik,jl}.P\{s = k\}}{P\{s \leq K - i\}} \tag{6-19}$$

Since the data threshold is determined by the number of talking voice users, we can change the subject of equation 6-19 to β' in order to reflect the data capacity transition statistics. Since the throughput performance of the data network is obviously zero for values of $\beta' \leq 0$, we are only interested in the values of t which satisfy $\beta' = (\beta - t) > 0$. If we define β'_n as the data threshold in the current slot n, and Λ_{kl} as the probability of the threshold changing from $\beta'_n = k$ to $\beta'_{n+1} = l$, then

$$\Lambda_{kl} = P\{\beta'_{n+1} = l | \beta'_n = k\} = \Pi'_{\beta-k,\beta-l} \tag{6-20}$$

The steady state probability of having a data threshold of $\beta' = l$ is given by

$$P\{\beta' = l\} = P\{t = \beta - l\} \qquad \textbf{(6-21)}$$

6.4.2 Computation of the modified data state transition probability matrices

The stochastic process which governs the data state x must now incorporate the fact that the data threshold can vary from slot to slot. For this purpose, we choose to characterise the process for x by the first-order joint-state Markov process $(x = j, \beta' = l)$. $P\{x = j, \beta' = l\}$ represents the steady state probability of being in state $x = j$ in the current slot while the threshold in the current slot is equal to $\beta' = l$.

The transition matrix for this joint state system, defined as $\boldsymbol{\pi}$, is expanded into four dimensions, and contains elements $P\{x_{n+1} = j, \beta'_{n+1} = l | x_n = i, \beta'_n = k\}$. To simplify the notation, we will define $\pi_{ik,jl} = P\{x_{n+1} = j, \beta'_{n+1} = l | x_n = i, \beta'_n = k\}$. Since the state transitions of β' are completely independent of x, we can rewrite $\pi_{ik,jl}$ as:

$$\begin{aligned}\pi_{ik,jl} &= P\{x_{n+1} = j, \beta'_{n+1} = l | x_n = i, \beta'_n = k\} \\ &= P\{\beta'_{n+1} = l | \beta'_n = k\}.P\{x_{n+1} = j | x_n = i, \beta'_n = k, \beta'_{n+1} = l\} \\ &= \Lambda_{kl}.P\{x_{n+1} = j | x_n = i, \beta'_n = k, \beta'_{n+1} = l\}\end{aligned} \qquad \textbf{(6-22)}$$

The values for $\pi_{ik,jl}$ can then be derived as follows

$$\pi_{ik,jl} = \Lambda_{kl} \times \begin{cases} 0 & i > \phi k,\ j > i \\ b(i-j, i, \mu_d) & i > \phi k,\ j \le i \\ b(i-j, i, \mu_d) & j \le i \le \phi k,\ j > l \\ \sum_{k=i-j}^{i} f_\infty(j-i+k).b(k,i,\mu_d) + b(i-j,i,\mu_d).\sum_{m=l-j+1}^{\infty} f_\infty(m) & j \le i \le \phi k,\ j \le l \\ \sum_{k=0}^{i} f_\infty(j-i+k).b(k,i,\mu_d) & i \le \phi k,\ j > i,\ j \le l \\ 0 & i \le \phi k,\ j > i,\ j > l \end{cases}$$

(6-23)

Following the logic of equation 4-16, we also require the probability that a reference packet sees itself being successfully received given that it sees the system joint states in the previous and current slots to be $(\hat{x} = i, \beta' = k)$ and $(\hat{x} = j, \beta' = l)$ respectively. Following the logic of equation 4-20, we also compute the probability that the base station counts a packet as being successful given that it sees the joint system states in the previous and current slots to be $(x = i, \beta' = k)$ and $(x = j, \beta' = l)$ respectively. These two probabilities are represented by $\hat{\pi}^{PS}_{ik,jl}$ and $\pi^{PS}_{ik,jl}$, respectively:

$$\hat{\pi}^{PS}_{ik,jl} = P\{\hat{x}_{n+1} = j, \beta'_{n+1} = l \big| \hat{x}_n = i, \beta'_n = k\}.P\{\text{packet success in slot } n+1)$$

$$= \Lambda_{kl} \times \begin{cases} 0 & i > \phi k - 1, j > i \\ b(i-j, i, \mu_d).P_{PS}(j+1+\beta-l) & i > \phi k - 1, j \le i \\ b(i-j, i, \mu_d).\sum\limits_{m=l-j}^{\infty} f_\infty(m).P_{PS}(j+m+1+\beta-l) & j \le i \le \phi k - 1, j > l-1 \\ \sum\limits_{k=i-j}^{i} f_\infty(j-i+k)b(k,i,\mu_d).P_{PS}(j+1+\beta-l) & \\ \qquad + b(i-j,i,\mu_d).\sum\limits_{m=l-j}^{\infty} f_\infty(m).P_{PS}(j+m+1+\beta-l) & j \le i \le \phi k - 1, j \le l-1 \\ \sum\limits_{k=0}^{i} f_\infty(j-i+k)b(k,i,\mu_d).P_{PS}(j+1+\beta-l) & i \le \phi k - 1, j > i, j \le l-1 \\ 0 & i \le \phi k - 1, j > i, j > l-1 \end{cases}$$

(6-24)

$$\pi^{PS}_{ik,jl} = P\{x_{n+1} = j, \beta'_{n+1} = l \big| x_n = i, \beta'_n = k\}.P\{\text{packet success in slot } n+1\}$$

$$= \Lambda_{kl} \times \begin{cases} 0 & i > \phi k, \ j > i \\ b(i-j, i, \mu_d).P_{PS}(j+\beta-l) & i > \phi k, \ j \le i \\ b(i-j, i, \mu_d).\sum\limits_{m=l-j+1}^{\infty} f_\infty(m).P_{PS}(j+m+\beta-l) & j \le i \le \phi k, \ j > l \\ \sum\limits_{k=i-j}^{i} f_\infty(j-i+k).b(k,i,\mu_d).P_{PS}(j+\beta-l) & \\ \qquad + b(i-j,i,\mu_d).\sum\limits_{m=l-j+1}^{\infty} f_\infty(m).P_{PS}(j+m+\beta-l) & j \le i \le \phi k, \ j \le l \\ \sum\limits_{k=0}^{i} f_\infty(j-i+k).b(k,i,\mu_d).P_{PS}(j+\beta-l) & i \le \phi k, \ j > i, \ j \le l \\ 0 & i \le \phi k, \ j > i, \ j > l \end{cases}$$

(6-25)

For ease of notation, we represent the joint state probability as $X(j,l) = P\{x = j, \beta' = l\}$. The steady joint state occupancy probabilities for $X(j,l)$ can be obtained by simultaneously solving the set of equations

$$X(j,l) = \sum_{i=0}^{\beta}\sum_{k=0}^{\beta} X(i,k).\pi_{ik,jl} \tag{6-26}$$

$$\sum_{i=0}^{\beta}\sum_{k=0}^{\beta} X(i,k) = 1 \tag{6-27}$$

6.4.3 Computation of the data message blocking probability

For computation of the expected data message blocking probability and message success probability, we require the probability of a reference message being blocked given that it sees the joint system state to be $(\hat{x} = j, \beta' = l)$ in its arrival slot. This probability is given by

$$P_{MB}(j,l) = \sum_{k=0}^{\beta}\begin{cases} \displaystyle\sum_{i=\phi k+1}^{\beta} \frac{X(i,k)}{X(j,l)}.\pi_{ik,jl} & \phi k < j < l-1 \\ \displaystyle\sum_{i=\phi k+1}^{\beta} \frac{X(i,k)}{X(j,l)}.\pi_{ik,jl} + \sum_{i=j}^{\phi k}\ \sum_{m=l+1-j}^{\infty} \frac{X(i,k)}{X(j,l)}.b(i-j,i,\mu_d).f_{\infty}(m) & j \le \phi k, j < l-1 \end{cases} \tag{6-28}$$

Obviously, a message is always blocked if it arrives to find the number of other transmitting users greater than or equal to the current threshold $(\beta' = l)$, thus $P_{MB}(j,l) = 1$ for all $j+1 \ge l$. The overall expected message blocking probability is then given by

$$P_{MB} = \sum_{j=0}^{\beta}\sum_{l=j+1}^{\beta} P_{MB}(j,l).P\{\beta' = l\} \tag{6-29}$$

6.4.4 Computation of the data throughput

Using the logic of section 4.4.4, we define $R_n(j,l)$ as the probability of success of the first n packets of a message, given that the n^{th} packet sees the joint system state to be $(\hat{x} = j, \beta' = l)$. The following recursive procedure is used to compute $R_n(j,l)$

$$R_1(j,l) = \begin{cases} X(j,l).[1 - P_{MB}(j,l)].P_{PS}(j+1+\beta-l) & j < l, l \leq \beta \\ 0 & j \geq l, l \leq \beta \end{cases} \qquad \textbf{(6-30)}$$

$$R_n(j,l) = \sum_{i=0}^{\beta-1}\sum_{k=0}^{\beta-1} R_{n-1}(i,k).\hat{\pi}_{ik,jl}^{PS} \qquad j < \beta, l \leq \beta \qquad \textbf{(6-31)}$$

For n=1, the probability of success of the first packet of a message is obviously equal to zero if the packet arrives to find more than β' users (including itself) transmitting since this would be a collision scenario. Note that in equations 6-24, 6-25 and 6-30 that we have included the number of transmitting voice users contributing to the MAI when computing the probability of packet success. If the number of transmitting data interferers is j and the data threshold is equal to l, then the number of transmitting voice users is obviously equal to $\beta - l$ and the total number of transmissions is $j + 1 + \beta - l$ (including the reference user itself). The probability of success of a message containing L packets is then

$$P_{MS}(L) = \sum_{j=0}^{\beta-1}\sum_{l=0}^{\beta} R_L(j,l) \qquad \textbf{(6-32)}$$

and the total average message success probability is

$$P_{MS} = \sum_{l=1}^{\infty} P_{MS}(l) \cdot P_{ML}(l) \qquad \textbf{(6-33)}$$

The expected data message throughput is

$$S_M = G.P_{MS} \qquad \textbf{(6-34)}$$

Using π^{PS} from equation 6-25, the average data packet throughput can be shown to be

$$S_P = \sum_{j=1}^{\beta}\sum_{i=0}^{\beta}\sum_{l=1}^{\beta}\sum_{k=0}^{\beta} j.X(i,k).\pi_{ik,jl}^{PS} \qquad \textbf{(6-35)}$$

6.4.5 Data performance without voice activity exploitation

If there is no exploitation of the voice activity factor then the voice model illustrated in figure 6.2 is simplified to that which appears in figure 6.4, where there are only two states: idle and active.

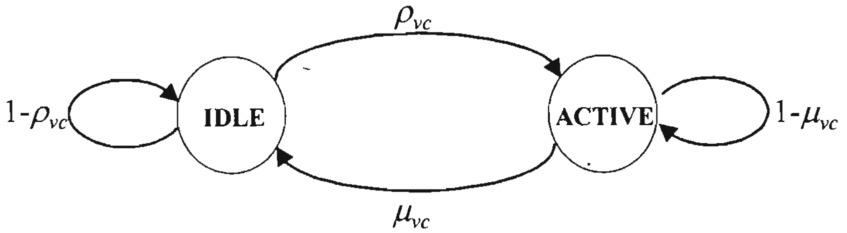


Figure 6.4 : Voice model for the case where there is no voice activity factor exploitation (or equivalently, the VAF=100%)

In this scenario, the admission of data users is dependent on the number of voice calls in progress rather than the number of talking voice users in progress. Each active voice terminal's transmitter is on continuously for the entire duration of the voice call. Alternatively, one can imagine this to be a scenario in which the voice activity factor of each terminal is always 100 percent. If v is the number of calls in progress, then the state transition probabilities for v, represented as $P\{v_{n+1}=j|v_n=i\}$, are given as

$$P\{v_{n+1}=j|v_n=i\} = \begin{cases} \sum_{n=0}^{i} b(n,i,\mu_{vc}).b(j-i+n,N_v-i,\rho_{vc}) & j>i \\ \sum_{n=i-j}^{i} b(n,i,\mu_{vc}).b(j-i+n,N_v-i,\rho_{vc}) & \\ \quad + \sum_{g=K+1-j}^{N_v-i} b(g,N_v-i,\rho_{vc}).b(i-j,i,\mu_{vc}) & j\le i \end{cases} \qquad \textbf{(6-36)}$$

The steady state occupancy probabilities for the number of voice calls in progress can be found by solving the set of linear equations given by

$$P\{v=j\}=\sum_{i=0}^{K}P\{v=i\}.P\{v_{n+1}=j|v_n=i\} \qquad \textbf{(6-37)}$$

$$\sum_{j=0}^{K}P\{v=j\}=1 \qquad \textbf{(6-38)}$$

Since the data threshold, β', is now determined only by v, the steady state probability of having $\beta'=l$ is now given by the probability of having l voice calls in progress

$$P\{\beta'=l\}=P\{v=\beta-l\} \qquad \textbf{(6-39)}$$

The probability that β' changes from k to l over successive slots, defined as Λ_{kl} in section 6.4.1, is now given by

$$\begin{aligned}\Lambda_{kl}&=P\{\beta'_{n+1}=l|\beta'_n=k\}\\&=P\{v_{n+1}=\beta-l|v_n=\beta-k\}\end{aligned} \qquad \textbf{(6-40)}$$

Equations 6-39 and 6-40 can then be used in the data analysis for the case in which there is no voice activity factor exploitation, instead of equations 6-21 and 6-20.

6.5 Results for the Integrated Traffic Model

6.5.1 Network parameters

The pertinent parameters for this network are provided in table 6.1. Other parameters associated with the data protocol are as outlined in table 4.2. The data protocol is the SR-DT/CDMA protocol with an infinite population. For the voice network, the voice call holding times and voice talkspurt lengths are assumed to be fixed, whereas the voice terminal idle times and speech silent periods are assumed to vary to provide different offered voice loads and voice activity factors, respectively.

For the mean voice call holding time, we assume the standard telephone conversation length of approximately 3 minutes which, for a slot length of 16ms, equals 11250 slots. For the mean talkspurt length, we assume a value of approximately 220ms (14 slots), according to measurements presented in [Lee & Un, 1986] and [Gruber, 1982]. The voice threshold is chosen to be K=20 (this is equal to the data collision threshold β=20 as found in section 4.5.2). K=20 reflects a maximum tolerable bit error rate of approximately 10^{-3} - a value which has commonly been assumed to be the worst case allowable bit error rate for digital voice (e.g. [Gilhousen *et al*, 1991] and [Yang & Geraniotis, 1994]).

Item (*units*)	**Symbol**	**Value**
Data population (***terminals***)	N_d	∞ (SR-DT/CDMA)
Voice population (***terminals***)	N_v	100
Maximum voice threshold (***transmissions***)	K	20
Maximum data threshold (***transmissions***)	β	20
Average data message length (***slots***)	$\overline{L}_d$	4
∴Message termination probability	μ_d	0.25
Offered data load	G_d	Variable
Average talkspurt length (***slots***)	N/A	14 (220 ms)
∴Talkspurt initiation probability	ρ_{vt}	0.07142
Average silent period length (***slots***)	N/A	variable
∴Talkspurt termination probability	μ_{vt}	variable
Average voice call holding time (***slots***)	$\overline{L}_{vc}$	11250 (3 minutes)
∴Call termination probability	μ_{vc}	$\mu_{vc} = [11250\times(1-VAF)]^{-1}$
Average voice terminal idle time (***slots***)	N/A	variable
∴Call initiation probability	ρ_{vc}	variable

Table 6.1 : Imaginary network parameters used in the analytical evaluation and simulation

6.5.2 Results for the voice network

The performance measures of interest for the voice network are the average carried voice loads and the expected voice call blocking probability for a given offered voice load. In figure 6.5, we plot the expected carried number of voice calls, $\bar{v}$, as computed in equation 6-12, and the expected number of talking speakers, $\bar{t}$, as computed in equation 6-14, for a given offered voice call arrival rate G_v. As one would expect, the level of voice traffic will be low whenever the probability of generating voice calls is low. As G_v increases (by increasing ρ_{vc}), $\bar{v}$ increases. Since the threshold for voice users is $K = 20$, the carried call load asymptotically approaches K as the voice call arrival rate becomes large. The curves for the number of talking speakers, $\bar{t}$, confirm our analysis of the voice activity factor, where it is clearly seen that the number of talking speakers is indeed equal tõ the number of calls in progress multiplied by the voice activity factor. The analysis of the admission policy is also correct, where it can be seen that the number of calls can never exceed K due to the blocking scheme.

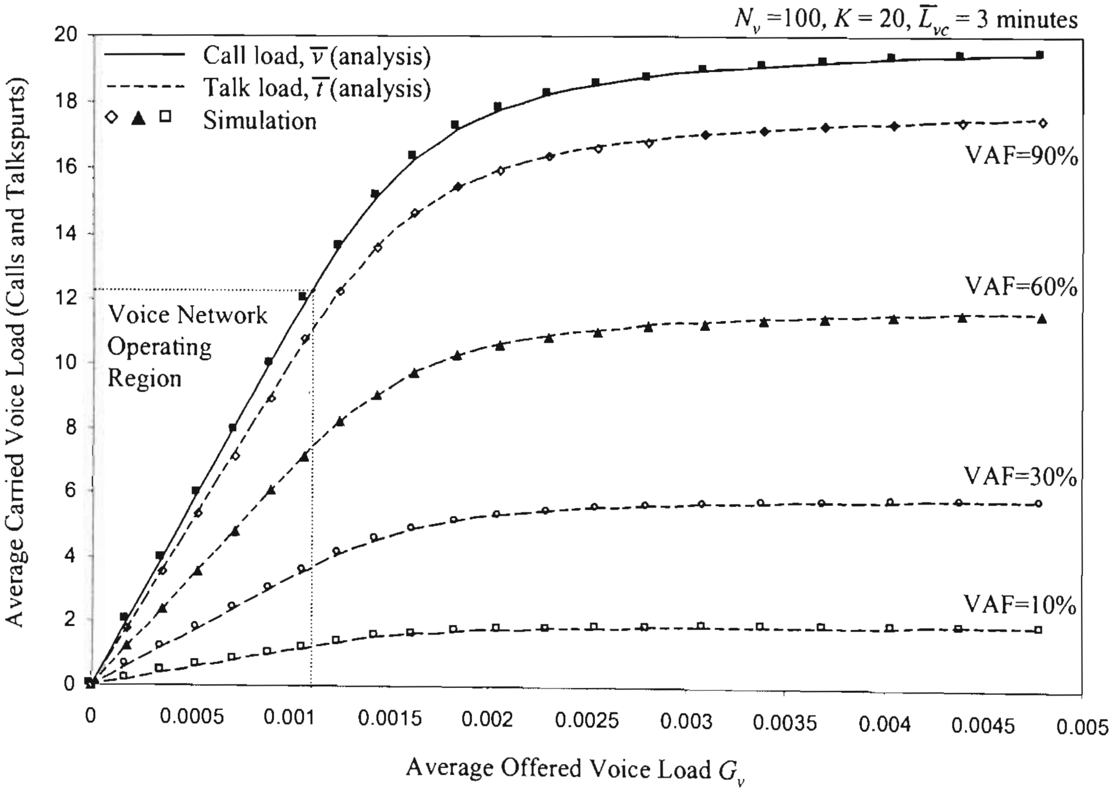


Figure 6.5 : Average carried voice and talkspurt loads ($\bar{v}$ and $\bar{t}$) versus the offered voice call load (G_v), for various voice activity factors (10%, 30%, 60%, and 90%)

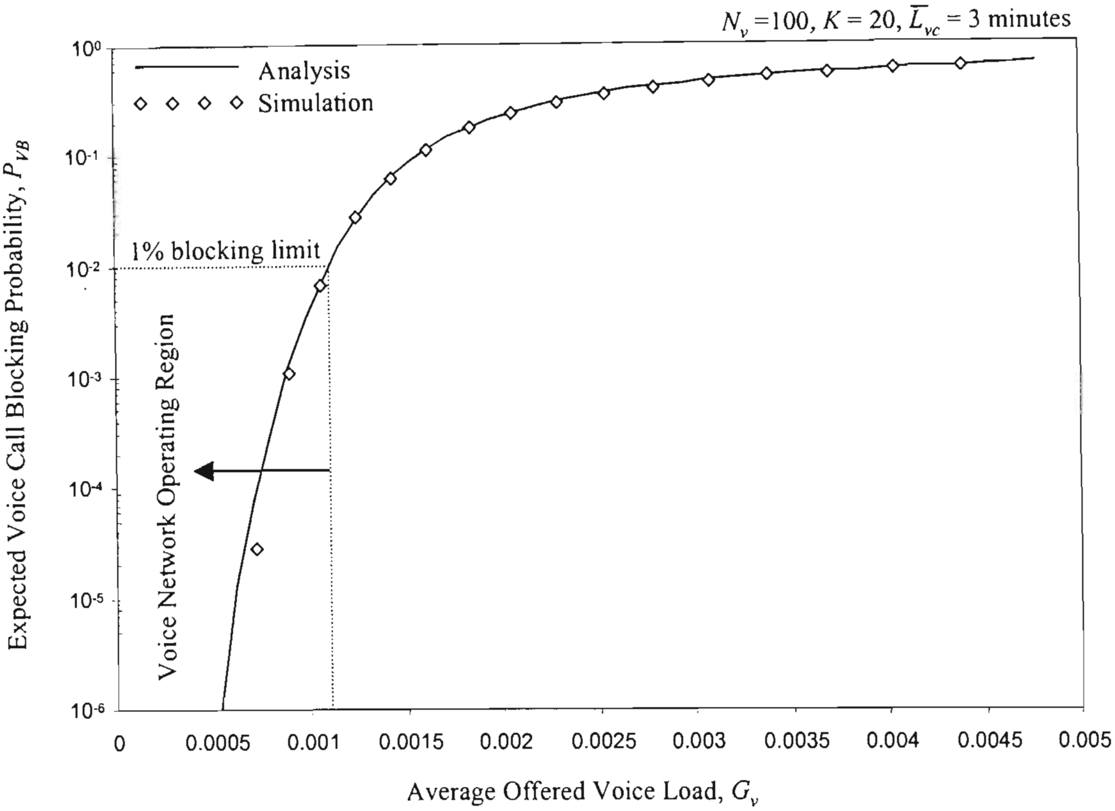


Figure 6.6 : Expected voice call blocking probability versus the offered voice call load

The expected call blocking probability P_{VB} (equation 6-15) versus the offered voice load is plotted in figure 6.6. In cellular systems supporting voice traffic, adequate service is usually associated with a blocking probability of less than 2 percent, and network engineers usually dimension the network in such a manner to ensure that this probability is rarely exceeded. In our network, if we assume that an expected blocking probability of 1 percent is rarely exceeded, then we can see that the operating region of the voice network is one in which the offered voice load does not exceed approximately $G_v \approx 0.001$ calls per slot. In terms of the exact network timing, this implies a peak call arrival rate of one call every 1000 slots, or every 16 seconds, for the given population size $N_d = 100$ and voice threshold K=20.

If we look back at figure 6.5, we see that for a maximum offered voice call load of $G_v \approx 0.001$, the expected number of calls in progress is approximately $\bar{v} \approx 12$. In terms of the data network, this implies that, for a maximum data threshold of $\beta = 20$, we can expect there to be approximately $\overline{\beta'} \approx 8$ channels available for data, on average, during the worst case voice load with no voice activity exploitation. If the standard VAF of approximately 30 percent is assumed [Lee & Un, 1986], then we would expect approximately 4 people out of the 12 to be talking on average, and hence an expected data threshold of $\overline{\beta}' = 16$, which is double that of the case when the VAF is not exploited.

6.5.3. Results for the integrated voice/data network

For the integrated voice data network, we examine the data performance under steady state voice conditions. This means that the mean offered voice load and voice activity factor are assumed to be time invariant. We consider two voice network scenarios, namely a light voice load case in which there are $\overline{v} = 6$ voice calls in progress on average, and a heavy voice load in which there are $\overline{v} = 15$ calls in progress on average. As mentioned in the previous section, the heavy load is beyond the expected operating region for the voice network, but is useful for examining the data performance under extreme voice cases.

In figures 6.7a and 6.7b, we plot the data packet throughput and data message throughput (respectively) for the light voice load case for various voice activity factors (VAF=0%, VAF=30%, VAF=60% and VAF=100%). The case VAF=0% is equivalent to the no-voice data case and hence full data capacity conditions, as obtained in Chapter 4. The case VAF=100% refers to the model and analysis outlined in section 6.4.5 (i.e. no voice activity exploitation). In figures 6.8a and 6.8b, we plot similar results for the heavy voice load case.

It is obvious that the data throughput must decrease as the number of voice calls increases (or the voice activity increases) due to the fact that the data threshold decreases. For light voice loads with a small voice activity factor, we obviously expect a minimal reduction in the performance of the data network. For a heavy voice load with a high voice activity factor (e.g. $\overline{v} = 15$ with VAF=100%), the data throughput can be seen in figure 6.8 to be severely reduced. Figures 6.7 and 6.8 reveal the relationship between the mean number of voice calls, mean capacity available to data and the data performance. For example, a mean voice load of $\bar{v} = 6$ calls with a VAF of 100 percent

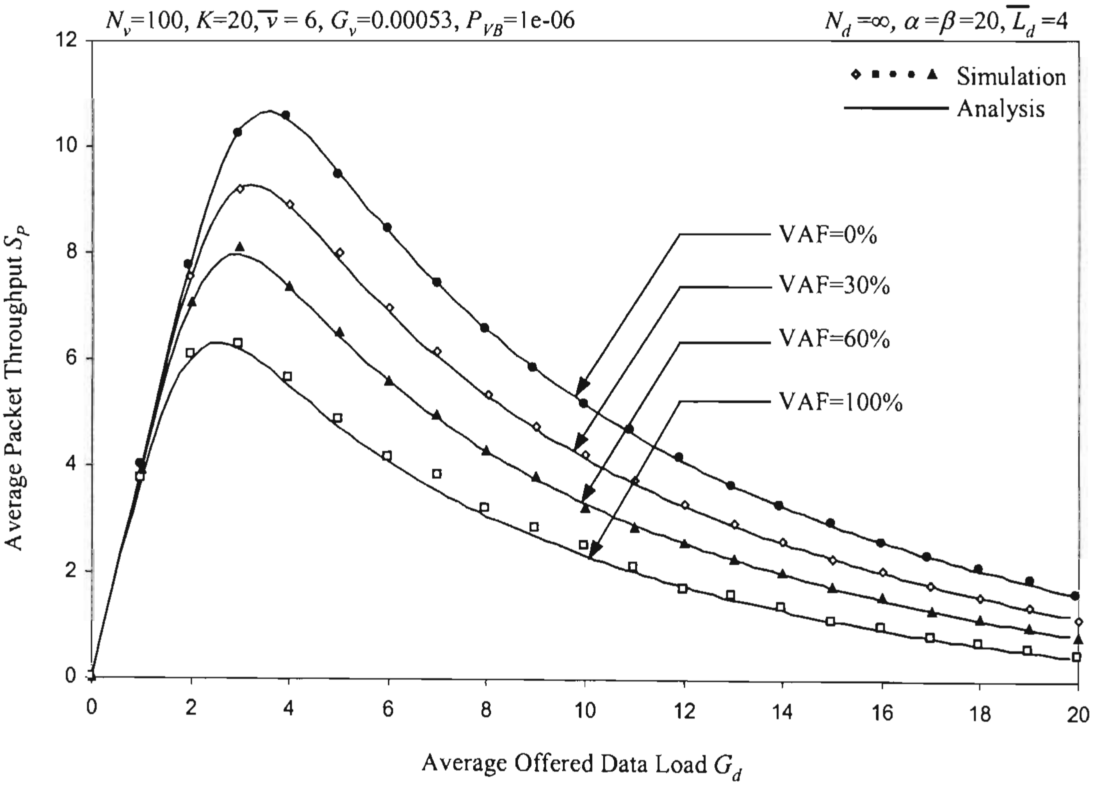


Figure 6.7a : Data packet throughput for a light voice load ($\bar{v}$ =6 calls) and various VAF

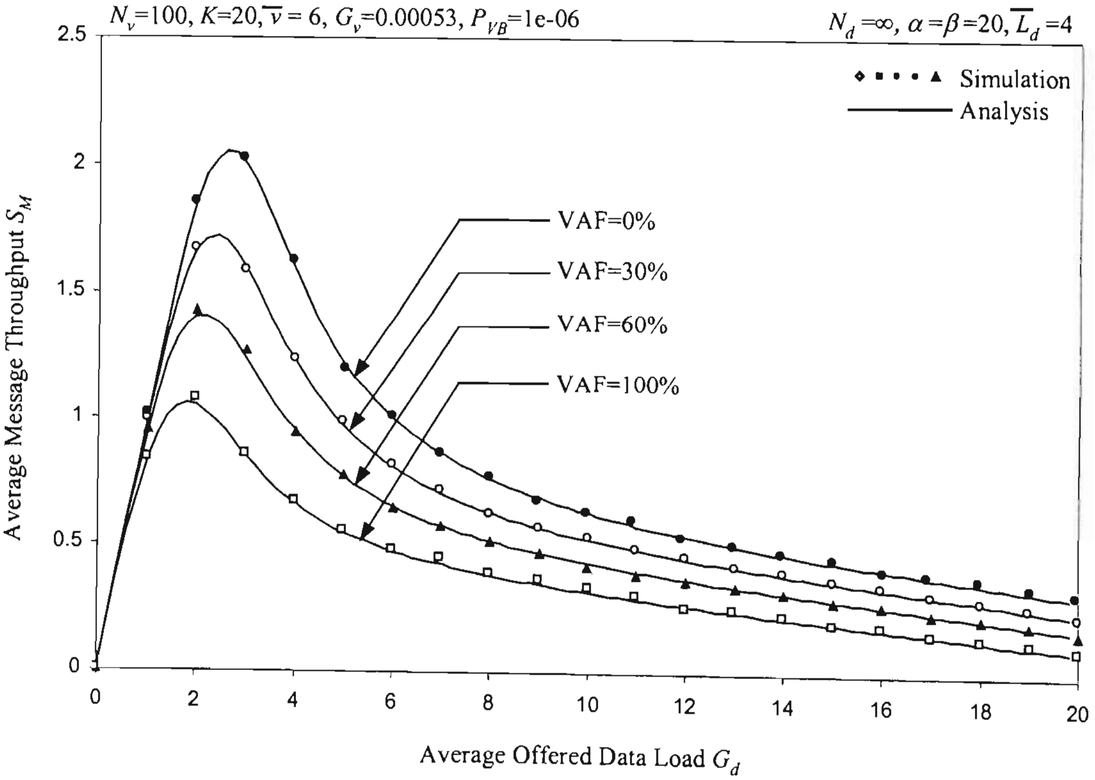


Figure 6.7b : Data message throughput for a light voice load ($\bar{v}$ =6 calls) and various VAF

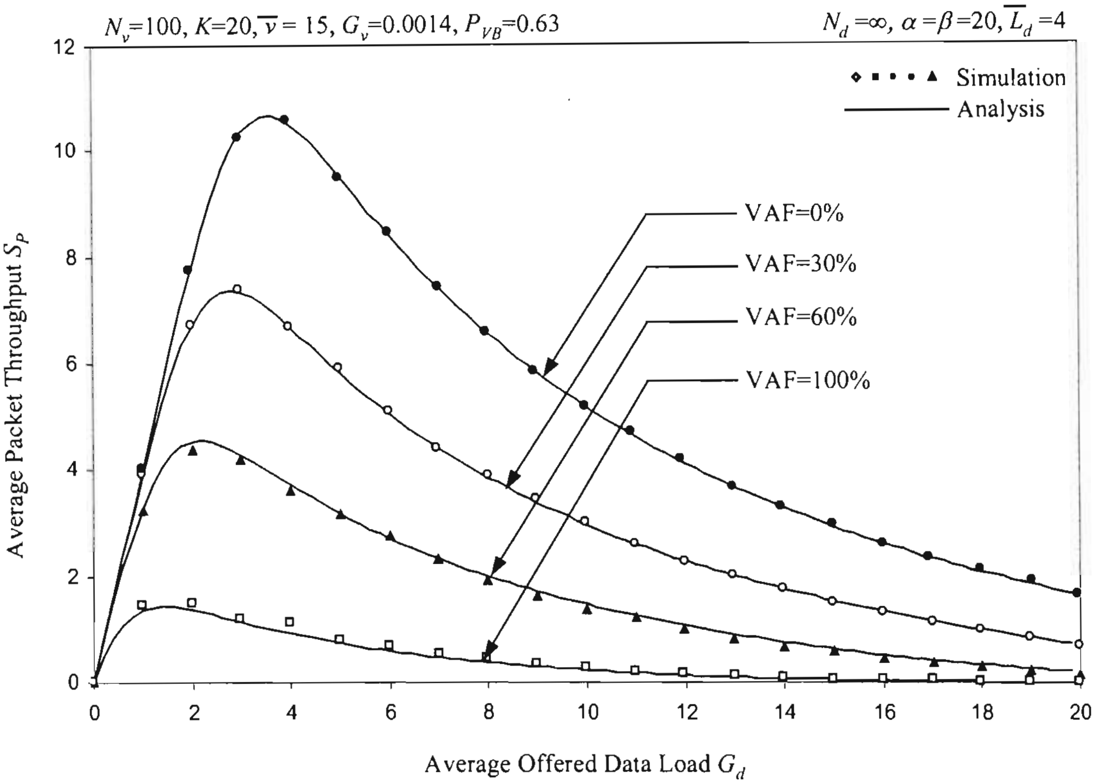


Figure 6.8a : Data packet throughput for a heavy voice load ($\overline{v}$ =15 calls) and various VAF

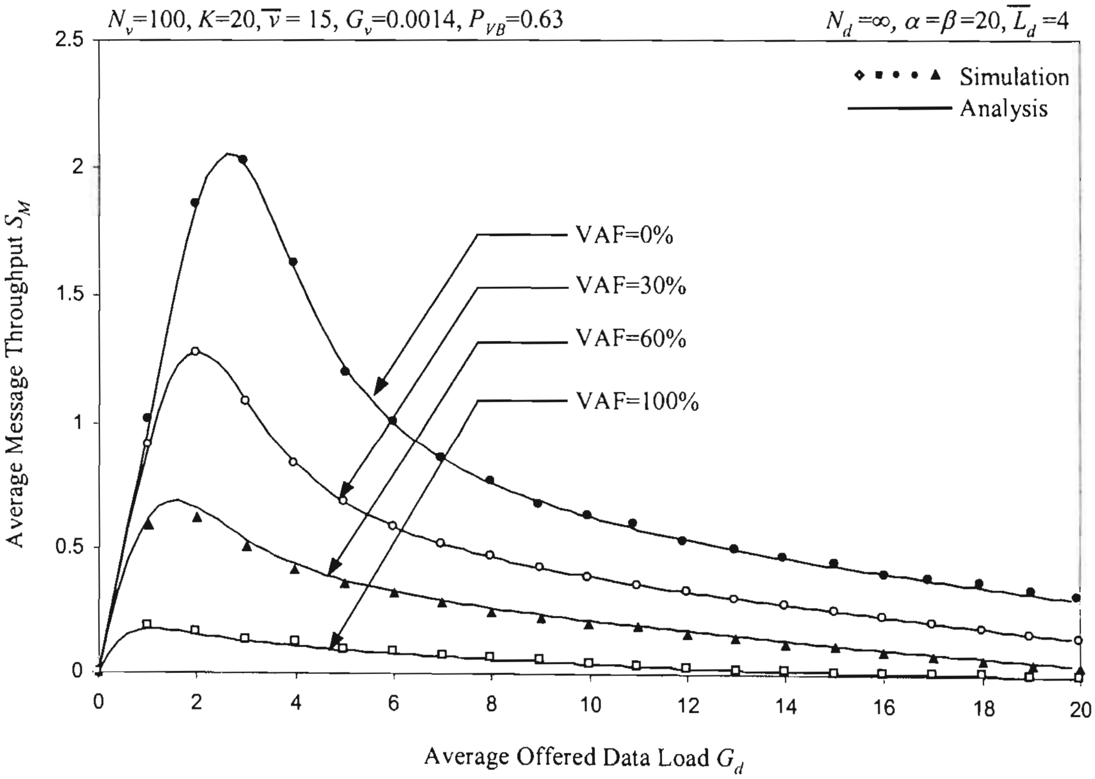


Figure 6.8b : Data message throughput for a heavy voice load ($\overline{v}$ =15 calls) and various VAF

implies a reduction in data capacity of 30 percent (since $\overline{\beta}'/\beta = 14/20 = 0.7$). Similarly, a mean voice load of 15 voice calls with a VAF of say 60 percent implies a reduction in data capacity of 55 percent. The reduction in the data throughput is seen to be always larger than the reduction in the data capacity, and the reason for this is that, in addition to the percentage throughput reduction due to the decrease in data capacity, there is an added reduction in data throughput due to the additional MAI contributed by the voice users.

Exploitation of the voice activity factor can be seen to provide a significant improvement over the case where the data capacity is determined by the number of calls in progress (VAF=100%). The exact percentage increase in capacity achieved is dependent on the maximum data capacity, β, and the number of calls in progress, v, and is given by

$$\frac{\overline{\beta}'_{VAF} - \overline{\beta}'_{NO\_VAF}}{\overline{\beta}'_{NO\_VAF}} = \frac{\overline{v} - VAF.\overline{v}}{\beta - \overline{v}} \qquad \textbf{(6-41)}$$

where $\overline{\beta}'_{VAF} = \beta - \overline{t}$ and $\overline{\beta}'_{NO\_VAF} = \beta - \overline{v}$ are the expected data capacities with and without voice activity exploitation. If we consider the standard VAF of 30 percent for normal human speech, and a mean voice load of 6 calls, we see that a capacity improvement of 30 percent occurs over the case where the VAF=100%. This capacity improvement increases for increasing $\overline{v}$. For $\overline{v} = 15$ and VAF=30%, the percentage capacity increase is 210 percent. The percentage improvement in the data throughput will always be larger than the percentage increase in the data capacity, due to the reduction in MAI that accrues as a result of suppressing each voice terminal's signal during the silent periods.

The high degree of correlation between the analytical and simulation results in figures 6.5 to 6.8 validates our analytical approach. In figure 6.9, we validate the analytical model, as well as show that the usual data performance enhancements occur at higher data loads, for the case where a blocking threshold $\alpha' < \beta'$ is employed. Since α' is a time varying threshold, we chose to plot results for a fixed threshold ratio, ϕ. The results

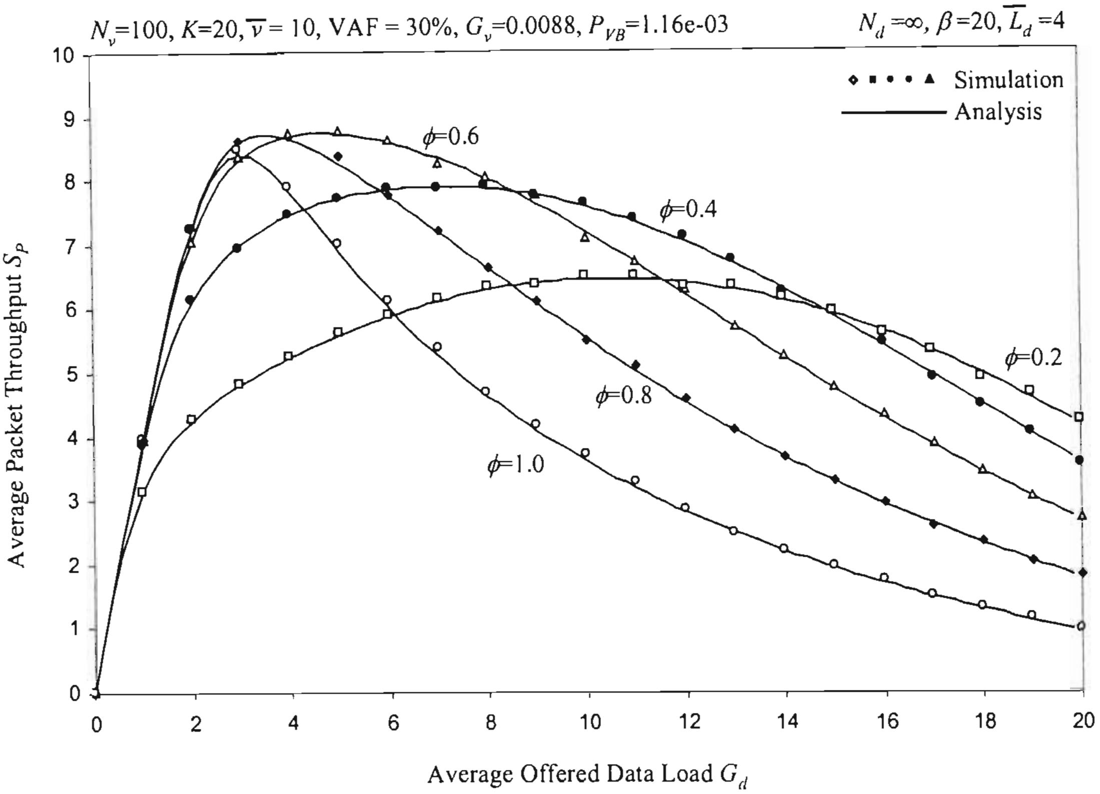


Figure 6.9 : Effect of α on the data throughput performance for the case where there is voice traffic

in figure 6.9 are for a mean voice load of $\overline{v} = 10$, a VAF of 30 percent and hence an expected data capacity of $\overline{\beta'} = 17$. So, for an expected capacity of $\overline{\beta'} = 17$ and a threshold ratio of $\phi = 0.6$, this translates into an expected blocking threshold of $\overline{\alpha'} = \phi\overline{\beta'} = 10.2$. As can be seen from figure 6.9, the blocking threshold α has the same performance enhancing effect on the data performance for the voice case as it did for the non-voice case. This result is expected, since the blocking threshold works only in conjunction with the collision threshold, where it does not matter what the collision threshold is, only that the frequency with which this threshold is exceeded is reduced. Again, we can see that for any given collision threshold, there is a corresponding value of α (or ϕ) which optimises the data throughput for a given offered data load.

6.5.4 Discussion on the choice of thresholds and exploitation of the VAF

For the network parameters considered in this chapter, we chose the maximum data threshold, β , and the voice threshold, K, to have the same value. This would not be the case if say, we required a lower bit error rate for voice, or if the data packet length was different. In sections 3.5.2 and 3.5.5, we discussed the fact that the multiple access threshold for data is a function of the message (and packet) dimensions. If K is smaller than β , then we have the case where there will always be capacity available for data, but the required bit error conditions for voice traffic are not always acceptable if the total number of transmitting users (voice and data) exceeds K. If K is larger than β , then there will be cases when there is no capacity available for data. In the latter case one may choose to decrease K to equal β to ensure that there is always capacity available to data. In decreasing K, we effectively set the maximum bit error rate to be lower, and hence ensure a better worst case quality of speech at the expense of a poorer call blocking rate. Either way, the combined fact that voice has unconditional priority over data together with the threshold assignment must ensure that the maximum number of users possible is always less than that which concedes the worst case BER for the traffic type which requires the lower BER.

Another point of discussion is exactly how one wishes to exploit the voice activity factor. In this thesis, we consider using the extra capacity to improve the data performance. Another option might be to increase the voice capacity K, to allow more simultaneous voice calls. For the accepted VAF of 30 percent, we could effectively increase K by a factor of 3.33. This, of course, would serve only to further reduce the performance of the data network. We know that the maximum number of simultaneous users in the network should never be larger than β , to ensure that the data throughput is optimised.

A suitable option might be to use some of the extra capacity due to the VAF exploitation in order to set $K=\beta$ (if K was originally less than β), and use the remaining exploited capacity to improve the data throughput. In this way, we have a single channel threshold that is the same for both traffic types, and represents the maximum number of transmissions allowed (both voice and data) such that the operating conditions for both traffic types are just adequate.

6.6 Summary

This chapter in its entirety dealt with our proposed model for integrating voice and data in a CDMA packet radio network. We began by presenting the imagined network architecture, the proposed multiple access policies for both service types, and the traffic models for both services. For the voice network, we considered a finite population of identical voice generating terminals, and modelled each terminal as a three-state, discrete-time Markov process. This three state model describes the call holding times and terminal idle times between calls as being exponentially distributed (or geometrically distributed in terms of packets), and also models the speech activity of each terminal during a voice call, where the talkspurt and silent period lengths are also geometrically distributed. For the data network, we considered an infinite population of identical data terminals, a Poisson data message arrival process, and employed our proposed SR-DT/CDMA data MAC protocol as presented in Chapter 4.

The second part of this chapter presented the derivation of an exact Markov analysis of the proposed network. For the voice network, we computed the expected distribution for the states of the voice population, expected values for offered and carried voice loads, and the expected voice call blocking probability. In the case of the data network, we computed the expected data state distribution, the mean data message blocking probability and the expected packet and message throughputs.

The third part of this chapter presented results for the proposed model as obtained from the analytical model and the simulator for a given set of network parameters. A high degree of correlation between simulation and analytical results validated our analytical approach and solution. For the voice network, we were able to find the maximum voice call arrival rate that causes blocking in excess of the widely excepted maximum (between 1 percent and 2 percent for cellular networks), and hence dimension for the expected operating level of voice traffic. For the data network, we showed the obvious fact that the performance of the data network degrades with the addition of voice users to the channel. This degradation is a result of the fact that the channel capacity for data is reduced by the number of transmitting voice users, as well as the fact that the transmitting voice users increase the level of MAI in the channel. However, we showed that this data performance degradation can be significantly reduced by exploiting the voice activity factor of human speech, especially if each voice terminal's transmitter is severely suppressed (or even switched off entirely) during the silent periods in the speech, such that the level of MAI contributed during the silent periods is negligible. For a standard voice activity factor of 30 percent, we can see that the reduction in data performance is not as severe as one might imagine, even for a heavy offered voice load which is beyond the normal operating region of the offered voice load curve. We also showed that our dual data threshold (α and β) system has the same performance enhancing effects on the data performance when voice is integrated into the channel.

We concluded the chapter with a discussion on the topic of the two traffic thresholds K and β, and discussed various options for setting these thresholds to simultaneously satisfy the channel requirements for both services. Although not dealt with in detail in this thesis, we also discussed how the voice activity factor may be exploited for the benefits of both voice and data services, instead of just data.

The contents of this chapter were presented by the author, in part, at the PIMRC'99 conference in Osaka, Japan, in September 1999 [Judge & Takawira, 1999a], and have been submitted for review to the journal "IEEE Transactions on Vehicular Technology" [Judge & Takawira, 1999b].

CHAPTER 7

EFFECTS OF A CORRELATED RAYLEIGH FADING CHANNEL ON THE PROPOSED SYSTEM

7.1 Introduction

In most preliminary analyses of a multi-access protocol, it is generally assumed that the channel is ideal. This is done in order to isolate the performance advantages gained through the application of a specific protocol, since protocol performance comparisons can be made under identical channel conditions. Thereafter, the protocol performance can be analysed under more realistic channel conditions. In a realistic, imperfect, mobile radio environment such as the DS-CDMA reverse link channel assumed in this thesis, the propagation and final reception of each user's transmitted signal is influenced and distorted by a variety of factors, the most prevalent and influential being: distance path loss, shadowing, multipath fading and, for a spread-spectrum system, the MAI. These influences are, by themselves and combined, all very interesting and have been the focus of much research. Since they are not specifically the focus of this chapter, we rather refer the reader to [Jakes, 1974] and [Kchao & Stüber, 1993] for details on, and accounts of, the effects of these influences.

In this chapter we analyse the effects of a Rayleigh fading channel on the performance of our multiple access DS-CDMA scheme. To aid in this, we propose a hidden Markov model (HMM) to model the stochastic channel process and, in particular, the error sequences that occur in the channel. Our proposed HMM is a CDMA, multi-user version of the popular narrowband, single-user Gilbert-Elliot HMM. Justification for the work in this chapter is based on the apparent lack of work in the area of Markov models for CDMA channels. A literature survey on the topic of Markov channel models revealed a large amount of work done for narrowband channels, but very little work, if any at all, done on Markov models for multi-user spread-spectrum channels.

In section 7.2 we present a very brief review of traditional Markov channel models, such as the popular Gilbert-Elliot model, that are used to simplify the modelling and analysis of imperfect narrowband, single-user channels. In section 7.3, we detail the statistical channel models used for the user and MAI amplitudes in the channel. In section 7.4 we then propose, discuss and compute approximations for the parameterisation of the HMM. We then apply the HMM to model the probability of packet and message success in our SR-DT/CDMA MAC protocol in order to quantify the effects that correlated Rayleigh fading has on the throughput performance of our network. In section 7.5, we evaluate the results and accuracy of the proposed HMM.

7.2 Traditional Markov Channel Models

Markov models (MM's) have been proven to be a very powerful tool for modelling a wide variety of stochastic processes. In the context of channel modelling, MM's have found great application in modelling channels with "memory". The term "memory" is given to channels in which the current channel state is dependent on, and correlated to, the channel's state history. Apart from being used to model the actual channel fading process, MM's are usually used to model the error process and error probabilities in channels carrying digital information. The reason for this is that communications engineers are often more concerned with the occurrence and statistics of the channel errors, rather than the statistics of the channel itself, even though the two are closely related. In most real digital communications channels that exhibit correlated fading, it is a well known phenomenon that channel errors tend to occur in clusters or bunches, separated by long error free gaps [Gilbert, 1960], [Kanal & Sastry, 1978]. These error clusters have been shown to be associated with the periods in which the channel is in a "deep fade" state (very low signal-to-interference ratios). If the error or non-error occurrences are laid out in time, then an *error sequence* is obtained. Since, in a digital communications system, a bit (or byte, or any sized sequence of bits for that matter) is either in error or not, the error sequence is essentially a binary sequence. Determining the statistical structure of these error sequences while, at the same time, characterising this behaviour of fading channels to exhibit these biased tendencies towards either being "too good" or "too bad", has been of significant interest over the year, and therein lies much motivation for channel modelling.

Gilbert initiated the study of MM's for channels which exhibit memory in the error process [Gilbert, 1960]. His work, together with that of Elliot [Elliot, 1963], gave birth to the popular Gilbert-Elliot channel model and, subsequently, a family of MM's for modelling channels with memory.

7.2.1 The Gilbert-Elliot model

The notion of a two-state binary error sequence gave rise to the original two-state Gilbert-Elliot model that considers defining the channel as either being "(*G*)ood" or "(*B*)ad". Figure 7.1 illustrates this Gilbert-Elliot Markov model.

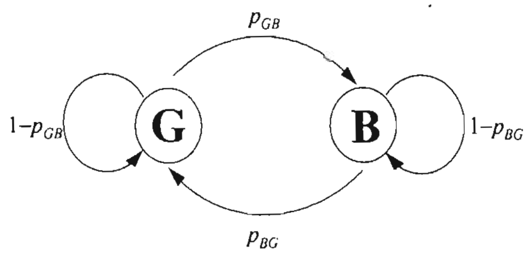


Figure 7.1 : The traditional two-state Gilbert-Elliot channel model

The exact resultant sequence of error occurrences depends on how one defines what happens during a good or a bad state. In the original model of [Gilbert, 1960], it was assumed that an error occurred during the *G* state with probability zero, and in the *B* state with some probability, *h*. In the model of [Elliot, 1963], it was assumed that errors occurred in both states: in the *G* state with small probability *k*, and in the *B* state with large probability *h*. Another common assumption is that errors occur in the *G* state with probability zero, and in the *B* state with probability one. This latter case would be an example of a non-hidden MM, whereas the first two cases are HMM's. Depending on which scheme is used, it is easy to map the states onto the error sequence.

The difficulty and main area of interest lies in parameterisation of the model, i.e. determining the values of p_{GB}, p_{BG}, *h* and *k* such that the model provides the best fit to the actual channel statistics. For the Gilbert-Elliot example, the parameters of interest for the channel statistics are the durations of the "good" and "bad" periods of the channel, which eventually manifest themselves in the state transition probabilities p_{GB} and p_{BG}. Illustrated in figure 7.2, is Gilbert's original concept, which is now very popular, for determining whether a narrowband single-user channel is good or bad.

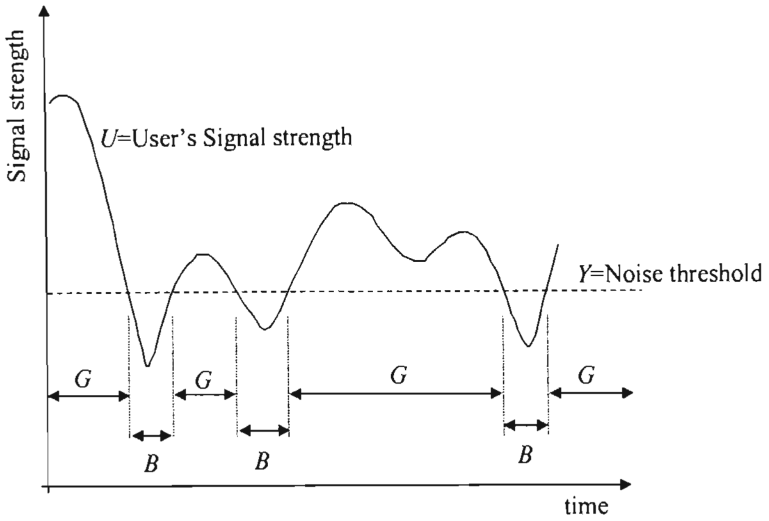


Figure 7.2 : Illustration of a narrowband channel being defined as "good" or "bad"

If the user's received signal power (*U*) is above some fixed and pre-defined noise level (*Y*), then the channel is in the good state *G*. If the received signal power is below the noise level, then the channel is in state *B*. In a narrowband single-user channel, the noise threshold level *Y* is almost always assumed to be fixed, usually to some fixed fraction of the AWGN component of the channel. Alternatively then, one can imagine that if the signal-to-noise ratio (SNR) exceeds some predefined value, then the channel is good, otherwise it is bad. Since a bad channel state occurs during periods of a deep fade (low SNR), we may choose to use the term "fade state" synonymously with "bad state", and "non-fade state" synonymously with "good state". If the statistical description of *U* is known (this usually includes the probability distribution of *U*, the rate of change of *U* over time and the associated auto-covariance and auto-correlation functions of *U*), then it is not difficult to obtain the state transition probabilities for the two-state model. For example, in [Jakes, 1974], details are provided for parameterisation of the model for a correlated Rayleigh channel. The Gilbert-Elliot model's simplicity and modest accuracy have made it extremely popular for modelling error sequences in fading channels.

7.2.2 Higher-order Hidden Markov Models

Based on the Gilbert-Elliot model, several extended state HMM's have been proposed that improve on the accuracy of the original model. The drawback of these models is that they become increasingly complicated as more states are added. The number of models is vast and even a brief overview is beyond the scope of this thesis. Instead, we discuss a generic model proposed by Fritchman [1967], which gave rise to, and can be modified into, many of these schemes. The Fritchman channel model is based on a finite number of N states, partitioned into a group A, of K error free states, and another group B, of N-K error states. This concept is illustrated in figure 7.3.

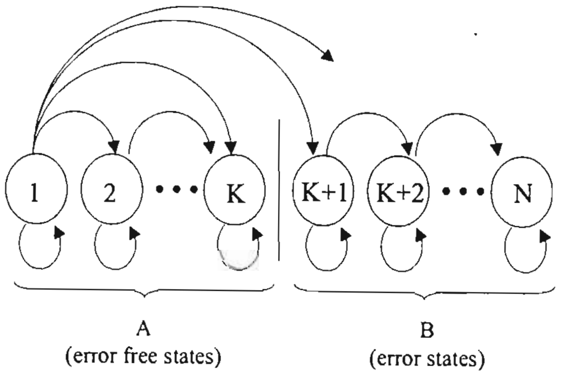


Figure 7.3 : The traditional generic Fritchman model

Kanal & Sastry [1978] give an excellent review of the Fritchman model and many of derivative higher-order HMM's based upon this scheme. In [Turin & van Nobelen, 1998], the applicability of high-order HMM's in accurately modelling fading channels is discussed. They also provide a good review of several general techniques for the parameter estimation of such models, and derive accurate expressions for the fade duration and level crossing distributions of a flat Rayleigh fading channel. The result of their work reveals that HMM's are highly successful in modelling fading channels.

7.3 Proposed Channel Signal Model

In this chapter we consider an imperfect fading reverse link channel that has the following assumed effects on the signals that are received at the base station:

1. Each terminal's transmitted signal is affected by *Rayleigh multipath fading*, meaning that the received signal from each user is *Rayleigh* distributed.
2. The Rayleigh fading is assumed to be *frequency non-selective*. This means that all frequencies are equally affected by a fade.
3. All users are assumed to share the *same fading parameters*. That is, the mean and variance of the Rayleigh distribution in assumption (1) are the same for all users.
4. Each terminal's communications link fades *independently* of every other user's link.
5. The fading process is assumed to be fairly *slow* in comparison to the system slot duration, such that the signal variations over a system time slot are negligible to the point that we assume that the signal amplitude remains constant over the duration of a slot. The received signal amplitudes in consecutive slots are also assumed to be reasonably highly correlated.
6. The fading channel is assumed to have no effect on the MAC protocol. This means that, even when a terminal is in a deep fade state, the base station is still aware of its presence, and includes it in the number of transmitting users when deciding which signal tones to broadcast.

Throughout the remainder of this chapter, we will refer to random variable U as the signal received by the base station from a single *reference user* in which we are interested, and random variable Y as the combined sum of received signals from the other users in the channel (i.e. the *Multiple Access Interference*). The aforementioned channel assumptions imply that U will be a Rayleigh distributed random value, and Y will be the sum of several Rayleigh distributed random values. What follows in the remainder of this section, are our signal models for U and Y.

7.3.1 User Signal Model (U)

In this thesis, we consider a popular and widely used model for a flat fading channel. The channel model considers the amplitude of signal U as being modelled as complex, stationary Gaussian process, $\alpha(t)$, with independent and identically distributed real and imaginary components, where the envelope of $|\alpha(t)|$ is Rayleigh distributed [Jakes, 1974], [Zorzi, Rao & Milstein, 1997]:

$$P_{sig}(u) = \frac{u}{b^2}\exp\left(\frac{-u^2}{2b^2}\right) \tag{7-1}$$

where subscript "*sig*" denotes the reference signal and Rayleigh parameter b determines the mean and variance of the distribution according to the following expressions:

$$\text{mean} = \mu_U = b\sqrt{\frac{\pi}{2}} \tag{7-2a}$$

$$\text{variance} = \sigma_U^2 = (2 - \frac{\pi}{2})b^2 \tag{7-2b}$$

If $\alpha(t)$ is assumed to have the popular bandlimited nonrational spectrum given by

$$S(f) = \begin{cases} \frac{2}{\pi}\left[1 - \left(\frac{f}{f_D}\right)^2\right]^{-1/2} & |f| < f_D \\ 0 & \textit{otherwise} \end{cases} \tag{7-3}$$

then the covariance function for $\alpha(t)$ is given as [Zorzi, Rao & Milstein, 1997]

$$K(\tau) = J_0(2\pi f_D|\tau|) \tag{7-4}$$

The parameter f_D is known as the channel Doppler bandwidth, and $J_0(.)$ is the Bessel function of the first kind and zeroth order. The channel coherence time is inversely proportional to f_D. In a cellular environment, the Doppler bandwidth is determined primarily by the speed and transmitting frequency of the mobile terminal. Parameter τ refers to the time between two consecutive samples of $|\alpha(t)|$. In equation 7-4, we see that when τ is small, the correlation between the two samples of $|\alpha(t)|$ will be large. Naturally, as τ increases, the amount of correlation between the two samples will decrease. The relationship between the Doppler frequency and the time between samples determines the "speed" of the fading process. Generally, if the product $f_D|\tau|$ is very small (<0.1) then the fading process is considered to be "slow". If $f_D|\tau|$ >0.2, then the channel fading is considered to be "fast" [Zorzi, Rao & Milstein, 1997]. In fast fading channels, two consecutive samples of $|\alpha(t)|$ will be almost independent. Assumption (5) of our signal model means that the duration of a slot (packet) is smaller than the coherence time of the channel. The duration of a slot size is defined in Chapter 3 as t_s and therefore the time between two consecutive channel "samples" is effectively $\tau = t_s$.

In order to account for the correlated nature of the samples in the fading channel in a probabilistic manner, we require an expression for the probability of receiving a certain amplitude u_n given the previous "history" of the channel. Recently, Wang & Chang [1996] have shown that a one-step Markov process is adequate in describing the complex Gaussian process described above. The fading process can then be considered to be first-order fading and the following expression, which has found common use throughout the years (e.g. [Jakes, 1974], [Bischl & Lutz, 1995] and [Turin & Nobelen, 1998]) can be used for the conditional PDF of the Rayleigh fading channel

$$P_{sig}(u_n|u_{n-1}) = \frac{P_{sig}(u_n,u_{n-1})}{P_{sig}(u_{n-1})} = \frac{u_n}{\mu_U^2(1-\rho_U^2)}.\exp\left(\frac{-\rho_U^2 u_{n-1}^2 - u_n^2}{2\mu_U(1-\rho_U^2)}\right).I_0\left(\frac{u_{n-1}u_n\rho_U}{\mu_U(1-\rho_U^2)}\right) \quad \textbf{(7-5)}$$

$P_{sig}(u_n,u_{n-1})$ is the joint probability density function for two consecutive samples:

$$P_{sig}(u_n,u_{n-1}) = \frac{u_n u_{n-1}}{\mu_U^2(1-\rho_U^2)}.\exp\left(\frac{-(u_n^2+u_{n-1}^2)}{2\mu_U(1-\rho_U^2)}\right).I_0\left(\frac{u_n u_{n-1}\rho_U}{\mu_U(1-\rho_U^2)}\right), \quad \textbf{(7-6)}$$

$I_0(.)$ is the modified Bessel function of zeroth order. The term ρ_U is known as the *correlation coefficient* of u_n and u_{n-1} and, for the Gaussian model considered, is given by [Jakes, 1974], [Zorzi, Rao & Milstein, 1997], i.e.

$$\rho_U = \frac{K(t_s)}{K(0)} = J_0(2\pi f_D t_s) \quad \textbf{(7-7)}$$

7.3.2 MAI Model (*Y*)

The interference amplitude received at the base station from the interfering terminals (*Y*) is obviously going to be a random variable which is a direct summation of the independent Rayleigh distributed amplitude levels received from each terminal. The probability distribution function for such a sum of independent Rayleigh random variables is not easy to compute however, especially when the means and variances of the constituent summands are different. This problem has been a topic of interest for over 70 years. We point the interested reader to [Beaulieu, 1990] and the references cited therein, for discussions on the problems associated with finding the distribution for *Y*. Various approximations and computations have been presented in the literature [Beaulieu, 1990], [Mason, Ginsburg & Brennan, 1960]. However, no simple closed form expressions for the distribution of *Y* have been found to date.

Fortunately, if the Rayleigh random variables are identically distributed however, a simple approximation for the problem exists. In our model, we have assumed that all users share identical Rayleigh fading properties and, as such, the means and variances of the constituent summands in *Y* are identical, as well as being independent. A common approximation, which is fairly accurate, is to assume that the distribution of *Y* is Gaussian. In [Geraniotis & Pursley, 1985], [Geraniotis & Pursley, 1986] and [Geraniotis, 1986], accurate expressions for the SIR and BER in multipath fading channels are derived. Their work also includes a rigorous comparison of the Gaussian approximation of the BER with more accurate measurements and approximations, where they assume that the MAI in the fading channel is in fact Gaussian (they consider both Rayleigh fading and Rician fading). Their results yield that the Gaussian approximation is satisfactory, giving only slightly optimistic BER results (slight underestimation of the MAI), even for a very low number of interfering users (j). The accuracy of the Gaussian approximation is seen to improve with increasing j.

The mean and variance of the Gaussian distribution is simply equal to the sum of the means and the sum of the variances of the constituent summand Rayleigh distributions, respectively. For example, the sum of k independent Rayleigh random variables, each with mean μ_U and variance σ_U^2 (equation 7-2), can be reasonably well approximated by a normal Gaussian distribution function, with mean $\mu_Y(k) = k.\mu_U$ and variance $\sigma_Y^2(k) = k.\sigma_U^2$. We define the expression for the probability distribution function for the total MAI signal amplitude (*Y*) from j interfering users (summands) as

$$P_{int}(y|j) = \begin{cases} \dfrac{y}{b^2}\exp\left(\dfrac{-y^2}{2b^2}\right) & j = 1 \\ \\ \dfrac{1}{\sqrt{2\pi j\sigma_U^2}}\exp\left(\dfrac{-(y - j\mu_U)^2}{2j\sigma_U^2}\right) & j > 1 \end{cases} \qquad \textbf{(7-8)}$$

where the Rayleigh distribution can be used when there is only one interferer. For $j > 1$ interferers, the Gaussian distribution with mean $j\mu_U$ and variance $j\sigma_U^2$ is used. The subscript *int* is used to indicate the received *interference* signal.

An expression for the conditional probability density function ($P\{y_n|y_{n-1}\}$) that captures the correlation between successive samples of the interference signal is far more difficult (and in fact analytically and computationally intractable) to obtain than for a single Rayleigh fading signal (equation 7-5). The main reason for this is that, although the total sum of the interference power is known, the individual power contribution from each constituent interferer is not known, and as these individual interferers enter and leave the channel, the exact changes they inflict on the interference sum are not known. In [Chuah, Nanda & Rege, 1998], they compute the one-step correlation coefficient for *Y* when the number of users remains constant and the channel is effected by log-normal shadowing, instead of Rayleigh fading. As soon as the number of users becomes variable, the problem becomes significantly more complex.

Since the goal of this work is to construct a simple channel Markov model, we wish to avoid complex and computationally intensive mathematical derivations that detract from this goal. A preliminary attempt at solving for $P\{y_n|y_{n-1}\}$ revealed a multi-dimensional integration problem that considers every possible user scenario, the contribution of each user to the MAI, and the user transition statistics (given by, say, equation 4-10). The immense increase in complexity incurred, together with the increased time required to solve the model, is certainly enough to preclude the use of such a mathematical burden in favour of a simple assumption that does not appear to be too unreasonable or limiting. Fortunately, the problem is simplified by the fact that the users enter and leave the channel in a reasonably uncorrelated random manner according to the multiple birth-death process (characterised by the transition matrix in, say equation 3-6 or equation 4-10). This fact that the number of constituent summands is fluctuating in a reasonably uncorrelated manner and at a reasonably high rate, even when the fading process for each user is fairly slow, means that the sequence of interference samples in signal *Y* should appear to be significantly less correlated than that of each constituent interfering signal *U*. In figure 7.4, we choose to test this assumption through simulation.

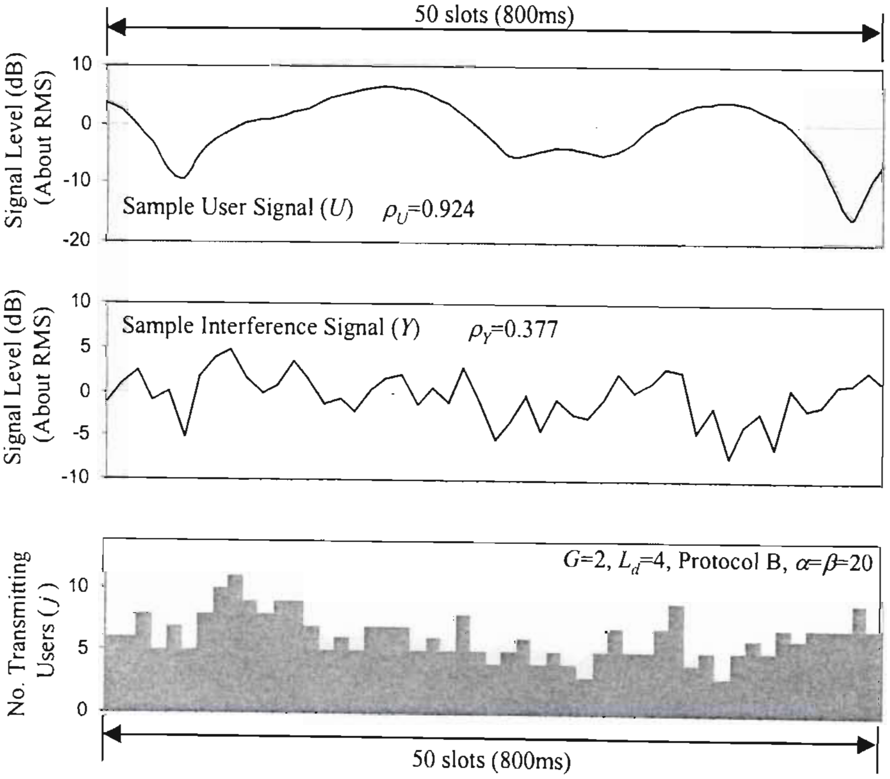


Figure 7.4 : Simulation sample sequence of values for *j*, *U* and *Y*

Figure 7.4 shows an example sample sequence of channel events, as obtained through simulation, over 50 slots. Included is a sample Rayleigh faded user signal (*U*), the total MAI (*Y*) and the number of transmitting users (*j*). The protocol used in this simulation is the SR-DT/CDMA protocol of Chapter 4 with the following parameters: mean message length $\overline{L}_d = 4$; thresholds $\alpha = \beta = 20$; expected message arrival rate $G_d = 2$ messages per slot (which is a reasonably low offered load in the context of the protocol); and a Rayleigh fading channel with parameters b=1, $f_D t_s$=0.05. The simulation algorithm used to generate sequences of correlated Rayleigh random variates with the covariance function given in equation 7-4 is the well known and popular method proposed by Jakes [Jakes, 1974].

As can be seen from figure 7.4, the protocol is such that the number of transmitting users (j) is fairly rapidly fluctuating from slot to slot and the sequence of j values appears to be reasonably uncorrelated over short sequences. This is important in the context of the solution of the network performance, where we are only interested in the channel statistics over the duration of a transmitted message. Since the expected length of a data message in this scenario is $\overline{L}_d = 4$, with long messages being rare, we are only interested in the network states over a handful of slots. We can also see that the trends in signal Y follow the trends in j. We should expect this, since as j increases, we expect to see Y increase and as j decreases, we expect to see Y decrease. If we compute the correlation coefficient of the given simulation sample sequence of Y values in figure 7.4, and define this as ρ_Y, we find $\rho_Y = 0.377$. The correlation coefficient of signal U with parameter $f_D t_s = 0.05$, defined in equation 7-7, was found to be $\rho_U = 0.924$. As can be seen, the correlation coefficient for the sequence of Y samples is significantly smaller than the correlation coefficient of the Rayleigh fading channel. This process was repeated for several network scenarios, and in all cases it was found that ρ_Y was always significantly smaller than ρ_U except, of course, for the cases when ρ_U is small (i.e. in cases where the Rayleigh channel is highly uncorrelated), in which case ρ_U and ρ_Y become comparable. Although it is not completely uncorrelated, the fact that the MAI signal Y is significantly less correlated than signal U might be reason enough to consider it practically uncorrelated. We thus make a significant simplifying assumption that the *levels of MAI in consecutive slots are independent*. This assumption will be shown to simplify the parameterisation of our model significantly, while still providing reasonable results.

7.4 Proposed Markov model for a Multi-user DS-CDMA Channel

In this section, we present our proposed hidden Markov fading model for a reference terminal in the DS-CDMA reverse link. We then apply our HMM to obtain new approximations for the probabilities of packet and message success in the fading channel, and test these approximations by employing them in our analysis of the data throughput of the network.

7.4.1 Description of the relationship between the SIR and the channel state

As with the narrowband Gilbert-Elliot model, we assume that the channel "observed" by each terminal in the network can exist in one of two states, either "Good" ("non-faded") or "Bad" ("faded"). Naturally, this channel state should be a function of the SIR experienced by the reference terminal, i.e. the ratio between *U* and *Y*. In figure 7.5, we present example signal samples, and introduce the relationship between the fade

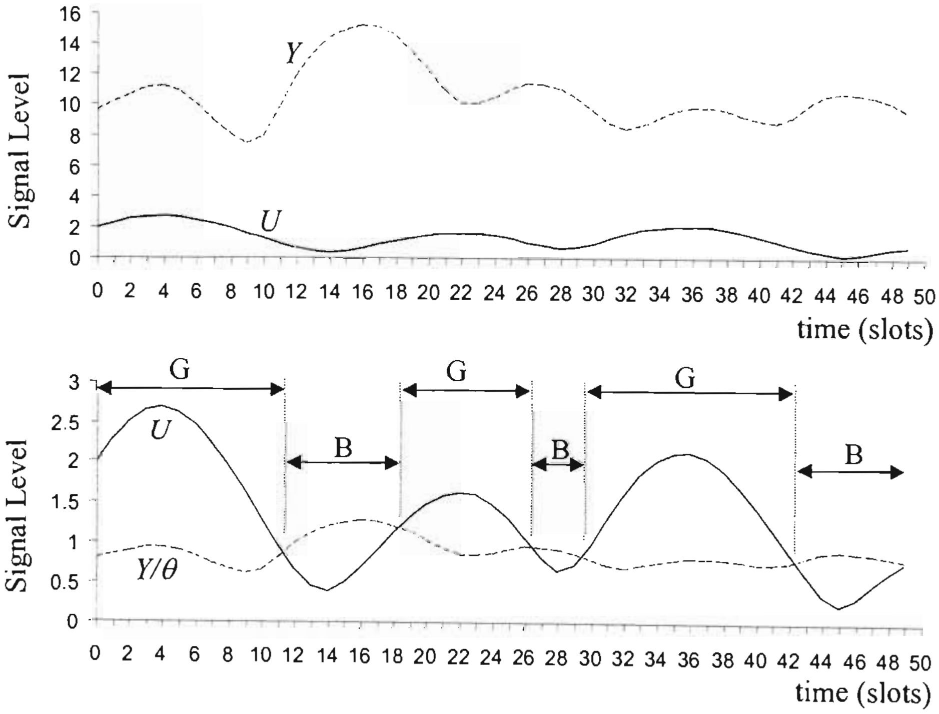


Figure 7.5 : Illustration of the relationship between the received signal amplitudes (arbitrary units) from both a reference user (*U*) and the total other-user interference (*Y*), and the resultant transitions between the imagined Good (G) and Bad (B) states. θ is the *fading ratio*.

state of a terminal and its received power (*U*) with respect to the MAI (*Y*). Since a spread-spectrum receiver is intended to operate at negative SIR's, we expect the level of MAI to be significantly larger than *U*. For example, in a perfect channel scenario in which there are *j* users transmitting in total, we would obviously expect *Y* to be exactly (*j*-1) times larger than *U*. For a non-perfect channel, we would expect to see variations in *U* and *Y* as in figure 7.5, where the *mean* of *Y* is approximately (*j*-1) times the mean of *U*.

Another important property of a spread-spectrum system, which is relevant to our channel model, is that the larger the ratio of MAI to user power (i.e. *Y*/*U*) the higher the probability of packet error is for the reference user. This brings us to the idea of imposing a threshold, defined as the ***fading ratio*** θ, on the ratio *Y*/*U*, such that a terminal is:

in the ***BAD (fade)*** state if $\frac{Y}{U} > \theta$,

or in the ***GOOD (non-fade)*** state if $\frac{Y}{U} < \theta$.

Alternatively, it is easier to see that the reference terminal is in the ***GOOD*** state if its received power *U* is less than Y/θ, or in the ***BAD*** state if *U* is greater than Y/θ, as illustrated in figure 7.5. We have assumed that the signals received at the base station remain constant over the duration of a system slot. Another assumption made in Chapter 3 is that all users are assumed to be exactly slot synchronised with the base station, such that the number of transmitting users in each slot is constant, and can only change at each slot boundary. As a result of these two assumptions, it is obvious that the state of the channel observed by each terminal remains the same over each slot, and can only change at each slot boundary.

7.4.2 Description of the Markov model

We begin by defining the fade state of the channel as seen by a reference terminal as Ω, where Ω belongs to the set of possible channel states defined as $\mathbf{\Omega} = \{\text{Bad}, \text{Good}\}$. We also define the joint channel state $(\Omega, \hat{x})$ as the case where the reference terminal observes the channel to be in state Ω while there are $\hat{x}$ interfering users in the channel. Based on this choice of state definitions, we can characterise the joint state of the channel by the stationary first-order Markov chain $(\Omega, \hat{x})$, which has steady state occupancy probabilities $P\{\Omega = C, \hat{x} = j\}$ and state transition probabilities $P\{\Omega_{n+1} = D, \hat{x}_{n+1} = j | \Omega_n = C, \hat{x}_n = i\}$, where $C, D \in \mathbf{\Omega}$.

For the sake of simplifying the notation, we define G_j as the joint state $(\Omega = \mathrm{Good}, \hat{x} = j)$ and B_j as the joint state $(\Omega = \mathrm{Bad}, \hat{x} = j)$. Also $P(\mathrm{G}_j) = P\{\Omega = \mathrm{Good}, \hat{x} = j\}$ and $P(\mathrm{B}_j) = P\{\Omega = \mathrm{Bad}, \hat{x} = j\}$. We also simplify the notation for the joint state transition probabilities by defining $\omega_{ij}^{CD} = P\{\Omega_{n+1} = D, \hat{x}_{n+1} = j | \Omega_n = C, \hat{x}_n = i\}$, where $C, D \in \mathbf{\Omega}$. For example, $\omega_{02}^{GB} = P\{\Omega_{n+1} = \mathrm{Bad}, \hat{x}_{n+1} = 2 | \Omega_n = \mathrm{Good}, \hat{x}_n = 0\}$ is the probability that the terminal transitions from the joint state of a Good channel with zero interferers in slot n (i.e. state G_0) to the joint state of a Bad channel with two interferers in slot n+1 (i.e. state B_2). Based on these simpler notations, the complete Markov chain description is illustrated graphically in figure 7.6. To avoid a messy diagram, only a few transition probabilities are included. In the full model, each state may transition to any other state.

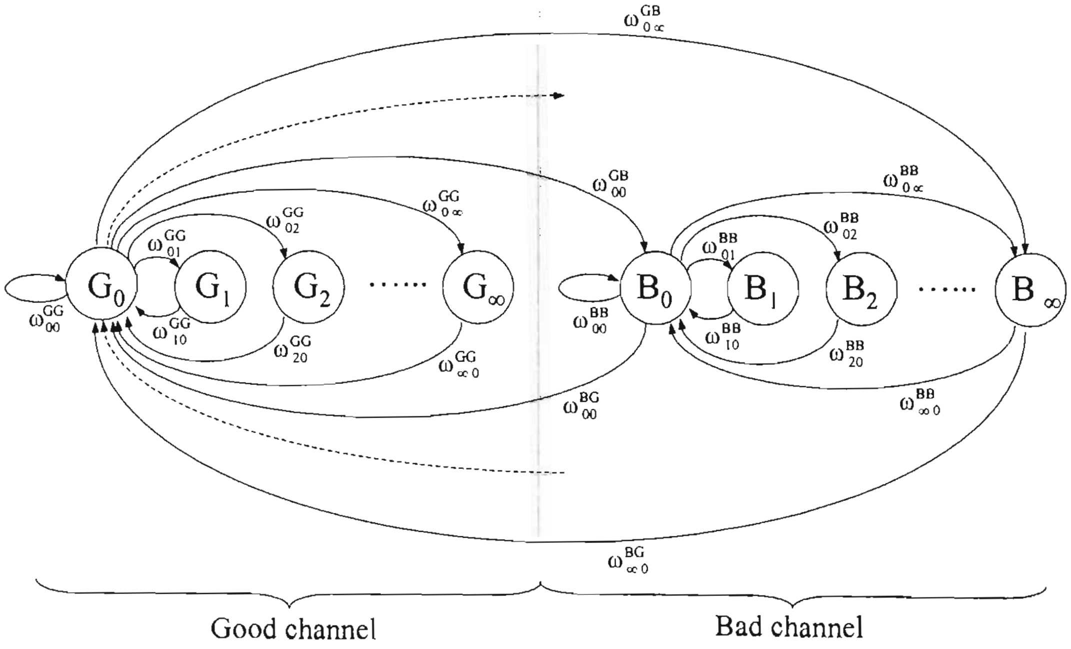


Figure 7.6 : Proposed multi-user Markov chain for the channel model

Mathematically, the following set of linear equations completes the description of the Markov chain

$$P(G_j) = \sum_{i=0}^{\infty} P(G_i).\omega_{ij}^{GG} + \sum_{i=0}^{\infty} P(B_i).\omega_{ij}^{BG} \tag{7-9}$$

$$P(B_j) = \sum_{i=0}^{\infty} P(G_i).\omega_{ij}^{GB} + \sum_{i=0}^{\infty} P(B_i).\omega_{ij}^{BB} \tag{7-10}$$

$$\sum_{i=0}^{\infty} \left[\mathrm{P}(G_i) + P(B_i)\right] = 1 \tag{7-11}$$

7.4.3 Computation of the stationary state occupancy and transition probabilities

The joint state occupancy probabilities $P(\mathrm{G}_j)$ and $P(\mathrm{B}_j)$ can be found as follows. Using Bayes' theorem, we can rewrite these joint state probabilities as

$$\begin{aligned} P(\mathrm{G}_j) &= P\{\Omega = \mathrm{Good}, \hat{x} = j\} \\ &= P\{\Omega = \mathrm{Good} | \hat{x} = j\}.P\{\hat{x} = j\} \end{aligned} \tag{7-12}$$

$$\begin{aligned} P(\mathrm{B}_j) &= P\{\Omega = \mathrm{Bad}, \hat{x} = j\} \\ &= P\{\Omega = \mathrm{Bad} | \hat{x} = j\}.P\{\hat{x} = j\} \end{aligned} \tag{7-13}$$

where $P\{\hat{x} = j\}$ is the steady state probability of finding j interferers in the channel (obtained by solving equations 4-10, 4-11 and 4-12 in the SR-DT/CDMA model), and $P\{\Omega = D | \hat{x} = j\}, D \in \mathbf{\Omega}$ is the conditional probability of being in state $\Omega = D$ given that there are j interferers. $P\{\Omega = D | \hat{x} = j\}, D \in \mathbf{\Omega}$ can be easily found as follows

$$P\{\Omega = \mathrm{Good} | \hat{x} = j\} = \int_{0}^{\infty} \int_{\frac{y}{\theta}}^{\infty} P_{int}(y|j).P_{sig}(u)dudy \tag{7-14}$$

$$P\{\Omega = \mathrm{Bad} | \hat{x} = j\} = 1 - P\{\Omega = \mathrm{Good} | \hat{x} = j\} \tag{7-15}$$

where $P_{int}(y|j)$ and $P_{sig}(u)$ are given by equations 7-8 and 7-1, respectively. In equation 7-14 we integrate over all possible levels of interference in the presence of j interferers and condition on the fact that the reference user's amplitude must be larger than the fading threshold (i.e. $u > y/\theta$) in order to satisfy the criterion for a Good (non-fade) state. The probability of the reference user being in the Bad (fade) state given j interferers is obviously one minus the probability of being in the Good state given j interferers. The joint state probabilities $P(\mathrm{G}_j)$ and $P(\mathrm{B}_j)$ can now be found by substituting equations 7-14 and 7-15 into equations 7-12 and 7-13, respectively. The joint state transition probabilities ω_{ij}^{CD} can be found as follows. Since the channel state depends on the number of transmitting users, whereas the number of transmitting users is completely independent of the channel state, we can rewrite ω_{ij}^{CD} as

$$\omega_{ij}^{CD} = P\{\Omega_{n+1} = D|\Omega_n = C, \hat{x}_{n+1} = j, \hat{x}_n = i\}.P\{\hat{x}_{n+1} = j|\hat{x}_n = i\} \quad C, D \in \mathbf{\Omega} \qquad \textbf{(7-16)}$$

The probability $P\{\hat{x}_{n+1} = j|\hat{x}_n = i\}$ is simply the transition probability for the number of users in the channel, and can be obtained from equation 4-10, i.e. $P\{\hat{x}_{n+1} = j|\hat{x}_n = i\} = \pi_{ij}$. The probability $P\{\Omega_{n+1} = D|\Omega_n = C, \hat{x}_{n+1} = j, \hat{x}_n = i\}$ represents the conditional probability of the channel being in state $\Omega_{n+1} = D$ in the current slot, given that the channel was in state $\Omega_n = C$ in the previous slot, and given that there were i interferers in the previous slot and j interferers in the current slot. Again, we can make use of Bayes' theorem and rewrite $P\{\Omega_{n+1} = D|\Omega_n = C, \hat{x}_{n+1} = j, \hat{x}_n = i\}$ as

$$P\{\Omega_{n+1} = D|\Omega_n = C, \hat{x}_{n+1} = j, \hat{x}_n = i\} = \frac{P\{\Omega_{n+1} = D, \Omega_n = C|\hat{x}_{n+1} = j, \hat{x}_n = i\}}{P\{\Omega_n = C|\hat{x}_n = i\}} \qquad \textbf{(7-17)}$$

where $P\{\Omega_{n+1} = D, \Omega_n = C|\hat{x}_{n+1} = j, \hat{x}_n = i\}$ represents the joint state probability of being in channel states $\Omega_n = C$ and $\Omega_{n+1} = D$ in consecutive slots n and n+1, given that the number of users in slot n is i and in slot n+1 is j. Since we are assuming a stationary channel, we can ignore subscript n and thus $P\{\Omega_n = C|\hat{x}_n = i\}$ is given by equations 7-14 and 7-15.

In order to compute the conditional probabilities $P\{\Omega_{n+1} = D, \Omega_n = C|\hat{x}_{n+1} = j, \hat{x}_n = i\}$, we need to know the conditional probability density functions for both signals U and Y. The reason for this is that the channel state transition probability is a function of the correlation between the consecutive signal amplitudes in the current and next slots. Obviously, if the channel fading is very slow, then we would expect the channel state of a terminal to remain the same in the next slot. If the channel fading is very fast, then the state of the terminal in the following slot would be more or less independent of the state of the terminal in the current slot. In section 7.3.1, we provided equation 7-5 for the conditional probability density function for the reference user's signal U. In section 7.3.2, we made the simplifying assumption that the level of sequence of Y samples is uncorrelated and hence successive Y samples are independent. Using this assumption of independence in Y, we can derive the following expressions for the four possible joint state conditional probabilities (Bad, Bad), (Good, Bad), (Bad, Good) and (Good, Good).

$$P\{\Omega_{n+1} = \text{Bad}, \Omega_n = \text{Bad}|\hat{x}_{n+1} = j, \hat{x}_n = i\}$$

$$= \int_0^\infty\int_0^\infty P_{sig}(u_n).[1 - P_{int}(y_n < \theta u_n|i)].P_{int}(y_{n+1}|j).\left[\int_0^{\frac{y_{n+1}}{\theta}} P_{sig}(u_{n+1}|u_n)du_{n+1}\right]du_n dy_{n+1}$$

(7-18)

$$P\{\Omega_{n+1} = \text{Good}, \Omega_n = \text{Bad}|\hat{x}_{n+1} = j, \hat{x}_n = i\}$$

$$= \int_0^\infty\int_0^\infty P_{sig}(u_n).[1 - P_{int}(y_n < \theta u_n|i)].P_{int}(y_{n+1}|j).\left[1 - \int_0^{\frac{y_{n+1}}{\theta}} P_{sig}(u_{n+1}|u_n)du_{n+1}\right]du_n dy_{n+1}$$

(7-19)

$$P\{\Omega_{n+1} = \text{Bad}, \Omega_n = \text{Good}|\hat{x}_{n+1} = j, \hat{x}_n = i\}$$

$$= \int_0^\infty\int_0^\infty P_{sig}(u_n).P_{int}(y_n < \theta u_n|i).P_{int}(y_{n+1}|j).\left[\int_0^{\frac{y_{n+1}}{\theta}} P_{sig}(u_{n+1}|u_n)du_{n+1}\right]du_n dy_{n+1}$$

(7-20)

$$P\{\Omega_{n+1}=\text{Good},\Omega_n=\text{Good}|\hat{x}_{n+1}=j,\hat{x}_n=i\}$$

$$=\int_0^\infty\int_0^\infty \underbrace{P_{sig}(u_n)}_{A}.\underbrace{P_{int}(y_n<\theta u_n|i)}_{B}.\underbrace{P_{int}(y_{n+1}|j)}_{C}.\underbrace{\left[1-\int_0^{\frac{y_{n+1}}{\theta}} P_{sig}(u_{n+1}|u_n)du_{n+1}\right]}_{D}du_n dy_{n+1} \qquad \textbf{(7-21)}$$

We explain the derivation of these equations using the following thought process. In equations 7-18 to 7-21, y_n and y_{n+1} represent the received interference amplitudes in slots n (i transmitting users) and n+1 (j transmitting users) respectively, while u_n and u_{n+1} represent the reference user's received signal in slots n and n+1 respectively. Consider the last case $P\{\Omega_{n+1}=\text{Good},\Omega_n=\text{Good}|\hat{x}_{n+1}=j,\hat{x}_n=i\}$ in equation 7-21. The probability of staying in the Good (non-fade) state depends only on the fact that the reference user's signal is above the fading threshold over both slots. In the current slot, we see that for a given reference user amplitude u_n (term A), the interference amplitude must be less than θu_n (term B). Then, if the interference amplitude in the following slot changes, independently of the amplitude in the current slot, to y_{n+1} (term C) then, in order for the reference user to remain in the Good state, its amplitude can not change from u_n to anything less than y_{n+1}/θ (term D). Using similar logic, the other three joint channel state probabilities can be easily found. The term $P_{int}(y<\phi|j)$ in equations 7-18 to 7-21 is given by

$$P_{int}(y<\phi|j)=\int_0^\phi P_{int}(y|j)dy \qquad \textbf{(7-22)}$$

Using equations 7-18 to 7-21, we can now compute the conditional channel state probabilities, $P\{\Omega_{n+1}=D|\Omega_n=C,\hat{x}_{n+1}=j,\hat{x}_n=i\}$, using equation 7-17, and hence the joint state transition probabilities ω_{ij}^{CD}, using equation 7-16. Of the four possible conditional channel transitions in equations 7-18 to 7-21, the one that interests us most is $P\{\Omega_{n+1}=\text{Good}|\Omega_n=\text{Good},\hat{x}_{n+1}=j,\hat{x}_n=i\}$, since this will be used when computing the success of a message containing several packets, where we require the channel to remain Good over all packets in the message. For this purpose, we define

$$C_{ij} = P\{\Omega_{n+1} = \text{Good} | \Omega_n = \text{Good}, \hat{x}_{n+1} = j, \hat{x}_n = i\} \quad \textbf{(7-23)}$$

In summary, the main differences between our MM and the narrowband Gilbert-Elliot model (figure 7.1) are the following: Firstly, in our model the interference level Y is now a time varying value that is a function of both the number of users in the channel and the amount of power that each user contributes to the level of interference, rather than a time invariant value that is fixed at some fraction of the narrowband channel AWGN. Secondly, and perhaps more importantly, since Y is a function of the number of interferers in the channel, our MM becomes application specific in that the application of an admission/transmission control protocol will affect the number of transmitting users and hence the parameters of the model. Fortunately, in our case the number of users in the channel is rapidly fluctuating due to the nature of the traffic model and admission scheme, and this allows us to assume that the level of MAI in the channel is uncorrelated. In the narrowband Gilbert-Elliot model, the choice of state was dependent only on the fading statistics and choice of the fixed noise threshold Y.

7.4.4 Proposed mapping of the channel state onto the packet error process

Although we have completed the description and statistics of the fading channel, we are ultimately more interested in the sequence of errors that occur in the channel, rather than the actual sequence of channel states. In section 7.2, we discussed the fact that the sequence of errors that is obtained by mapping the sequence of Good/Bad (non-fade/fade states) onto the error sequence depends on the error process that occurs in each state. If we recall that $\mathbf{\Omega} = \{\text{Good}, \text{Bad}\}$ is the set of possible channel states, and define $\mathbf{E} = \{\text{Error}, \text{Correct}\}$ as the set of possible error observations based on $\mathbf{\Omega}$, then $\mathbf{\Psi}$ is a matrix containing the probability of error output $E \in \mathbf{E}$ conditioned on each $\Omega \in \mathbf{\Omega}$, such that

$$\mathbf{P(E)} = \mathbf{\Psi} \cdot \mathbf{P(\Omega)} \quad \textbf{(7-24)}$$

Since the two sets $\mathbf{E}$ and $\mathbf{\Omega}$ have only two states each, matrix $\mathbf{\Psi}$ will be a two-by-two matrix containing a function of two mapping probabilities. If we consider that the only possible mapping scheme possible for this scenario is as follows:

- In the Bad channel state an error occurs with large probability h, or alternatively, information is correct with probability 1-h
- In the Good channel state, an error occurs with small probability k, or alternatively, information is correct with probability 1-k

then
$$\mathbf{\Psi} = \begin{bmatrix} (1-k) & (1-h) \\ k & h \end{bmatrix} \qquad \textbf{(7-25)}$$

Since our previous analyses have been conditioned mostly on the number of transmitting users in the channel (e.g. equation 4-13, 4-14 and 4-15), we require expressions for the probability of error conditioned on the number of transmitting users. Since matrix $\mathbf{\Psi}$ is independent of $\hat{x}$, we can compute the conditional probability density function for the error event as

$$P\{E = \text{correct}|\hat{x} = j\} = P\{\Omega = \text{Good}|\hat{x} = j\}.(1-k) + P\{\Omega = \text{Bad}|\hat{x} = j\}.(1-h) \qquad \textbf{(7-26)}$$

$$P\{E = \text{error}|\hat{x} = j\} = P\{\Omega = \text{Good}|\hat{x} = j\}.k + P\{\Omega = \text{Bad}|\hat{x} = j\}.h \qquad \textbf{(7-27)}$$

7.4.5 Computations of the packet success probability

So far, our HMM has considered the sequence of errors and channel states that occur from slot to slot, as seen by a reference terminal. Since this observed channel is also used by the terminal to transmit its data, where each packet in the data message is transmitted in a slot, it is obvious that the success of a transmitted packet is simply equal to the probability that an error event does not occur in that slot. So, conditioning on the number of users in a slot, the probability of packet success is given simply by

$$P_{PS}^{MM}(j) = P\{E = \text{correct}|\hat{x} = j\} \qquad \textbf{(7-28)}$$

where the superscript "*MM*" denotes that this is an approximation obtained from the application of our *M*arkov *M*odel. In this thesis, we will only examine the case (h=1, k=0) in equations 7-26 and 7-27, but show that surprisingly accurate results can be obtained from this approximation.

As a matter of interest, and to conclude the error analysis, we also compute the *EX*act probability of packet success for the fading channel in the presence of $\hat{x} = j$ interferers (denoted by superscript "*EX*") as

$$P_{PS}^{EX}(j) = \int_0^\infty\int_0^\infty P_{sig}(u).P_{int}(y|j).\left[1-\frac{1}{2}erfc\left(\sqrt{\frac{3Nu}{2y}}\right)\right]^Q dudy$$

$$= \begin{cases} 1 & j=0 \\ \dfrac{1}{b^4}\displaystyle\int_0^\infty\int_0^\infty uy.\exp\left(\frac{-u^2-y^2}{2b^2}\right).\left[1-\frac{1}{2}erfc\left(\sqrt{\frac{3Nu}{2y}}\right)\right]^Q dudy & j=1 \\ \dfrac{1}{b^2\sqrt{2\pi j\sigma_R^2}}\displaystyle\int_0^\infty\int_0^\infty u.\exp\left(\frac{-(y-j\mu_R)^2}{2j\sigma_R^2}-\frac{u^2}{2b^2}\right).\left[1-\frac{1}{2}erfc\left(\sqrt{\frac{3Nu}{2y}}\right)\right]^Q dudy & j>1 \end{cases}$$

(7-29)

Equation 7-29 is obtained by integrating over the distributions for the received signal (u) and the interference signal (y) in the SGA expression for the probability of packet error for a packet that contains Q bits, and as such effectively considers all possibilities for the exact SIR. When there is only one other transmitting user (i.e. $j=1$) then the Rayleigh distribution is used for the distribution of the MAI, while the Gaussian approximation in equation 7-8 is used for two or more interferers. If there are no interfering users, then the probability of packet success is assumed to be one. Equation 7-29 can then be used to judge the success of the Markov model approximation in equation 7-28. Although equation 7-29 should be more accurate than equation 7-28 in determining the probability of an error event in a slot, it does not model the correlated fading in the channel in any way. In our Markov model, we have sacrificed accuracy in order to model the statistics of the *sequence of error events* in the fading channel, instead of just the probability of an error event.

7.4.6 Computation of the message success probability

A new expression for the probability of success of a reference message has to be derived based on the new expressions for the probability of packet success, as well as the probability that the reference terminal remains in the non-fade state over the duration of the message transmission. We modify equations 4-14 and 4-15 in order to accommodate the MM. Obviously, the success of the first packet in the message (refer to equation 4-14) is now also conditioned on the fact that the reference terminal is in the non-fade state given that it finds j other messages transmitting. Using equation 7-28 now to express the probability of packet success, we can rewrite equation 4-14 as

$$R_1(j) = X(j).[1 - P_{MB}(j)].P_{PS}^{MM}(j) \qquad \textbf{(7-30)}$$

The recursive procedure to compute the probability of success of the remaining packets in the message is now slightly different since we have to account for the cases where there are channel collisions. In C_{ij} (equation 7-23), the probability of remaining in a non-fade state over two consecutive slots is conditioned on the number of transmitting terminals in each of those two slots. Recalling that when a "collision" occurs, we assume that those users that arrive to cause the collision (in equations 4-10, 4-13 and 4-16 the dummy variable m refers to those users) add to the MAI in the slot in which they arrive before being dropped. This is explained after equations 4-10 and 4-13. Before, we ignored these users since a collision in slot t-1 had no affect on the probability of success of packets in slot t. Now however, the probability of remaining in a non-fade state in slot t depends on these additional users since they determined whether the reference terminal was in a fade state or a non-fade state in slot t-1.

The following steps in the recursive algorithm can be derived as follows:

$$R_2(j,m) = \begin{cases} \sum_{i=0}^{\beta-1} R_1(i).\hat{\pi}_{ijm}.C_{im} & m = j \text{ or } m \geq \beta \\ 0 & m \neq j \text{ and } m < \beta \end{cases} \qquad \textbf{(7-31)}$$

$$R_n(j,m) = \begin{cases} \sum_{i=0}^{\beta-1} \sum_{l=0}^{\infty} R_{n-1}(i,l).\hat{\pi}_{ijm}.C_{lm} & n > 2, m = j \text{ or } m \geq \beta \\ 0 & n > 2, m \neq j \text{ and } m < \beta \end{cases} \qquad \textbf{(7-32)}$$

where we must now record the fact that a collision can occur during transmission of the current packet. While j is the number of other transmitting users as seen by the reference message *after* message dropping has occurred (as in equation 4-16), m is the number of transmitting users as seen by the reference message *before* the offending collision messages are dropped. This means that, in the n^{th} slot of a message's transmission (n>1), j other admitted messages were in the process of being transmitted when $m-j$ new users arrived, caused a collision, and were then dropped before the end of the slot. As in equation 4-16 these $m-j$ additional users must be considered when determining the probability of success of the reference packet. This is captured in the transition probability term $\hat{\pi}_{ijm}$, which now represents the probability that the n^{th} packet in the reference message sees i other *admitted* transmissions in the previous slot, j other *admitted* transmissions in the current slot and an *additional* $m-j$ "contending collision" users in the current slot

$$\hat{\pi}_{ijm} = \begin{cases} 0 & i,j > \beta - 1 \\ 0 & j \le i, m \ne j, m \le \beta - 1 \\ 0 & j > i, m \ne j \\ b(i-j,i,\mu_d) & m = j, j \le i, i > \alpha - 1 \\ \sum\limits_{k=(0,i-j)^+}^{i} f_\infty(j-i+k).b(k,i,\mu_d) & m = j \le \beta - 1, i \le \alpha - 1 \\ b(i-j,i,\mu_d).f_\infty(m-j) & m > \beta - 1, j \le i, i \le \alpha - 1 \end{cases} \qquad \textbf{(7-33)}$$

In equations 7-31 and 7-32, we must also include the conditional joint probability (C_{lm} defined in equation 7-23) that the reference terminal remains in the non-fade state given that in the previous slot there were l transmitting terminals in total (including collision users), and m terminals in total in the current slot.

The probability of success of a message of length L packets is then given by

$$P(S|L) = \sum_{j=0}^{\beta-1} \sum_{m=0}^{\infty} R_L(j,m) \tag{7-34}$$

while the expected probability of success of any message, and the expected message throughput are given in equations 3-25 and 3-27.

7.4.7 Computation of the expected packet throughput

The expected packet throughput is comparatively easy to compute since it is not a function of the degree of correlation in the channel. This is obvious, since the success of a single packet is a function only of the state of the network in the current slot, and does not depend on the success of previous packets, or the state of the network in the previous slot. As in 4-19, we can use the expression

$$S_P = \sum_{j=0}^{\beta} \sum_{i=0}^{\beta} j.X(i).\pi_{ij}^{PS} \tag{7-35}$$

for the expected packet throughput, where $\boldsymbol{\pi}^{PS}$ is the system state transition probability matrix as seen by the base station. Following the logic of equation 4-20, we can give transition probability π_{ij}^{PS} as

$$\pi_{ij}^{PS} = \begin{cases} 0 & i,j \geq \beta \\ \sum_{k=(0,i-j)^+}^{i} f(j-i+k).b(k,i,\mu_d).P_{PS}^{MM}(j-1) & i \leq \alpha, j > i \\ \sum_{k=(0,i-j)^+}^{i} f(j-i+k).b(k,i,\mu_d).P_{PS}^{MM}(j-1) & \\ \quad + b(i-j,i,\mu_d).\sum_{m=\beta+1-j}^{N_d-j} f(m).P_{PS}^{MM}(j+m-1) & i \leq \alpha, j \leq i \\ b(i-j,i,\mu_d).P_{PS}^{MM}(j-1) & i > \alpha, j \leq i \end{cases} \tag{7-36}$$

7.5 Results

In this section, we will evaluate the success of our channel HMM by comparing the *MM* approximation for the probability of packet success (equation 7-28) to simulation results and the near-exact *EX* computation (equation 7-29). We will also evaluate the effect that a correlated Rayleigh fading channel has on our network's message throughput.

7.5.1 Network Parameters

The network parameters used for the results of this chapter are as presented in table 4.1. For the data protocol, we use the SR-DT/CDMA protocol with an infinite population (Poisson model). For the Rayleigh fading process, we assume a Rayleigh parameter b=1 for each user's received signal at the base station. For the error process, we assume that errors occur in the fade state with probability h=1, and in the non-fade state with probability k=0.

7.5.2 Determination of the fading threshold ratio θ.

In this section, we find a value for the fading threshold ratio, θ, which gives the closest analytical approximation to simulation results. At first glance, it is not easy to imagine what value of θ is appropriate, or even in what order of magnitude it should lie. Does an SIR of –40dB (Y/U=100) constitute a fade state, or is say an SIR of -6dB (Y/U=2) enough to constitute a fade? Recall that in this analysis, we compute the SIR in terms of the received signal amplitudes (the signal amplitude is directly proportional to the square of the signal power).

The easiest way to determine θ is to attempt to match equations 7-28 and 7-29 for the probability of packet success, i.e. find the value of θ for which the mean difference between $P_{PS}^{MM}(j)$ and $P_{PS}^{EX}(j)$ is minimised over the region of interest. Alternatively, we can obtain values for the probability of packet success from the simulation, and use these values instead of $P_{PS}^{EX}(j)$. In figure 7.7, we plot all three sets of results

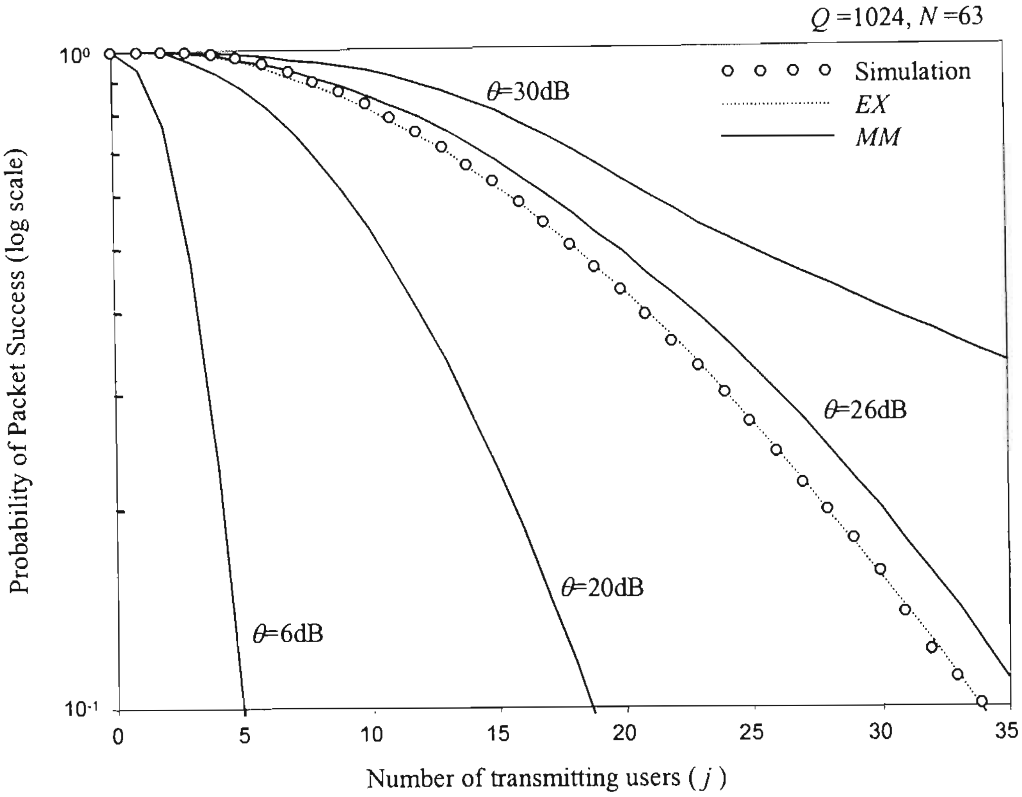


Figure 7.7 : Probability of packet success vs. the number of transmitting users for various θ (course resolution).

(simulation, $P_{PS}^{EX}(j)$ and $P_{PS}^{MM}(j)$) for several preliminary (course) values of θ. As can be seen, the near-exact probability expression $P_{PS}^{EX}(j)$ matches up to the success probabilities obtained from the simulation. This result confirms that the Gaussian approximation we made in section 7.3.2 for the interference distribution, $P_{int}(y|j)$, causes negligible error in the result of equation 7-29. These preliminary results in figure 7.7 reveal that θ should be slightly less than 26dB.

The shape and trend of the curve for the *MM* approximation for θ=26dB also appear to be surprisingly close to the *EX* and simulations results (*MM* and *EX* imply the use of $P_{PS}^{MM}(j)$ and $P_{PS}^{EX}(j)$ respectively). In figure 7.8, we "zoom" into a portion of figure 7.7 for greater resolution in matching the curves, and plot various *MM* results for $\theta \leq 26\text{dB}$, in steps of 0.5dB. At this resolution, it appears that θ ≈25dB gives the closest match between *MM* and *EX*.

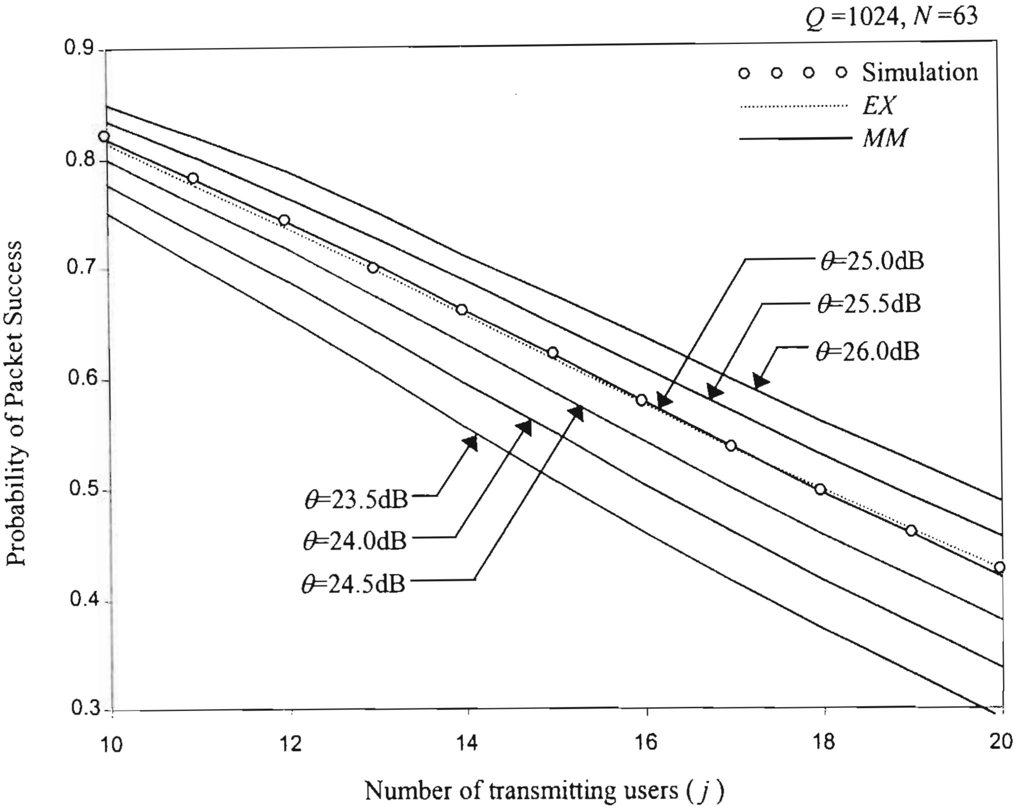


Figure 7.8 : Probability of packet success vs. the number of transmitting users for various θ (fine resolution).

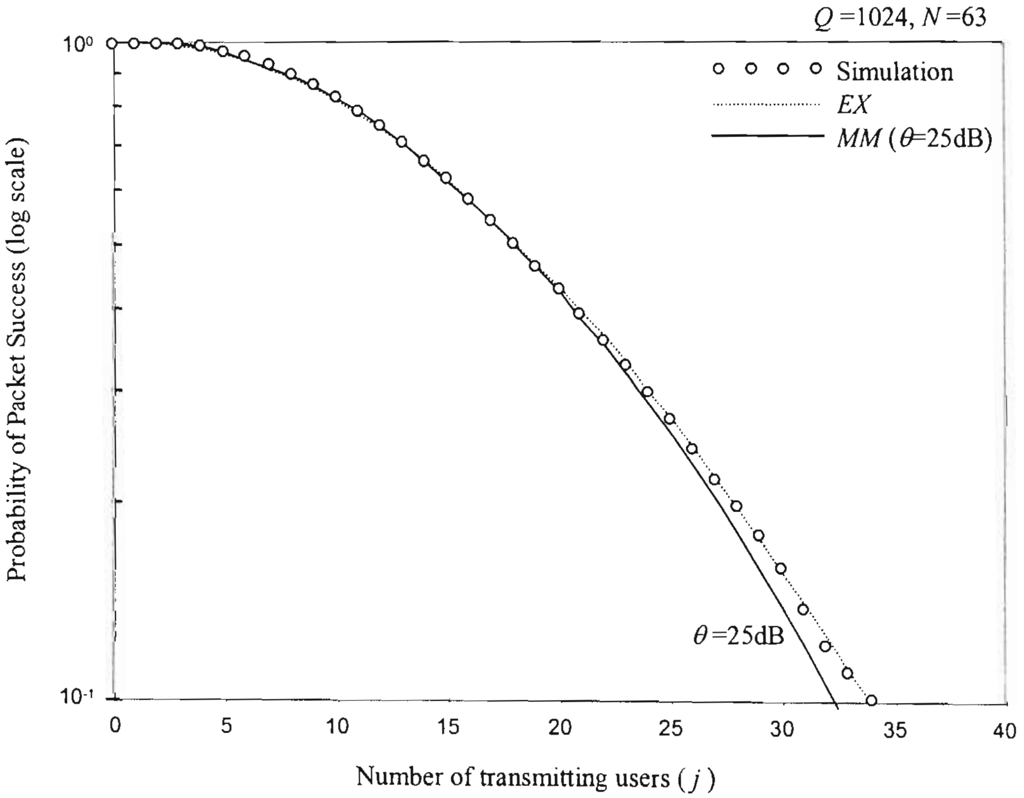


Figure 7.9 : Packet success probabilities for the *MM* approach for the value of θ which provides the closest match to the *EX* and simulation results.

In figure 7.9, we plot a more complete curve to confirm this match. The high degree of correlation between *MM* and *EX* results for $\theta \approx 25$dB is astonishing when one considers the gross "digitisation" of the SIR and resultant error process that was made in our discrete two-state Markov fading model. As can be seen, the *MM* response appears to be more accurate for smaller j and underestimates for larger j. This, of course, is for $\theta \approx 25$dB. For a slightly larger value such as $\theta \approx 25.5$dB, we might expect to see a match for larger j, with the *MM* slightly overestimating for smaller j. Since our SR-DT/CDMA protocol attempts to avoid the number of users exceeding the multiple access threshold $\beta = 20$, we see that $\theta \approx 25$dB provides a satisfactory match for the cases when $j < 20$.

In figure 7.10a and 7.10b, we investigate the accuracy of the HMM for different packet lengths since we expect that the value of θ should be dependent on the number of bits per packet. In equation 7-29, we see that the probability of packet success is very much dependent on the number of bits per packet, as well as the expected BER. For a given BER (related directly to θ through the SIR), we would obviously expect a shorter packet to have a higher probability of success than a longer packet. We thus expect to see that θ should increase as the size of the packet decreases, and should decrease as the packet size increases. Figure 7.10a considers a packet length of $Q = 10000$ bits, while figure 7.10b considers a packet length of $Q = 1$ bit.

For the long packet case ($Q = 10000$ bits), a quantitative trial-and-error approach revealed that a value of $\theta \approx 22$dB appeared to provide a good match over the operating region of interest. By the region of interest, we mean that we expect the system to be operating at the lower values of j far more often than for larger values of j. For example, we would hope the network would very seldom be operating in conditions where the expected mean packet success rate is 10^{-1} (for $j > 20$ users in figure 7.10a). Again, we see that our *MM* model tends to be accurate for low j (the operating region of interest), and underestimates the probability of success for larger j.

The case $Q = 1$ bit equivalently means that we are attempting to model the bit error (or success) rate directly using the HMM. The trial-and-error approach revealed that a significantly larger value of $\theta \approx 44$dB provided the best match for $Q = 1$. Again, the

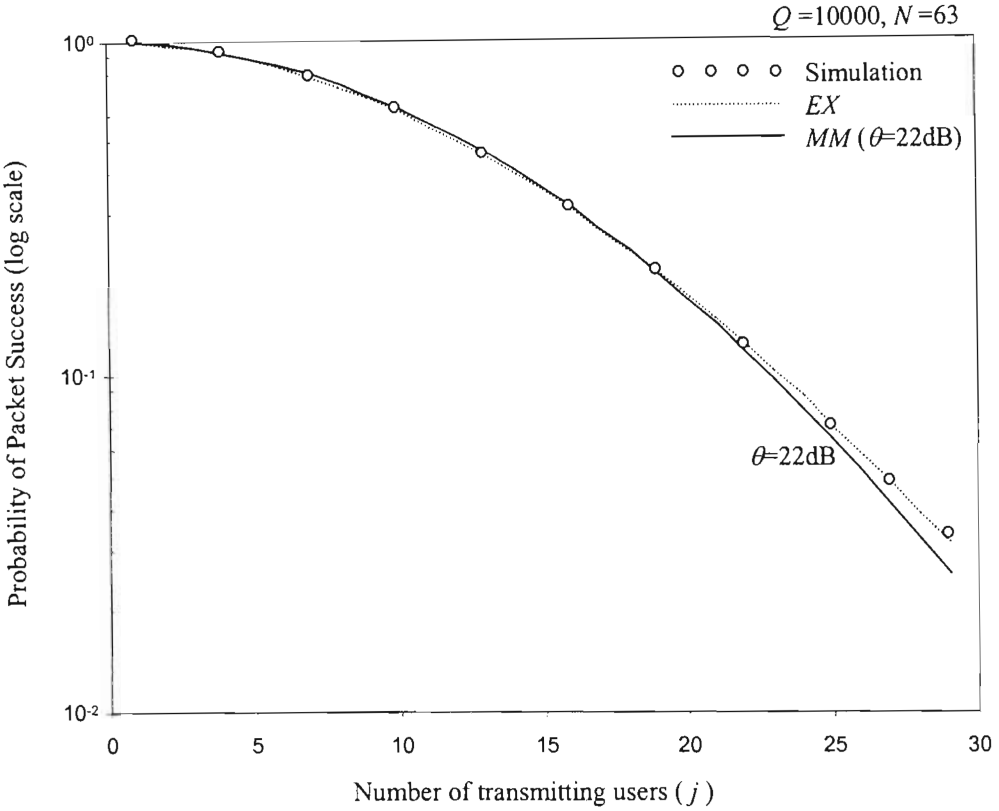


Figure 7.10a : Packet success probabilities from the *MM* for Q=10000 bits. $\theta \approx$ 22dB provides a satisfactory match to the exact (*EX*) and simulation results.

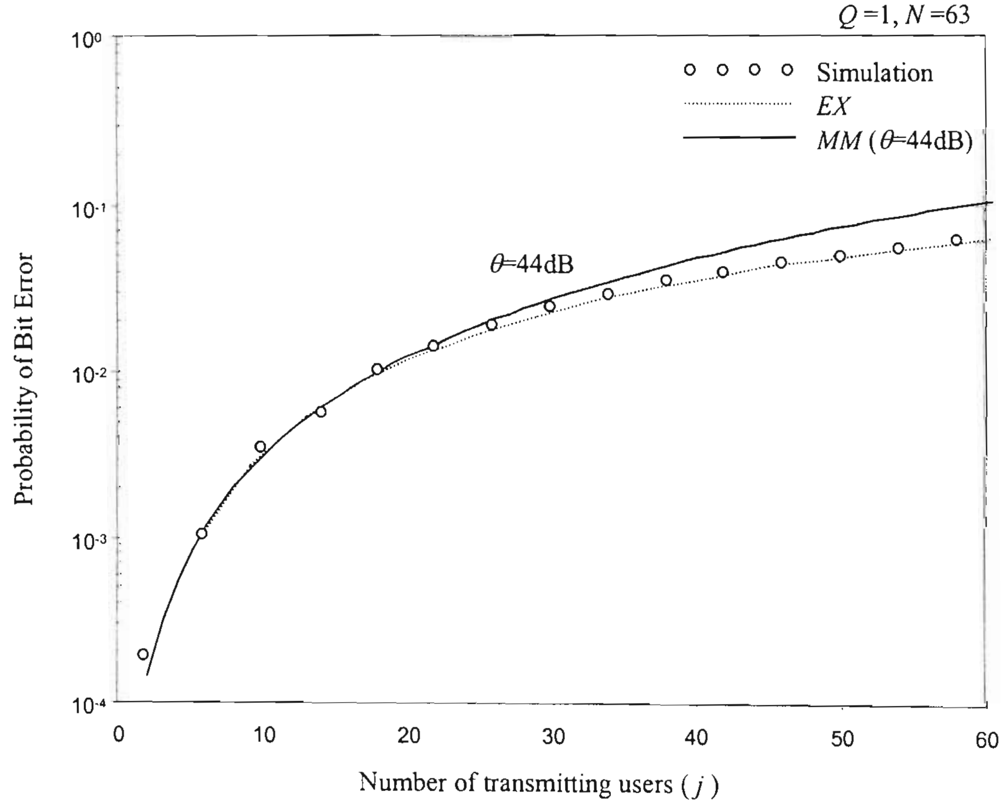


Figure 7.10b : Bit error probability (Q=1) for the *MM* (θ=44dB)

model is seen to underestimate the probability of success for larger *j*, but again however, we hope the network will seldom be operating at a BER less than 10^{-2}!

7.5.3 Effect of correlated Rayleigh fading on the data network throughput

In this section, we evaluate the accuracy of our HMM, as well as the effect that the Rayleigh fading has on the throughput performance of our data protocol. In figure 7.11, we plot simulation and analysis results of the data packet throughput for the Rayleigh fading channel (i.e. with Rayleigh parameter b=1) for several values of $f_D t_s$, as well as the perfect channel (as obtained in Chapter 4).

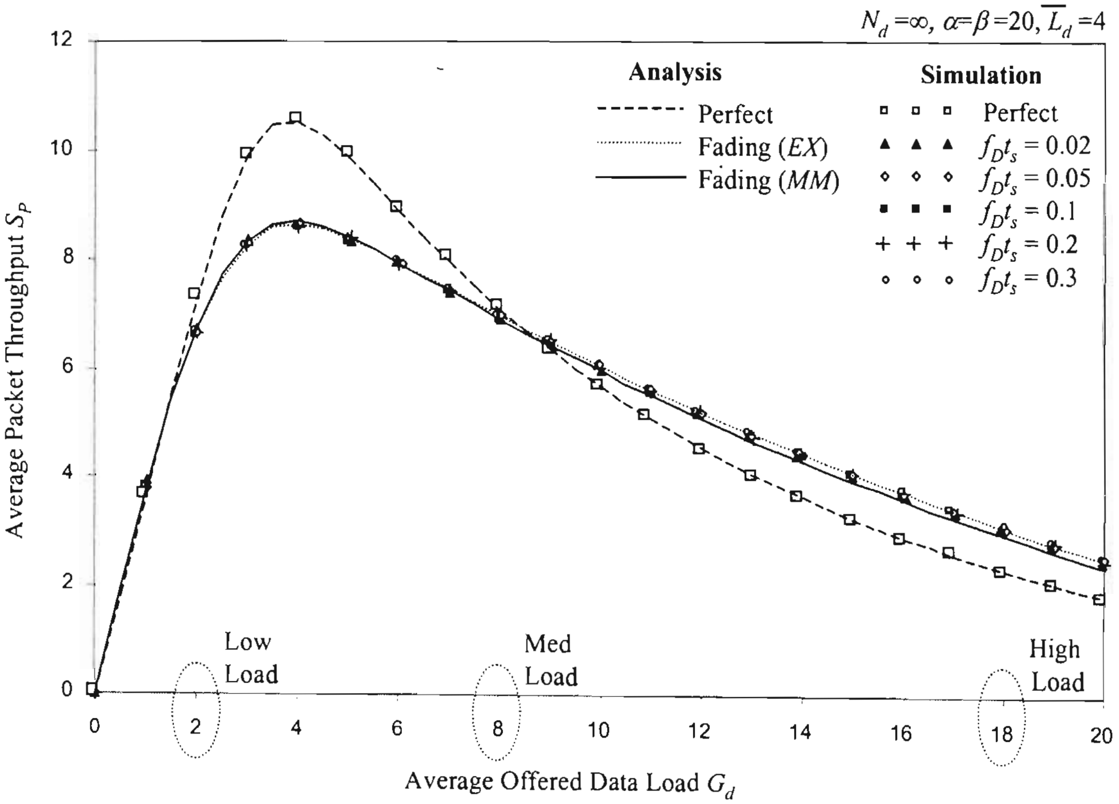


Figure 7.11 : Data packet throughput for the perfect and Rayleigh fading channels (collision detection only, no blocking : $\alpha=\beta=20$).

Intuitively, we expect that the degree of correlation in the channel should have absolutely no effect on the packet throughput of the system. This is because the success of a data packet is dependent entirely on the number of transmitting users in the current slot, and completely independent on the number of transmitting users (or their associated powers) in the previous slot(s). This is reflected in our estimation for the packet throughput (equation 7-35), where no reference is made to any conditional channel state probabilities (such as the transition probability C_{ij}). The simulation results, which overlap exactly for all values of $f_D t_s$ in figure 7.11, prove this intuition to be true. As can be seen our *MM* model is very close to the *EX* and simulation results. As reflected in figure 7.9, the *MM* model underestimates the throughput slightly for larger loads where the number of transmitting users, and the frequency of collisions, is expected to be greater. For later use, we also define three different load points on figure 7.11, namely: a low load scenario, a medium load scenario, and a high load scenario.

A very interesting result from figure 7.11 is that the fading channel actually provides a better system throughput than the perfect channel at higher loads for these particular network parameters (α=β). What we are seeing here is a classic case of channel "capture". In narrowband wireless ALOHA networks where all users are received with equal power, then simultaneous packet transmissions are always unsuccessful. If one user's signal is significantly stronger than the other simultaneous signals, then that user may actually be received correctly. As discussed in section 2.4.6, this signal capture phenomenon has been shown to improve the performance of wireless random access networks. In a CDMA channel, the degree of success varies as a function of the SIR. There may be several users that are more successful than others due to the fact that they are received with a higher power. In these cases the average packet success, when all users are considered, is actually larger than if all users had the same power and probability of success. This accounts for the improved throughput during high MAI for the fading channel.

In figure 7.12 we plot results for an example case when the blocking threshold is employed (α=12, β=20). As can be seen, the advantage of channel capture is now overwhelmed by the advantages of employing the blocking threshold. This is to be expected since the capture phenomenon usually only occurs in very poor channel conditions. Since we can consider a large number of transmitting users to be a poor channel condition (high MAI), the α threshold negates signal capture by avoiding having a large number of users transmitting, i.e. message access is denied before the

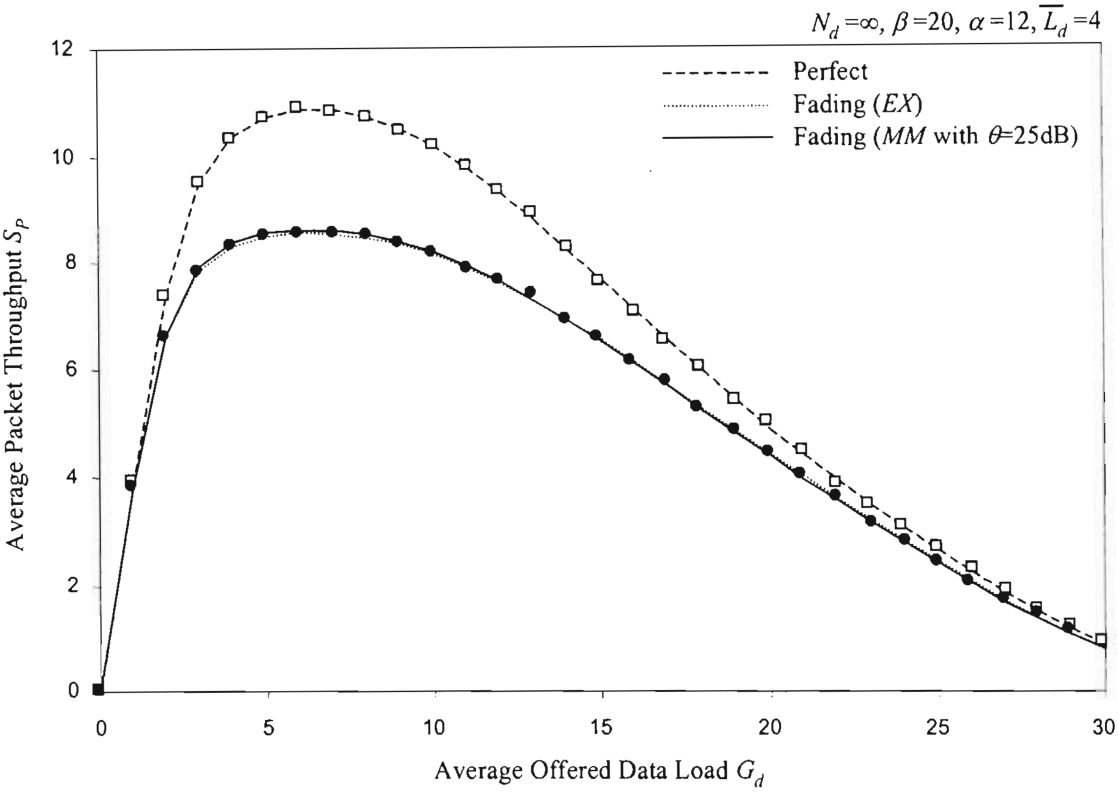


Figure 7.12 : Data packet throughput for the perfect and Rayleigh fading channels (collision detection, with blocking : α=12, β=20).

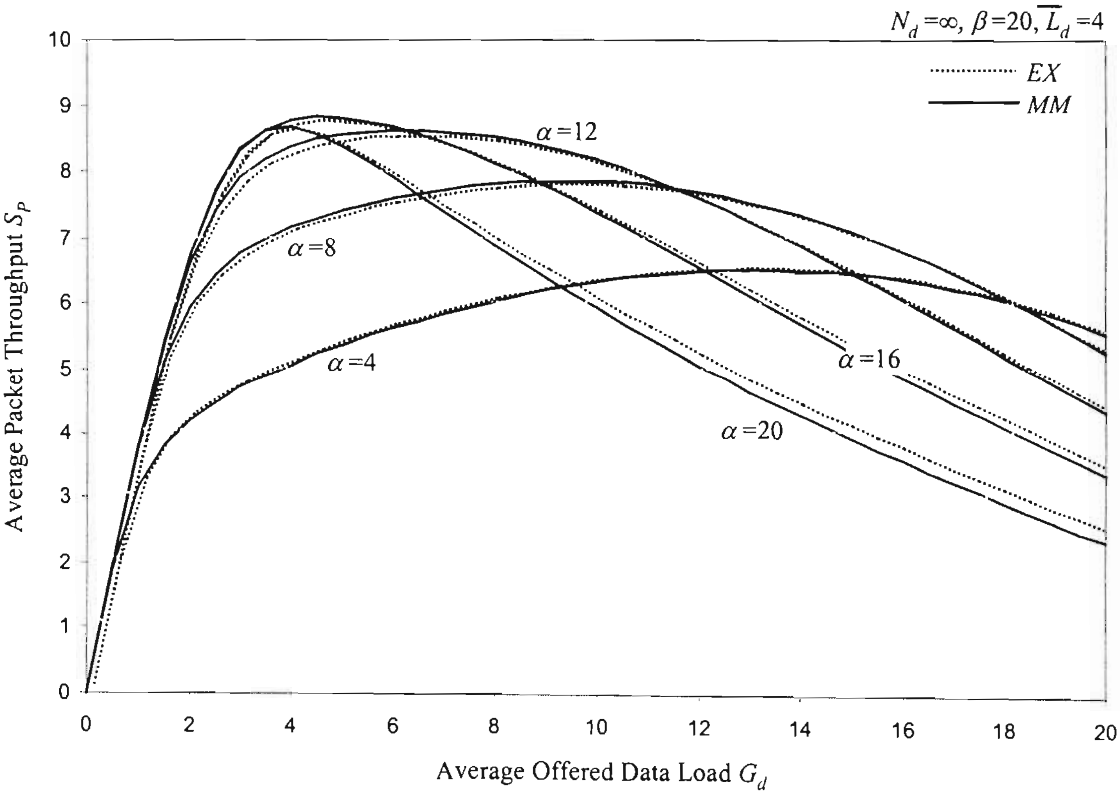


Figure 7.13 : Optimum data packet throughput for the Rayleigh fading channel obtained by using a variable α threshold

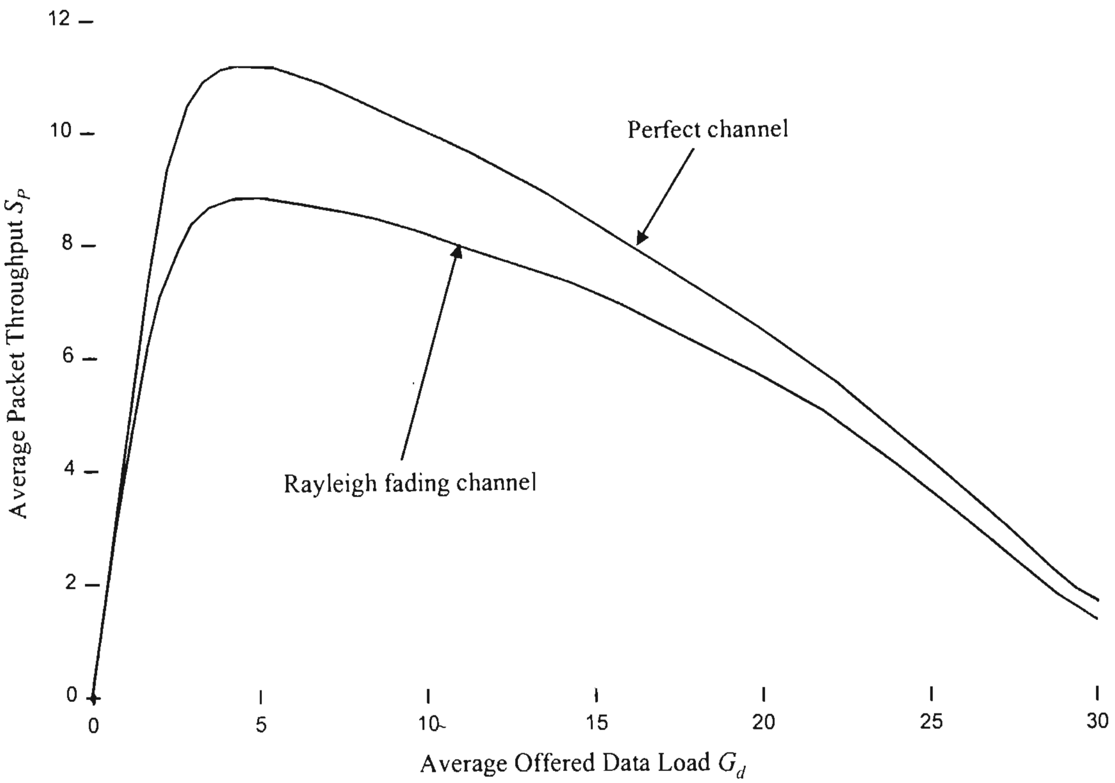


Figure 7.14 : Comparison of the optimum data packet throughputs for the perfect and Rayleigh fading channels

effects of capture can occur. This is not a disadvantage though, since the throughput enhancement gained from α appears to far exceed the performance enhancement gained from capture.

In figure 7.13, we evaluate our *MM* model, as well as the protocol performance for a larger range of α values. As can be seen, our *MM* model is accurate under all load and threshold conditions. The advantages of implementing a variable α threshold are also seen to apply for a fading channel, as they did for the perfect channel in Chapter 4.

If we compare the optimum throughput curve for the fading channel (imagine the envelope of the curves in figure 7.13), to the optimum throughput curve for the perfect channel (figure 4.9), then we see that the perfect channel is superior to the fading channel we have considered here. This comparison is illustrated in figure 7.14.

7.5.4 Evaluation of the HMM for a correlated channel

In order to evaluate the degree to which our Markov model in figure 7.6 is successful in modelling a correlated fading channel, we require some means to test the transition parameters of the model, in particular the expression for conditional probability of remaining in the good state over two consecutive slots (defined as C_{ij} and computed using equation 7-21). One way to do this is to use the recursive procedure in equations 7-30 to 7-32 which effectively computes the success of n consecutive packets in the channel and makes use of the conditional transition probability C_{ij}.

In theory, the message throughput should also be independent of the degree of channel fading. This was proved in the finite population model of Chapter 4, where we saw that the message throughput was always equal to the packet throughput divided by the expected number of packets per message (at network equilibrium). So, an advantage of our protocol is that it is completely unaffected by the degree of correlation of the channel fading between consecutive packets. However, equation 7-34 still represents the probability of having L correct packets in sequence in the channel, regardless of the fact that it may be a subset of a previously transmitted message, and is thus a useful means to test the accuracy of the state transition probabilities of the HMM.

In figure 7.15, we plot distributions for the probability of success of a message conditioned on its length for different load values and different degrees of channel correlation, as obtained from the simulation model and the HMM (i.e. $P_{MS}(l)$ from equation 7-34). Figure 7.15a plots $P_{MS}(l)$ vs. l for the low load scenario (as we defined earlier on figure 7.11) for different values of $f_D t_s$. Figure 7.15b and 7.15c do the same for the medium and high load scenarios. For the low load case, we see that the perfect channel is more efficient than the fading channel. This corresponds to the results in figure 7.11. For the high load case, the reverse is seen to be true due to capture, also as seen in figure 7.11. For the medium load case, we omit the perfect channel case since the results were seen to be almost exactly the same as, and clashed with, the plot for $f_D t_s = 0.3$. From all three plots, our model is seen to be fairly accurate. Interestingly,

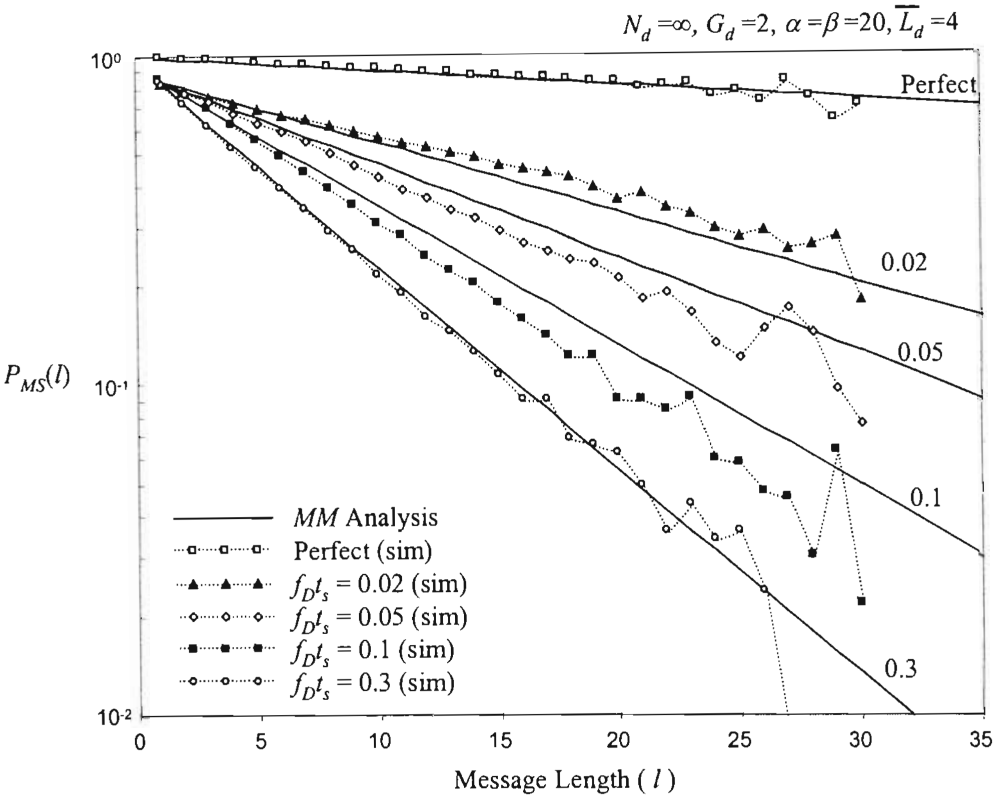


Figure 7.15a : Message success probabilities for the low offered load scenario

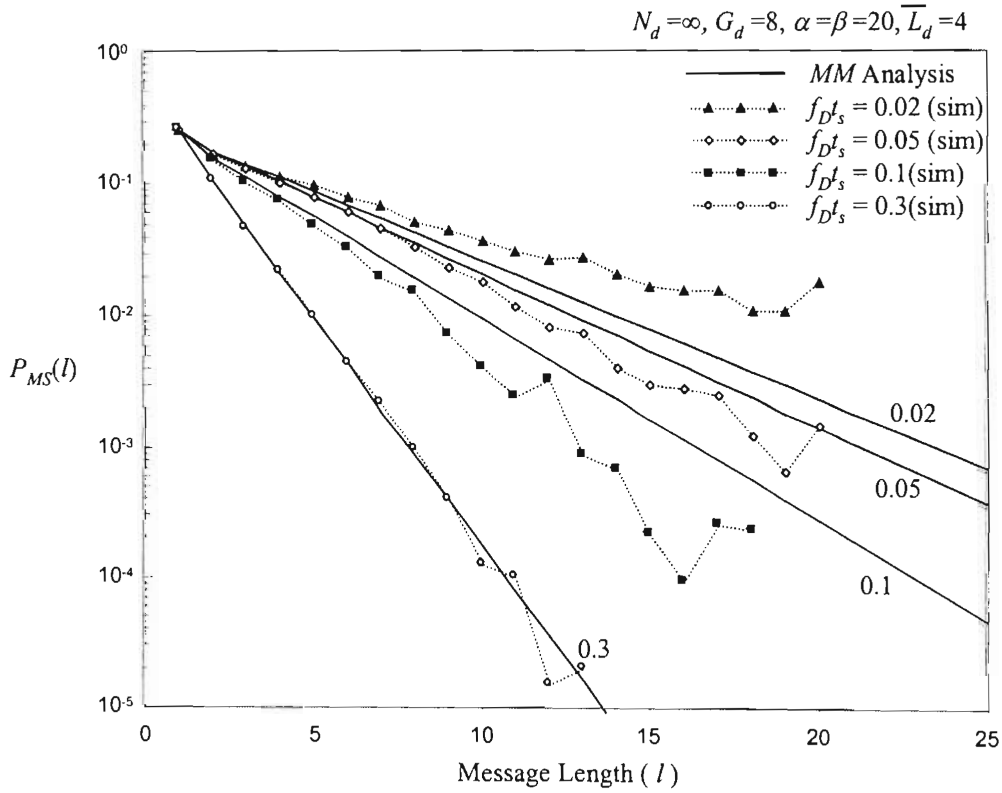


Figure 7.15b : Message success probabilities for the medium offered load scenario

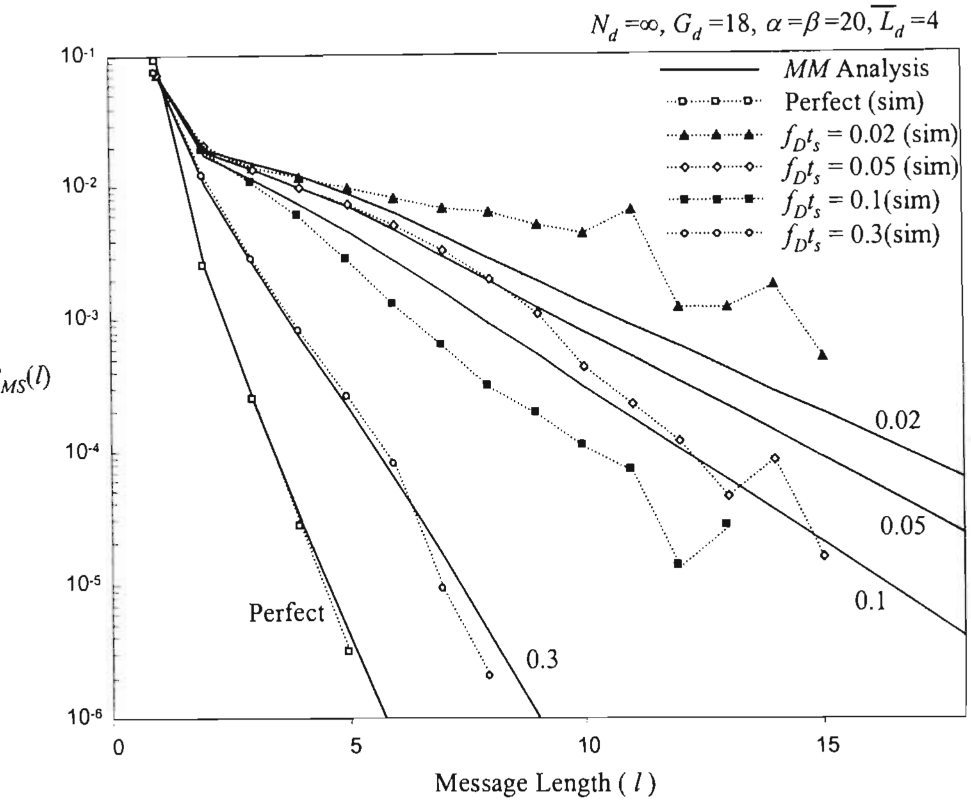


Figure 7.15c : Message success probabilities for the high offered load scenario

we see from the simulation that the probability of message success is practically a geometric function of its length. This is seen by the fact that the distribution, when plotted on a logarithmic scale, is a straight line. Our HMM is quite successful in capturing this phenomenon. From all three load plots, we see that the HMM starts by underestimating the probability of message success for very slow fading channels ($f_D t_s < 0.03$), tends towards overestimating the probability of message success for intermediate fading channels ($0.03 < f_D t_s < 0.2$), but is practically perfect for highly uncorrelated channels ($f_D t_s > 0.2$). For values of $f_D t_s > 0.2$, the results do not change much and the channel can be assumed to be completely uncorrelated.

With respect to l, we see that the accuracy of our model decreases with increasing l. Fortunately, this is not so serious since the probability of having a message of length l diminishes with increasing l, due to the fact that the distribution for message lengths is geometric. This can be seen in the erratic simulation results for large l, where the occurrence of longer messages is too rare (even over a considerable simulation time) for the simulation to converge to a decent average. The most probable reason for

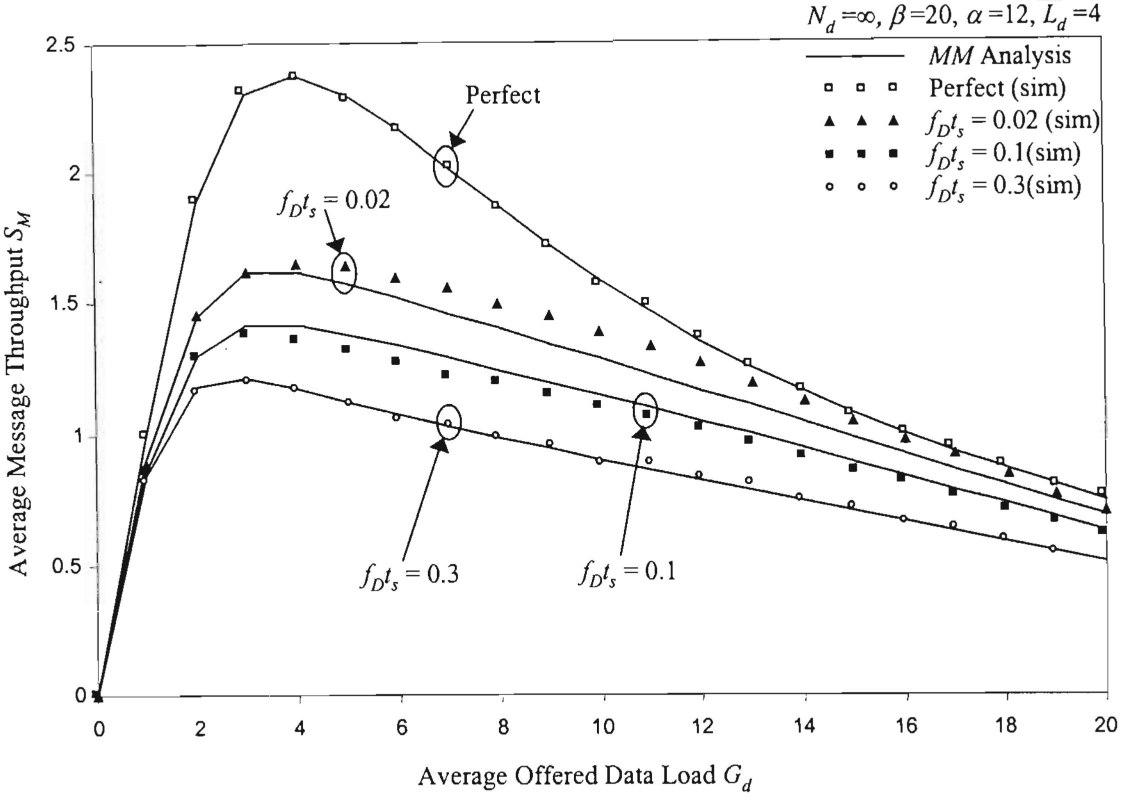


Figure 7.16 : Message throughput for the correlated Rayleigh fading channel for several different values of f_Dt_s.

this inaccuracy in the HMM is due to the significant simplifying assumption we made in section 7.3.2 that the level of MAI in the channel is completely uncorrelated. However, the complexity avoided in parameterising the HMM outweighs this loss of accuracy, and, despite its simplicity, the HMM appears to be modestly accurate.

For all results in figure 7.15, we see that a highly correlated channel provides a message with a better chance of being successful. This is so because the probability of a message remaining in the non-fade state over all packets is higher than for a more uncorrelated channel. One may argue that it also means that the probability of remaining in the non-fade is also higher. However, it only takes one unsuccessful packet to render a sequence of packets "unsuccessful" and, in a more uncorrelated channel where the transition rates between states are larger, we expect a message to have a higher probability of experiencing a fade state during its transmission.

In figure 7.16, we evaluate the accuracy of the HMM in terms of the message throughput for the infinite population model. We choose the infinite population model because the message throughput will, although unrealistically, be dependent on the degree of channel correlation. Results yield exactly what we would expect from the results of figure 7.15. For highly correlated channels, the HMM underestimates the probability of message success and hence the message throughput. For the intermediate fading channel, the HMM overestimates slightly. For a highly uncorrelated channel, the HMM is most accurate.

7.6 Summary

In this chapter, we have modified the simple narrowband two-state Gilbert-Elliot Markov model to aid in modelling the error process in a multi-user CDMA channel in which each user's signal is affected by correlated Rayleigh fading. Our proposed HMM differs from the narrowband Gilbert-Elliot model in that the level of interference in the channel is a time-varying value due to the changing MAI, rather than a fixed value. The HMM is based on representing the state of the channel as seen by each user as a discrete signal waveform containing only two states, namely Bad (fade) and Good (non-fade), which are determined by considering whether the user's SIR is below or above some pre-defined value.

Using a well-known first-order Gaussian model of a correlated Rayleigh channel, together with several simplifying assumptions, all steady-state and transient parameters of the HMM were computed. The simplifying assumptions made were that the MAI due to several Rayleigh faded users was Gaussian in nature, and that the sequence of MAI samples was uncorrelated over time. The error mapping scheme employed was one in which a packet error was assumed to occur in the Bad channel state with probability one, and in the Good channel state with zero probability.

The proposed HMM, despite its simplicity, was shown to provide a surprisingly accurate approximation of the overall probability of packet and message successes, when compared to simulation results and a near exact computation of the packet success probability. In particular, the two-state HMM, in conjunction with the error mapping scheme, was found to provide a very satisfactory approximation for the probability of success of a packet, especially when the number of users in the channel is low. The model was found to be less successful in modelling the correlated fading aspect of the channel, most probably due to the fact that our assumption that the MAI in the channel is uncorrelated is not entirely valid. However, the significant mathematical and analytical complexity avoided by making this assumption far outweighed the slight loss of accuracy of the model.

The HMM, in conjunction with simulation results, was then used to evaluate the effect that an imperfect Rayleigh fading channel has on the performance of our proposed data network's performance. As was expected, we saw that a perfect channel provides a significant performance advantage over a fading channel. In particular, it was shown that the probability of having several contiguously successful packets in a channel increases with an increase in the correlation coefficient of the fading process.

The contents of this chapter, in part, have been submitted for review as a journal paper to "Wireless Networks" (ACM/Baltzer Science Publishers) [Judge & Takawira, 1999c].

CHAPTER 8

CONCLUSIONS

8.1 Summary and Conclusions

This thesis examined the topic of medium access control protocols for voice/data integration in a CDMA packet radio network. A detailed mathematical and statistical performance analysis of a CDMA MAC protocol was presented, beginning with a data-only system and extending it into an integrated voice/data system, and considering both a perfect channel case and a correlated Rayleigh fading channel case. Simulation results for the proposed MAC protocol were also provided.

A literature survey of traditional and modern MAC protocols was presented in Chapters 2 and 5 of this thesis. Chapter 2 considered traditional MAC protocols that have been proposed for single traffic type networks (data-only or voice-only networks), while Chapter 5 considered more recent MAC protocols proposed for multimedia traffic networks.

Based on a random access MAC protocol proposed in a previous study [Resheff & Rubin, 1990] and an observation made in [Toshimitsu *et al*, 1994], a new MAC protocol for the reverse link of a centralised CDMA data-only packet radio network (wireless LAN or cell) was derived, which we name "Dual-Threshold CDMA" (DT/CDMA). We detailed the proposed network architecture and proposed data-only MAC policy for DT/CDMA in Chapter 3 of this thesis. For the data traffic model used for the study of DT/CDMA, we assumed an infinite population with a Poisson message arrival process and geometrically distributed random data message lengths. A Markov analysis of the proposed MAC protocol was derived and presented, the outputs of which provide predicted measurements such as the probability distribution for the number of transmitting data users, the expected data message blocking probability, the expected data message and packet throughputs, and the expected message delay.

Analytical performance measurements were then obtained for a typical imaginary data network scenario with a 64kbps data rate, a 4.096Mcps CDMA chip rate, a mean data packet length of 1024 bits and a mean data message length of 4 packets (4096 bits). Initially, we also considered a perfect, non-fading channel in which all users are received with equal power at the base station. For comparison, results from a custom-built software simulator for the network were also provided. Analytical and simulation results match very closely, thus validating our Markov analysis and analytical approach.

For our original DT/CDMA MAC protocol, the results showed firstly that, by implementing a channel overload (collision) detection scheme that aborts all data messages when the number of transmitting users exceeds some pre-specified user overload threshold, β, a significant throughput performance improvement can be obtained over the equivalent slotted-SS/ALOHA MAC protocols that do not employ any access restrictions. Through experimentation, we showed that the value of the β threshold should be chosen carefully to reflect the *optimum overall system throughput*. We then showed that further performance improvements can be made by implementing a second message blocking threshold, α, where $\alpha \leq \beta$, the purpose of which is to reduce the frequency of channel overloads by denying any further channel access whenever the number of transmitting users exceeds α. We showed that the α threshold should be variable and dependent on the expected data message arrival rate. For low offered loads, when the frequency of collisions is low, the value of α should be large (minimal blocking) whereas for higher loads, where the frequency of collisions is high, the value of α should be lower (increased blocking). This novel implementation of a fixed overload detection threshold β in conjunction with a load dependent blocking threshold α, is shown to be far superior to the existing SS/ALOHA protocols which employ neither blocking nor collision detection, or either blocking only or collision detection only.

We also discussed the fact, however, that our proposed DT/CDMA MAC policy is highly susceptible to instability, due to the nature of the chosen message retransmission scheme. Our initial DT/CDMA protocol specifies that if a data message is corrupted by MAI, then the entire message is retransmitted, regardless of the number of corrupt packets in the message. This assumption has several implications, the most significant being that it makes the Markov analysis of the protocol extremely difficult for a finite population model with a population dependent, non-Poisson, message arrival process.

In Chapter 4, we presented an improved MAC policy that considers a more efficient selective-repeat retransmission scheme in which only those packets in a message that are corrupt are retransmitted. This improvement on our original DT/CDMA MAC protocol results in what we name "Selective-Repeat Dual-Threshold CDMA" (SR-DT/CDMA). The improved selective-repeat retransmission scheme allowed us to make some simplifying assumptions regarding the distribution of the length of retransmitted messages in the network, namely that it is approximately geometric in nature. This assumption hence made it possible to perform a combined Markov and equilibrium-point analysis for a finite population model with a population dependent, non-Poisson message arrival process. An approximate statistical analysis based on a flow equilibrium model was presented for the SR-DT/CDMA protocol with a finite population traffic model. For direct comparison with the original DT/CDMA policy, we also provided a Markov analysis of SR-DT/CDMA for an infinite population model with a Poisson arrival process. Results showed that the improved selective-repeat retransmission scheme provides a superior performance over the DT/CDMA scheme under all load conditions. For the finite population model, simulation and analytical results from the combined Markov and equilibrium-point analysis were seen to compare satisfactorily, despite the simplifying assumptions that were made in order to make solving the model easier.

Having laid the groundwork for the data-only MAC protocol and its associated Markov analysis, we then extended the model to a joint voice/data scheme. This joint traffic model and it's associated analysis and results were presented in Chapter 6. We presented our imagined network architecture, traffic models and MAC admission policies for both traffic types. The proposed scheme is based on a two-service threshold model for voice and data traffic, as proposed in [Soroushnejad & Geraniotis, 1995]. The graceful degradation property of CDMA is used to accommodate several simultaneous spread-spectrum voice users up to some maximum multiple access threshold K, after which blocking is applied. The MAC protocol for voice users is similar to that of SR-

DT/CDMA for the data-only network, the only difference being that there is no α threshold, and corrupt voice messages are not retransmitted. Data users contend for any remaining channel capacity using the SR-DT/CDMA data MAC policy proposed in Chapter 4. In this thesis, we assumed that all voice traffic has unconditional priority over data traffic with regards to channel access.

For the voice traffic model, we assumed a finite population of terminals and assumed that each terminal can exist in one of three states: idle, silent and talking. This model considers the call idle and call holding times to be approximately geometrically distributed, and also models the voice activity factor of human speech. In this thesis, we considered a scheme in which a voice terminal's transmission is severely suppressed during the silence periods, and hence only contributes MAI to other uses during the periods in which the speaker is talking. A Markov analysis was derived for the joint voice/data protocol. The outputs of this analysis predict the state distributions and joint state distributions for all pertinent states of the network. For the voice sub-network, the expected carried number of voice calls, expected number of talking voice users, the expected offered call arrival rate and the expected voice call blocking probability were computed. For the data sub-network, the throughput measurements in the presence of voice traffic were computed.

Analytical and simulation performance measurements were then obtained for an imaginary network scenario using the following network parameters: 64kbps voice and data rate, a 4.096Mcps CDMA chip rate, a mean data packet length of 1024 bits, a mean data message length of 4 packets (4096 bits), a mean voice call length of 3 minutes and a perfect, non-fading channel. Several voice activity factors were considered in order to gauge this effect on the performance of the network. Again, analytical and simulation results matched very closely. For the voice network, we were able to find the maximum voice call arrival rate that causes blocking in excess of the widely excepted maximum (between 1 percent and 2 percent for cellular networks), and hence dimension for the expected operating level of voice traffic. For the data network, the expected effect of adding voice traffic results in a decrease in the data performance. This reduction is a result of the fact that the voice users decrease the channel capacity available to data, as well as contribute MAI to the channel and hence increase the probability of data packet error. In exploiting and varying the voice activity factor, we can see that a significant improvement in the data performance can be obtained, especially for low voice activity factors. If the standard voice activity factor of 30 percent is considered, then the reduction in performance of the data network in the presence of voice traffic is seen to be severely reduced, even under high voice load conditions when the number of calls in the network is high. This result stresses and confirms the significant capacity and performance improvements that can be obtained by suppressing a voice terminal's transmission during the silent periods inherent in, and which make up 70 percent of, a typical human conversation.

Lastly, we relaxed the assumption that the channel is perfect, and instead considered a correlated multi-path fading channel in which each user's received signal is assumed to be Rayleigh distributed. To aid in the performance analysis of the network operating in such a channel, we developed a simple hidden Markov model (HHM) based on the traditional Gilbert-Elliot Markov channel model. The model is essentially a discrete-time joint-state Markov chain which considers a channel to be either "good" (non-faded) or "bad" (faded). Our model differs from the original Gilbert-Elliot model in that firstly, it considers a variable MAI level that is related to the number of CDMA interferers in the channel, rather than a fixed interference level as assumed by Gilbert. Secondly, the steady state occupancy probabilities for the channel are dependent on the number of users in the channel due to the fact that the number of users determines the MAI. Based on a suitable signal model for a correlated Rayleigh fading channel, the state occupancy and state transition parameters of the HMM were computed. We also proposed a simple scheme to map the channel state sequence onto the sequence of errors that occur in the channel for a reference user. The results obtained from the proposed HMM, when compared to simulation results and a more accurate expression for the packet error rate, were found to be surprisingly accurate, despite the simplicity of the model. With regards to the performance of the network in the presence of correlated Rayleigh fading, the following conclusions were drawn: Firstly, the perfect channel scenario was shown to be more efficient than the Rayleigh fading channel, as expected. Secondly, it was shown that the degree of correlation in the fading channel has an influence on the message success probability, where slower fading channels produce a higher probability of message success than faster fading channels.

8.2 Possible Extensions and Topics of Further Study

Finally, to conclude this dissertation, we present some possible extensions and/or further topics of study to this research:

1. In chapters 3 and 4 we obtained the multiple access ("collision") threshold β through experimentation, and discussed the fact that a prediction of β is a function of many factors: the packet length in bits, the FEC coding schemes, the message lengths, the retransmission scheme (e.g. selective-repeat yielded a slightly higher threshold), the MAC protocol etc... As a possible extension, we propose the derivation of an expression to predict β based on a given set of network parameters.

2. Similarly, we suggest the derivation of an expression to predict the optimum value for the blocking threshold α, based on the given network parameters, the offered load, etc...

3. Furthermore, the possibility of having a soft threshold (i.e. variable β) may also be of interest although, as we point out in section 4.5.2, the advantages gained by varying β are minimal. However, a soft collision threshold may be useful in cases where the maximum allowable bit error rate is changing, as it might do when the traffic type is changing. For example, when only voice users are transmitting, a higher threshold may be used than for the case when both voice and data are transmitting.

4. Providing for other QoS (Quality of Service) requirements other than the maximum acceptable bit (or packet) error rate due to MAI. Consideration might be given to the maximum acceptable delay for a data packet or data message.

5. Extension of the proposed MAC protocol to incorporate other multi-media traffic types such as video. In doing this, one would have to re-evaluate the threshold selection for each traffic type in each slot, such that the maximum packet error rate

in a given slot satisfies the traffic type with the most stringent BER requirements (as discussed in section 5.4.4).

6. Although well understood and well documented in the literature for random access networks, one might be interested to study the stability of the proposed data MAC protocol in more detail. In this thesis, we briefly touched on the topic of network stability and the influence that the message retransmission or backoff scheme has on the stability of the network.

References

Abdelmonem, A H & Saadawi, T N, (1989), "Performance analysis of spread spectrum packet radio networks with channel load sensing", *IEEE Journal on Selected Areas in Communications*, Vol. 7, No. 1, pp. 161-166, January 1989

Abramson, N, (1970), "The ALOHA system - another alternative for computer communications", *AFIPS Conference Proceedings & Fall Joint Computer Conference*, Vol. 37, pp. 281-285, 1970

Azad, H, Aghvami, A H & Chambers, W C, (1999), "Multiservice/Multirate Pure CDMA for Mobile Communications", *IEEE Transactions on Vehicular Technology*, Vol. 48, No. 5, pp. 1404-1413, September 1999

Akyildiz, I F, McNair, J, Martorell, L C, Puigjaner, R & Yesha, Y, (1999), "Medium Access Control Protocols for Multimedia Traffic in Wireless Networks", *IEEE Network Magazine*, pp. 39-47, July/August 1999

Beaulieu, N C, (1990), "An Infinite Series for the Computation of the Complementary Probability Distribution Function of a Sum of Independent Random Variables and Its Application to the Sum of Rayleigh Random Variables", *IEEE Transactions on Communications*, Vol. 38, No. 9, pp. 1463-1474, September 1990

Bellini, S & Borgonovo, F, (1980), "On the throughput of an ALOHA channel with variable length packets", *IEEE Transactions on Communications*, Vol. COM-28, pp. 1130-1143, November 1980

Bischl, H & Lutz, E, (1995), "Packet Error Rate in the Non-Interleaved Rayleigh Channel", *IEEE Transactions on Communications*, Vol. 43, No. 2/3/4, pp. 1375-1382, Feb/Mar/Apr 1995

Brand, A E, & Aghvami, A H, (1998), "Multidimensional PRMA with Prioritized Bayesian Broadcast – A MAC Strategy for Multiservice Traffic over UMTS", *IEEE Transactions on Vehicular Technology*, Vol. 47, No. 4, pp. 1148-1161, November 1998

Bullington, K & Fraser, J M, (1962), "Engineering Aspects of the TASI System", *Bell Systems Technical Journal*, pp. 1439-1454, July 1962

Capetanakis, J I, (1979), "The multiple access broadcast channel: Protocol and capacity considerations", *IEEE Transactions on Information Theory*, Vol. IT-25, pp. 505-515, September 1979

Carleial, A B & Hellman, M E, (1975), "Bistable behaviour of ALOHA-type systems", *IEEE Transactions on Communications*, Vol. COM-23, pp. 401-410, April 1975

Chen, J S J & Li, V O K, (1989), "Reservation CSMA/CD : A multiple access protocol for LAN's", *IEEE Journal on Selected Areas in Communications*, Vol. 7, No. 2, pp. 202-210, February 1989

Cheun, K, (1997), "Performance of direct-sequence spread-spectrum RAKE receivers with random spreading sequences", *IEEE Transactions on Communications*, Vol. 45, No. 9, pp. 1130-1143, September 1997

Chuah, M C, Nanda, S & Rege, K M, (1998), "Analytical Models for a Hybrid Simulation of the CDMA Reverse Link", *Advances in Performance Analysis*, Vol. 1, No. 2, pp. 111-140, 1998

Crisler, K & Needham, M, (1995), "Throughput Analysis of Reservation ALOHA Multiple Access", *Electronic Letters*, Vol. 31, No. 2, pp. 87-89, January 1995

Crowther, W, Rettberg, R, Walden, Ornstein, S & Heart, F, (1973), "A system for broadcast communication: Reservation-ALOHA", *Proceedings of 6th HICSS*, University of Hawaii, Honolulu, January 1973

Elliot, E O, (1963), "Estimates of error rates for codes on burst-noise channels", *Bell Systems Technical Journal*, Vol. 42, pp. 1977-1997, September 1963

Fapojuwo, A O, (1993), "Radio Capacity of Direct Sequence Code Division Multiple Access Mobile Radio Systems", IEE Proceedings, Vol. 140, No. 5, pp. 402-408, October 1993

Fritchman, B D, (1967), "A Binary Channel Characterisation Using Partitioned Markov Chains", *IEEE Transactions on Information Theory*, Vol. IT-13, pp. 221-227, April 1967

Furguson, M J, (1977), "An approximate analysis of delay for fixed and variable length packets in an unslotted ALOHA channel", *IEEE Transactions on Communications*, Vol. COM-25, pp. 644-654, July 1977

Gallager, R G, (1985), "A perspective on multiaccess channels", *IEEE Transactions on Information Theory*, Vol. IT-31, No. 2, pp. 124-142, March 1985

Geraniotis, E, (1986), "Direct-Sequence Spread-Spectrum Multiple-Access Communications Over Non-selective and Frequency Selective Rician Fading Channels", *IEEE Transactions on Communications*, Vol. COM-34, pp. 756-764, August 1986

Geraniotis, E & Pursley, M B, (1982), "Error Probability for Direct-Sequence Spread-Spectrum Multiple Access Communications – Part II: Approximations", *IEEE Transactions on Communications*, Vol. COM-30, pp. 985-995, May 1982

Geraniotis, E & Pursley, M B, (1985), "Performance of Coherent Direct-Sequence Spread-Spectrum Communications Over Specular Multipath Fading Channels", *IEEE Transactions on Communications*, Vol. COM-33, pp. 502-508, June 1985

Geraniotis, E & Pursley, M B, (1986), "Performance of Non-coherent Direct-Sequence Spread-Spectrum Communications Over Specular Multipath Fading Channels", *IEEE Transactions on Communications*, Vol. COM-34, pp. 219-226, March 1986

Geraniotis, E, Soroushnejad, M & Yang, W B, (1995), "A multiple-access scheme for voice/data integration in hybrid satellite/terrestrial packet radio networks", *IEEE Transactions on Communications*, Vol. 43, No. 2/3/4, pp. 1756-1767, Feb/Mar/Apr 1995

Gilbert, E N, (1960), "Capacity of a burst-noise channel", *Bell Systems Technical Journal*, Vol. 39, pp. 1253-1266, September 1960

Gilhousen, K S, Jacobs, I M, Padovani, R, Viterbi, A J, Weaver, L A, Wheatley III, C E, (1991), "On the Capacity of a Cellular CDMA System", *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 2, pp. 303-312, May 1991

Glisic, S & Vucetic, B, (1997), *Spread Spectrum CDMA Systems for Wireless Communications*, Artech House, Boston. London, 1997

Goodman, D J, (1990), "Cellular Packet Communications", *IEEE Transactions on Communications*, Vol. 38, No. 8, pp. 1272-1280, August 1990

Goodman, D J, (1991), "Second Generation Wireless Information Networks", *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 2, pp. 366-374, May 1991

Goodman, D J, Valenzuela, R A, Gayliard, K T & Ramamurthi, B, (1989), "Packet reservation multiple access for local wireless communications", IEEE *Transactions on Communications*, Vol. 37, No. 8, pp. 885-890, August 1989

Goodman, D J & Wei, S X, (1991) "Efficiency of Packet Reservation Multiple Access", *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, pp. 170-176, February 1991

Gruber, J G, (1982), "A comparison of measured and calculated speech temporal parameters relevant to speech activity detection", *IEEE Transactions on Communications*, Vol. COM-30, No. 4, pp. 728-738, April 1982

Habbab, I M I, Kavehrad, M & Sundberg, C E W, (1989), "ALOHA with capture over slow and fast fading channels with coding and diversity", *IEEE Journal on Selected Areas in Communications*, Vol. 7, No.1, pp. 79-88, January 1989

Holtzman, J M, (1992), "A simple accurate method to calculate spread-spectrum multiple access error probabilities", *IEEE Transactions on Communications*, Vol. 40, No. 3, pp. 461-464, March 1992

Jain, R & Routhier, S A, (1986), "Packet Trains - Measurements and a New Model for Computer Network Traffic", *IEEE Journal on Selected Areas in Communications*, Vol. SAC-4, No. 6, pp. 986-995, September 1986

Jakes, W C, (Editor) (1974), *Microwave Mobile Communications*, New York: IEEE Press, 1974

Jamalipour, A, Katayama, M, Yamazato, T & Ogawa, A, (1995), "Performance of an integrated voice/data system in non-uniform traffic low earth-orbit satellite communication systems", *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 2, pp. 465-472, February 1995

Jangi, S & Merakos, F, (1994), "Performance analysis of reservation random access protocols for wireless access networks", *IEEE Transactions on Communications*, Vol. 42, No. 2/3/4, pp. 1223-1234, Feb/Mar/April 1994

Jeon, W S & Jeong, D G, (1998), "An Integrated Services MAC Protocol for Local Wireless Communications", *IEEE Transactions on Vehicular Technology*, Vol. 47, No. 1, pp. 352-363, February, 1998

Jeong, D G & Jeon, W S, (1995), "Performance of an Exponential Backoff Scheme for Slotted-ALOHA Protocol in Local Wireless Environment", *IEEE Transactions on Vehicular Technology*, Vol. 44, No. 3, pp. 470-479, August, 1995

Jenq, Y C, (1980), "On the stability of slotted ALOHA systems", *IEEE Transactions on Communications*, Vol. COM-28, No. 11, pp. 1936-1939, November 1980

Joseph, K & Raychaudhuri, D, (1993), "Throughput of unslotted direct-sequence spread-spectrum multiple-access channels with block FEC coding", *IEEE Transactions on Communications*, Vol. 41, No. 9, pp. 1373-1378, September 1993

Judge, G & Takawira, F, (1996), "A protocol for voice/data integration over a CDMA packet radio network", *Proceedings of TELETRAFFIC'96 Symposium*, Durban S.A., pp. 214-220, September 1996

Judge, G & Takawira, F, (1997), "Performance analysis of a multiple-access protocol for a packet switched CDMA packet radio network using channel load sensing", *Proceedings of IEEE COMSIG'97 Conference*, Grahamstown S.A., pp. 121-126, September 1997

Judge, G & Takawira, F, (1998a), "Performance analysis of a channel load sensing protocol for CDMA packet radio networks with finite population and variable length messages", *Proceedings of IEEE ISSSTA'98 Conference*, Sun City S.A., pp 282-286, September 1998

Judge, G & Takawira, F, (1998b), "Spread-Spectrum CDMA Packet Radio MAC Protocol Using Channel Overload Detection and Blocking", *To appear in Wireless Networks*, ACM/Baltzer Publishers

Judge, G & Takawira, F, (1999a), "Performance Analysis of a Novel Channel Load Sensing Protocol for CDMA Packet Radio Networks Supporting Voice and Data Traffic Integration", *Proceedings of PIMRC'99 Conference*, Osaka, Japan, pp 1295-1299, September 1999

Judge, G & Takawira, F, (1999b), "A Markov Analysis of a Voice/Data MAC Protocol for a CDMA Packet Radio Network", *Submitted to IEEE Transactions on Vehicular Technology*, December 1999

Judge, G & Takawira, F, (1999c), "A Simple Hidden Markov Model for a CDMA Channel with Correlated Rayleigh Fading", *Submitted to Wireless Networks* (ACM/Baltzer Publishers), December 1999

Kanal, L N & Sastry, A R K, (1978), "Models for Channels with Memory and Their Applications to Error Control", *Proceedings of the IEEE*, Vol. 66, No. 7, pp. 724-744, July 1978

Kchao, C & Stüber, G L, (1993), "Analysis of a Direct-Sequence Spread-Spectrum Cellular Radio System", *IEEE Transactions on Communications*, Vol. 41, No. 10, pp. 1507-1516, October 1993

Kleinrock, L, (1975a), *Queuing Systems. Volume I: Theory*, John Wiley and Sons, Wiley-Interscience Publication, 1975

Kleinrock, L, (1975b), *Queuing Systems. Volume II: Computer Applications*, John Wiley and Sons, Wiley-Interscience Publication, 1975

Kleinrock, L & Lam, S S, (1975), "Packet switching in a multi-access broadcasting channel: Performance evaluation", *IEEE Transactions on Communications*, Vol. COM-23, pp. 410-423, April 1975

Kleinrock, L & Tobagi, F A, (1975), "Packet switching in radio channels: Part I: CSMA modes and their throughput-delay characteristics", *IEEE Transactions on Communications*, Vol. COM-23 No. 12, pp 1400-1416, December 1975

Lam, P M & O'Farrell, T, (1992), "Delay/Throughput performance of CDMA embedded in a one-persistent channel sense and collision (overload) detection protocol", *Proceedings of ICCS/ISITA'92 Conference*, Singapore, pp. 491-495, 1992

Lam, S S, (1980), "Packet broadcast networks - A performance analysis of the R-ALOHA protocol", *IEEE Transactions on Communications*, Vol. COM-29, No. 7, pp. 596-603, July 1980

Lee, H H & Un, C K, (1986), "A study of On-Off Characteristics of Conversational Speech", *IEEE Transactions on Communications*, Vol. COM-34, No. 6, pp. 630-637, June 1986

Lee, J H & Un, C K, (1996), "Performance Analysis of Nonpersistent Idle-signal Casting Multiple Access with Collision-Detection (ICMA/CD) Protocol", *IEE Proceedings-Communications*, Vol. 143, No. 3, pp. 162-166, June 1996

Lutz, E, (1992), "Slotted ALOHA Multiple Access and Error Control Coding for Land Mobile Satellite Networks", *International Journal of Satellite Communications*, Vol. 10, pp. 275-281, 1992

Lyghounis, E, Poretti, I & Monti, G, (1974), "Speech Interpolation in Digital Transmission Systems", *IEEE Transactions on Communications*, Vol. COM-22, pp. 1179-1189, September 1974

Mason, W C, Ginsburg, M & Brennan, G D, (1960), "Tables of the Distribution Function of Sums of Rayleigh Variables", M.I.T. Lincoln Lab., Lexington, MA., March 1960

Morrow, R K & Lehnert, J S, (1989), "Bit-to-bit error dependence in slotted DS/SSMA packet systems with random signature sequences", *IEEE Transactions on Communications*, Vol. 37, No. 10, pp. 1052-1061, October 1989

Morrow, R K & Lehnert, J S, (1992), "Packet throughput in slotted ALOHA DS/SSMA radio systems with random signature sequences", *IEEE Transactions on Communications*, Vol. 40, No. 7, pp. 1223-1230, July 1992

Nanda, S, Goodman, D J & Timor, U, (1991), "Performance of PRMA: A Packet Voice Protocol for Cellular Systems", *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 3, pp. 584-598, August 1991

Polydoros, A & Silvester, J A, (1987), "Slotted Random Access Spread Spectrum Networks - An Analytical Framework", *IEEE Journal on Selected Areas in Communications, A Special Issue on Performance Evaluation of Multiple-Access Networks*, Vol. SAC-5, No. 5, pp. 989-1002, July 1987

Prasad, R, (1991), "Performance analysis of mobile packet radio networks in real channels with inhibit-sense multiple access", *IEE Proceedings-I*, Vol. 138, No. 5, pp. 458-464, October 1991

Prasad, R, (1996), *CDMA for Wireless Personal Communications*, Artech House, Boston. London, 1996

Prasad, R, Nijhof, J A M & Çakil, H I, (1997), "Performance Analysis of the Hybrid TDMA/CDMA Protocol for Mobile Multi-Media Communications", *Proceedings of the ICC'97 Conference*, pp. 1063-1067.

Press, W H, Teukolsky, S A, Vetterling, W T & Flannery, B P, (1992), *Numerical recipes in C: The art of scientific computing*, 2nd Edition, Cambridge University Press, 1992

Proakis, J G, (1989), *Digital Communications*, New York: McGraw-Hill, 1989

Pursley, M B, (1977), "Performance evaluation for phase-coded spread-spectrum multiple-access communication - Part I : System Analysis", *IEEE Transactions on Communications*, Vol. COM-25, No. 8, pp. 795-799, August 1977

Pursley, M B, Sarwate, D V & Stark, W E, (1982), "Error Probability for Direct-Sequence Spread-Spectrum Multiple Access Communications – Part II: Approximations", *IEEE Transactions on Communications*, Vol. COM-30, pp. 985-995, May 1982

Pursley, M B & Taipale, D J, (1987), "Error probabilities for spread spectrum packet radio with convolutional codes and Viterbi decoding", *IEEE Transactions on Communications*, Vol. COM-35, No. 1, pp. 1-12, January 1987

Raychaudhuri, D, (1981), "Performance analysis of random access packet-switched Code Division Multiple Access systems", *IEEE Transactions on Communications*, Vol. COM-29, No. 6, pp. 895-901, June 1981

Raychaudhuri, D, (1984), "ALOHA with Multipacket Messages and ARQ-Type Retransmission Protocols – Throughput Analysis", *IEEE Transactions on Communications*, Vol. COM-32, No. 2, pp. 148-154, February 1984

Raychaudhuri, D, (1987), "Stability, throughput, and delay of asynchronous selective reject ALOHA", *IEEE Transactions on Communications*, Vol. COM-35, No. 7, pp. 767-772, July 1987

Raychaudhuri, D, French, L J, Siracusa, R J, Biswas, S K, Yuan, R, Narasimhan, P & Johnston, C A, (1997), "WATMnet: A Prototype Wireless ATM System for Multimedia Personal Communication", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No.1, pp. 83-95, January 1997

Resheff, S & Rubin, I, (1990), "Performance of a coded band-limited spread-spectrum multiple-access scheme using channel load sensing", *IEEE Journal on Selected Areas in Communications*, Vol. 8, No. 4, pp. 515-528, May 1990

Roberts, L, (1973), "Dynamic allocation of satellite capacity through packet reservation", *AFIPS Conference Proceedings*, Vol. 42, pp. 711-716, 1973

Roorda, P D & Leung, V C M, (1996), "Dynamic control of time slot assignment in multi-access reservation protocols", *IEE Proceedings-Communications*, Vol. 143, No. 3, pp. 167-175, June 1996

Sanchez, J, Martinez, R & Marcellin, W, (1997), "A Survey of MAC Protocols Proposed for Wireless ATM", *IEEE Network Magazine*, pp. 52-62, November/December, 1997

Soroushnejad, M & Geraniotis, E, (1995), "Multi-access strategies for an integrated voice/data CDMA packet radio network", *IEEE Transactions on Communications*, Vol. 43, No. 2/3/4, pp. 934-945, Feb/Mar/Apr 1995

Sousa, E S & Silvester, J A, (1988), "Spreading code protocols for distributed spread spectrum packet radio networks", *IEEE Transactions on Communications*, Vol. 36, No. 3, pp. 272-281, March 1988

Stavrakakis, I & Kazakos, D, (1989), "A Limited Sensing Protocol for Multiuser Packet Radio Systems", *IEEE Transactions on Communications*, Vol. 37, No. 4, pp. 353-359, April 1989

Stern, H P & Sobol, H, (1995), "Design and performance analysis of an advanced, narrowband integrated voice/data mobile radio system", *IEEE Transactions on Communications*, Vol. 43, No. 1, pp. 107-115, January 1995

Storey, J S & Tobagi, F A, (1989), "Throughput performance of an unslotted direct-sequence SSMA packet radio network", *IEEE Transactions on Communications*, Vol. 37, No. 8, pp. 814-823, August 1989

Sunay, M O & McLane, P J, (1996), "Calculating error probabilities for DS CDMA systems: When not to use the Gaussian approximation", *Proceedings of GLOBECOM'96 Conference*, London, pp. 1744-1749, 1996

Taub, H & Schilling, D L, (1986), *Principles of communications systems,* 2nd Edition, McGraw-Hill International Editions, 1986

Thomopoulos, S C A, (1988), "A simple and versatile decentralised control for slotted ALOHA, Reservation ALOHA, and Local Area Networks", *IEEE Transactions on Communications*, Vol. 36, No. 6, pp. 662-674, June 1988

Tobagi, F A & Hunt, V B, (1980), "Performance Analysis of Carrier Sense Multiple Access with Collision Detection", Computer Networks, Vol. 4, pp. 245-259, October-November, 1980

Toshimitsu, K, Yamazato, T, Katayama, M & Ogawa, A, (1994), "A novel spread slotted ALOHA system with channel load sensing protocol", *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 4, pp. 665-672, May 1994

Turin, G L, (1984), "The Effects of Multipath Fading on the Performance of Direct-Sequence CDMA Systems", *IEEE Transactions on Vehicular Technology*, Vol. VT-33, pp. 213-219, August 1984

Turin, W & van Nobelen, R, (1998), "Hidden Markov Modelling of Flat Fading Channels", *IEEE Journal on Selected Areas in Communications*, Vol. 16, No.9, pp. 1809-1817, December 1998

van Nee, R D J, van Wolfswinkel, R N & Prasad, R, (1995), "Slotted ALOHA and code division multiple access techniques for land-mobile satellite personal communications", *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 2, pp. 382-388, February 1995

Viterbi, A M & Viterbi, A J, (1993), "Erlang Capacity of a Power Controlled CDMA System", *IEEE Journal on Selected Areas in Communications*, Vol. 11, No.6, pp. 892-900, August 1993

Wang, H S & Chang, P C, (1996), "On Verifying the First-Order Markovian Assumption for a Rayleigh Fading Channel Model", *IEEE Transactions on Vehicular Technology*, Vol. 45, No. 2, pp. 353-357, May 1996

Wieselthier, J E & Ephremides, A, (1995), "Fixed- and movable-boundary channel-access schemes for integrated voice/data wireless networks", *IEEE Transactions on Communications*, Vol. 43, No. 1, pp. 64-74, January 1995

Yang, W B & Geraniotis, E, (1994), "Admission policies for integrated voice and data traffic in CDMA packet radio networks", *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 4, pp. 654-664, May 1994

Yin, M & Li, V O K, (1990), "Unslotted CDMA with fixed packet lengths", *IEEE Journal on Selected Areas in Communications*, Vol. 8, No.4, pp. 529-541, May 1990

Zdunek, K J, Ucci, D R & Locicero, J L, (1989), "Throughput of Non-persistent Inhibit Sense Multiple Access with Capture", *Electron Letters*, Vol. 25, pp. 30-32, January 1989

Zorzi, M, Rao, R R & Milstein, L B, (1997), "ARQ Error Control for Fading Mobile Radio Channels", *IEEE Transactions on Vehicular Technology*, Vol. 46, No. 2, pp. 445-455, May 1997

APPENDIX

A.1 Introduction

The purpose of telecommunications modelling is to provide methods or solutions that can predict (or at least provide a rough estimate of) the performance of a communications system when presented with certain operating conditions. To model every intricate detail of real world systems is, to all extent and purposes, an almost impossible task given the overbearing number of variables and stochastic processes involved. For this purpose, we make various modelling assumptions that allow us to simplify the analysis such that the number of variables in the system can be reduced and such that the various stochastic processes can be modelled in forms that are well known and easy to compute. A highly detailed and intricate model that is impossible to solve is useless. Simplifying assumptions also allow for less computational effort and thus more rapid results. In some cases, even vast simplifications often provide results not too dissimilar from the real world result.

Simulators provide a means by which the assumptions made in the analytical model can be validated. Unlike the analytical model which obtains results through averaging mathematical representations of stochastic processes, the simulator models the actual stochastic processes in a dynamic fashion. In the analytical model, some form of steady state behaviour is assumed, and the system solution is found through averaging out results for all possible system states in a deterministic manner. In real world networks however, network events are random and dynamic. For this reason, simulators are based on measuring the system state variables at different *event times*, and at the end of the simulation run provide some form of average or count of the variables for the represented duration that the simulation was run for. Event times may be the start of a time slot (such as in a slotted-ALOHA system), or the arrival of a data message in the network. In the first case, the time between system events is more than likely fixed whereas, in the second case, the time between events might be a random process with a certain distribution. In either case, the simulator must provide some form of time index such that the various counts can be averaged over the respective time period for which the simulation was run.

For example, each time a voice call is generated in the network, the simulator might increment a certain count variable by one (e.g. CALLS=CALLS+1). If the call is blocked, then another variable will be incremented by one (e.g. BLOCKED_CALLS=BLOCKED_CALLS+1). If the simulation is run over say 100 units of time and, at the end of the simulation, the values of CALLS and BLOCKED_CALLS were 20 and 10 respectively, say, then we can easily compute the average voice call arrival rate to be 20/100=0.2 calls per unit time. The average voice call blocking probability can also be easily computed to be 10/20 = 50 percent blocking. The type of simulation model described above is known as a Monte Carlo model.

As in real telecommunications networks, the *events* themselves are more than likely stochastic in nature. For this reason, we need to incorporate some form of stochastic process into the simulator to ensure that events in the network occur in a stochastic manner. We can also tailor the stochastic process such that the time between events, or the nature or "size" of the event fits some form of distribution (e.g. Poisson, geometric, negative exponential etc...). Stochastic processes are dealt with in a computer programming context through the use of a *random number generator*. Strictly speaking, computers are incapable of producing purely random numbers. They are, however, capable of generating sequences of numbers which, when viewed over a certain time period, appear to be random in nature. These sequences of numbers are known as *pseudo-random number sequences*. Various deterministic recurrence algorithms exist for producing pseudo-random number sequences. By recurrence, it is meant that the sequence of numbers generated by the algorithm will eventually repeat itself. The *period* of the pseudo-random number sequence is defined as the number of values in the sequence before the sequence starts repeating itself again. Sequences with larger periods are generally better than sequences with short periods, since they avoid the case of repeatedly presenting the simulation with the same sequence of "random" numbers. There are various tests available for determining the relative "randomness" of a pseudo-random sequence. We do not provide further details, other than to point the interested reader to [Press *et al*, 1992] for a thorough explanation.

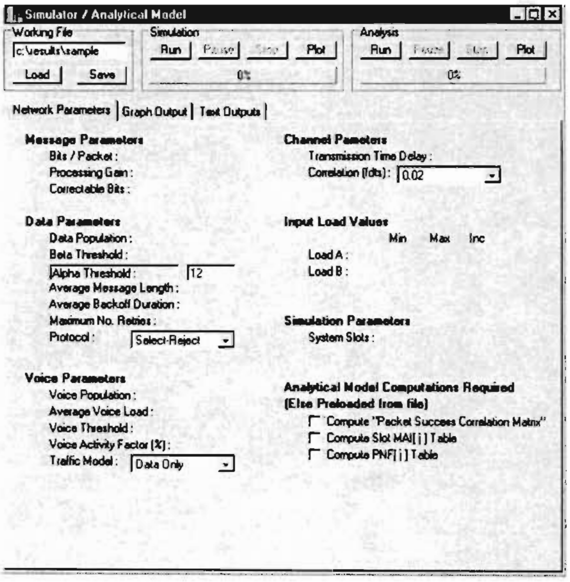
A.2 Custom Simulator / Analytical Model Package

The simulator and analytical models derived for this thesis were incorporated into a single 32-bit Windows 95 / Windows NT package using Borland C++ Code Builder. Justification for building a custom analytical results package was based on the fact that most equations were relatively simple in nature, and that C++ could provide rapid computation of the equations. In most cases, exact computation was obtainable, although in some cases (such as equations involving summations to infinity - e.g. equation 3-16) numerical methods were employed to detect when the summation had converged to a usable value, i.e. when further summation yielded a negligible change in the result.

Use was also made of the *IMSL®* (Integrated Mathematics & Statistics Libraries) library functions developed by the software company *Numerical Libraries*. This library contains several numerical and statistical functions that were very helpful in compiling the code to solve the analytical model equations and simulation, and include:

- excellent random number generators for most distributions (e.g. uniform, geometric, exponential, normal, and Poisson) for use in the simulator.
- several matrix functions that are used to generate solutions to a set of linear equations. These functions were useful in obtaining solutions to equations such as 4-11 and 6-26.
- numerical methods to solve the various integration functions presented in Chapter 7, including functions to solve double integrations.
- other specialised functions such as Bessel functions (used in equations 7-14 and 7-18) and the error complimentary function (*erfc*) (used in the standard Gaussian approximation for the probability of bit error)

An example screenshot of the software package's front-end is given in figure A.1. The user is able to specify all network input parameters via simple edit boxes, and is also able to save network parameters and results to a file. Two separate processor threads were used in the design of the package to allow for simultaneous, yet completely independent, simulation and analytical computation of the network. As can be seen from figure A.1, the package also allows for simulation and analytical results to be viewed, both in tabular form and graphical form.



Working File: c:\results\sample

Message Parameters
Bits / Packet :
Processing Gain :
Correctable Bits :

Data Parameters
Data Population :
Beta Threshold :
Alpha Threshold : 12
Average Message Length :
Average Backoff Duration :
Maximum No. Retries :
Protocol : Select-Reject

Voice Parameters
Voice Population :
Average Voice Load :
Voice Threshold :
Voice Activity Factor (%) :
Traffic Model : Data Only

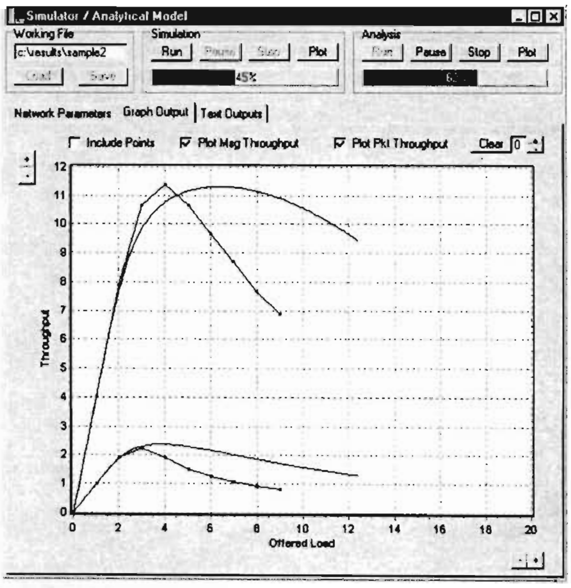
Channel Pameters
Transmission Time Delay :
Correlation (Idts) : 0.02

Input Load Values

	Min	Max	Inc
Load A :			
Load B :			

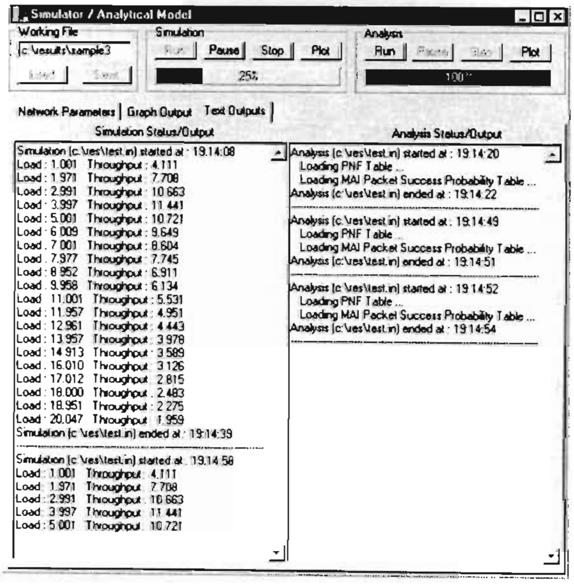
Simulation Parameters
System Slots :

Analytical Model Computations Required (Else Preloaded from file)
□ Compute "Packet Succees Correlation Matrix"
□ Compute Slot MAI[i] Table
□ Compute PNF[i] Table



Working File: c:\results\sample2

□ Include Points ☑ Plot Msg Throughput ☑ Plot Pkt Throughput 0



Working File: c:\results\sample3

Simulation Status/Output

```
Simulation (c:\res\test.in) started at : 19:14:08
Load : 1.001   Throughput : 4.111
Load : 1.971   Throughput : 7.708
Load : 2.991   Throughput : 10.663
Load : 3.997   Throughput : 11.441
Load : 5.001   Throughput : 10.721
Load : 6.009   Throughput : 9.649
Load : 7.001   Throughput : 8.604
Load : 7.977   Throughput : 7.745
Load : 8.952   Throughput : 6.911
Load : 9.958   Throughput : 6.134
Load : 11.001   Throughput : 5.531
Load : 11.957   Throughput : 4.951
Load : 12.961   Throughput : 4.443
Load : 13.957   Throughput : 3.978
Load : 14.913   Throughput : 3.589
Load : 16.010   Throughput : 3.126
Load : 17.012   Throughput : 2.815
Load : 18.000   Throughput : 2.483
Load : 18.951   Throughput : 2.275
Load : 20.047   Throughput : 1.959
Simulation (c:\res\test.in) ended at : 19:14:39
-----------------------------------------------------------
Simulation (c:\res\test.in) started at : 19:14:58
Load : 1.001   Throughput : 4.111
Load : 1.971   Throughput : 7.708
Load : 2.991   Throughput : 10.663
Load : 3.997   Throughput : 11.441
Load : 5.001   Throughput : 10.721
```

Analysis Status/Output

```
Analysis (c:\res\test.in) started at : 19:14:20
  Loading PNF Table ...
  Loading MAI Packet Success Probability Table ...
Analysis (c:\res\test.in) ended at : 19:14:22
-----------------------------------------------------------
Analysis (c:\res\test.in) started at : 19:14:49
  Loading PNF Table ...
  Loading MAI Packet Success Probability Table ...
Analysis (c:\res\test.in) ended at : 19:14:51
-----------------------------------------------------------
Analysis (c:\res\test.in) started at : 19:14:52
  Loading PNF Table ...
  Loading MAI Packet Success Probability Table ...
Analysis (c:\res\test.in) ended at : 19:14:54
-----------------------------------------------------------
```

Figure A.1 : Example screenshots of the custom simulator and analytical model package

A.3 Workings of the simulator

The simulator's running mode of operation is based on a fixed increment time process. This means that the time between each measurement of the system is fixed. Since the system is time slotted, it makes the most sense to evaluate the system states on a slot-by-slot basis. The following list contains some of the more important variables used to keep counts and measurements of the system state:

no_gen_voice : incremented each time a new voice call is generated
no_blk_voice : incremented each time a new voice call is blocked

no_ new_data : incremented each time a new data message arrives
no_ret _data : incremented each time an old data message retransmits
no_succ_data : incremented each time a data message succeeds
no_blk_data : incremented each time a data message is blocked (new or old)
no_succ_pkt : incremented each time a data packet is received uncorrupted

no_trans_data : variable holding the number of transmitting data users
no_trans_voice : variable holding the number of transmitting voice users

total_voice_load : increments by **no_trans_voice** at each slot
total_data_load : increments by **no_trans_data** at each slot

user_delay[n] : stores the number of slots since terminal n generated its last new message (this is the delay suffered by user n).
total_delay : increments by **user_delay[n]** each time user n successfully transmits his message.

msg_L_gen[500] : where **msg_L_gen** [X] represents the number of messages of length X that have been generated
msg_L_succ[500] : similarly for number of messages of length X that are successful
msg_L_ret[500] : similarly for distribution of retransmitted lengths

slot	: the number of the current slot
last_slot	: the total number of slots for which the system must run

The various system performance measures can then be easily computed using the various counts obtained over the system run. Of paramount importance in such simulations is the length of time the system is run for (size of **last_slot**). The longer the simulation duration, the more accurate the results are. This is due to the fact that more counts are obtained for each state occupancy, and thus when averaged over time, a more accurate aggregate is obtained. Another important reason why Monte Carlo simulations should be run for a lengthy period of time is that the system needs time to reach a steady state equilibrium point. The simulation should really only start counting and measuring system states once the system has reached a steady state, and continue doing so for a long time thereafter. This avoids including the unusual state counts that occur at the outset of the simulation when steady state has not been reached.

At the outset of the simulation, all counts are zeroed and various lookup tables are established. The various lookup tables hold values such as $P_{BE}^{SGA}(k)$ (equation 2-6), hence avoiding repetitive re-computation of values that are required often. The variable **slot** is then incremented by one, starting at zero and ending on **last_slot**. At each increment, the simulation calls upon the random number generators to determine the number of arriving voice calls and new data messages. The details of data messages which are to be retransmitted (such as length, number of times it has tried to transmit previously, the delay it has suffered to date, etc...) are stored in a separate array. The admission policies are then applied and the various counts are incremented to record how many new voice calls were blocked, how many data messages were successfully received etc...

After the simulation has run, the various performance measures are computed through simple averaging equations. We include the following set of equations to show how some of the more important system performance measures are obtained. We include pertinent measurements only. Other less important measurements are equally trivial to compute.

- The average carried voice load

$$\bar{v} = \frac{\textbf{total\_voice\_load}}{\textbf{last\_slot}} \qquad \textbf{(A-1)}$$

- The average carried data load

$$\bar{x} = \frac{\textbf{total\_data\_load}}{\textbf{last\_slot}} \qquad \textbf{(A-2)}$$

- The average voice call arrival rate

$$G_v = \frac{\textbf{no\_gen\_voice}}{\textbf{last\_slot}} \qquad \textbf{(A-3)}$$

- The average voice call blocking probability

$$P_B^{voice} = \frac{\textbf{no\_blk\_voice}}{\textbf{no\_gen\_voice}} \qquad \textbf{(A-4)}$$

- The average data message throughput

$$S_M = \frac{\textbf{no\_succ\_data}}{\textbf{last\_slot}} \qquad \textbf{(A-5)}$$

- The average data packet throughput

$$S_p = \frac{\textbf{no\_succ\_pkt}}{\textbf{last\_slot}} \qquad \textbf{(A-6)}$$

- The average offered data load

$$G = \frac{\textbf{no\_new\_data} + \textbf{no\_ret\_data}}{\textbf{last\_slot}} \qquad \textbf{(A-7)}$$

- The average data message blocking probability

$$P_B^{data} = \frac{\textbf{no\_blk\_data}}{\textbf{no\_new\_data} + \textbf{no\_ret\_data}} \tag{A-8}$$

- The probability of having a data message of length l attempting to transmit

$$P\{L_d = l\} = \frac{\textbf{msg\_L\_gen}[l] + \textbf{msg\_L\_ret}[l]}{\textbf{no\_new\_data} + \textbf{no\_ret\_data}} \tag{A-9}$$

- The probability of success of a data message of length L

$$P(S|L) = \frac{\textbf{msg\_L\_succ}[L]}{\textbf{msg\_L\_gen}[L] + \textbf{msg\_L\_ret}[L]} \tag{A-10}$$

- The overall probability of data message success

$$P(S) = \frac{\textbf{no\_succ\_data}}{\textbf{no\_new\_data} + \textbf{no\_ret\_data}} \tag{A-11}$$

- The expected data message delay

$$\overline{D} = \frac{\textbf{total\_delay}}{\textbf{no\_new\_data}} \tag{A-12}$$