# Call Admission Control in Wireless Communications Networks

by

Michael Cheung

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2000

0-612-53488-X

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

# Abstract

This thesis is concerned with call admission control (CAC) in wireless communication networks. A salient feature of mobile wireless communications is the support of roaming. However, user mobility has a profound effect on QoS provisioning. A study of the effect of user mobility on QoS provisioning is first studied. It provides a fundamental understanding of the effect of user roaming on utilization performance and capacity requirement. For a CDMA environment, a linkage between the link layer specification and network layer QoS performance is established. Second, a proposed joint-levels resource allocation (RA) scheme is studied. This scheme is based on simultaneous satisfaction of the packet-level and call-level QoS constraints. It is shown that the utilization is greatly increased. Third, a proposed movable boundary (MB) channel allocation scheme for two traffic classes, real-time and non-real-time, is discussed. By allowing non-real-time traffic to borrow idle real-time channels, utilization is enhanced. Fourth, a combined scheme, by integrating the joint-levels RA scheme and MB channel allocation scheme, is presented. Results have shown that the combined scheme improves the utilization performance drastically. Lastly, a CAC mechanism based on the combined scheme is structured.

# Acknowledgements

My course of Ph.D. studies is memorable and enjoyable. Because of the wonderful people at University of Waterloo, my journey has been a smooth one. I would like to take this opportunity to give my gratitude to those who have helped me throughout the last four years.

First, I would like to thank my supervisor, Professor Jon W. Mark, for his guidance and support throughout the course of my studies. Your patience and encouragement give me continuous support. You are a crucial part of my personal growth and professional development. Your sincerity and dedication have taught me how to become a better person. Without you, completion of this degree would never be possible. You are a truly unmatchable scholar. I would not have learned more from any other person than from you. Thank you.

I would also like to thank my Ph.D. committee members for their advice and suggestions, namely Dr. Johnny Wong, Dr. Xuemin Shen, Dr. Anwar Hasan and Dr. Terry Todd. I also owe my deep gratitude to Dr. Zhuang and Dr. Yoon for their timeless support and thoughtful advice. My thanks also go to Dr. Anthony Vannelli, Chair of the department. Thank you for your encouragement.

I must also acknowledge the administrative support staff, namely Wendy Boles, Jennifer Werth, Gini-Ivan-Roth, Donna O'Brecht, Richard Leah, Wendy Stoneman, Simone Thibodeau, Wendy Gauthier, Shaz Rahaman, Chris Lorenz and Monica Ruppert. You have made my life as a graduate student so much easier.

I would like to thank all the colleagues in the group for their advice, namely, Dr. Nasir Ghani, Dr. Robert Lehr, Dr. Jianping Pan, Dr. Shihua Zhu, Mr. Jiangzhu Zhang, Mr. Joseph Zhang, Mr. James Qiu, Mr. Shamir Mukhi, Mr. Nainesh Agarwal, Mr. Yu Zeng, Mr. Jun Ye, Mr. Liang Xu, and Ms. Dongmei Zhao.

I need to thank all my personal friends for their support and encouragement. In particular, I want to thank Mr. Michael Kwok, Mr. Anthony Ho, Mr. Otto Yung, Ms. Grace Wong, Mr. Kenneth Cheung, Mr. Patrick Lau, Dr. Vincent Wong, Mr. Kai Kwok, and Mr. Jack Wong. I know I have missed a few. But many thanks to all my friends. You have made the degree possible. I should also thank all my badminton buddies. We have enjoyable moments.

I need to acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC), Canadian Institute of Telecommunications Research (CITR), Communications and Information Technology Ontario (CITO) and the Ontario government for the financial support. I also thank the University of Waterloo for the various awards.

Finally, I owe my deepest gratitude to my dear parents. Thank you for your understanding and patience. Without you, the completion of this thesis will never be possible. Thank you for your love.

# Contents

1 Introduction 1

2 Background and Motivation 7

2.1 Review on Resource Allocation . . . . . . . . . . . . . . . . . . . . 9

2.1.1 Packet-Level Resource Allocation . . . . . . . . . . . . . . . . 9

2.1.2 Call-Level Resource Allocation . . . . . . . . . . . . . . . . . 11

2.2 Review on Call Admission Control . . . . . . . . . . . . . . . . . . 13

2.2.1 CAC in Network Layer . . . . . . . . . . . . . . . . . . . . . 14

2.2.2 CAC in Link Layer . . . . . . . . . . . . . . . . . . . . . . . 16

2.3 Motivation and Goals . . . . . . . . . . . . . . . . . . . . . . . . . 17

2.3.1 Effect of Mobility on QoS Provisioning . . . . . . . . . . . . 17

2.3.2 Coupling of the Link and Network Layer Requirements . . . 18

2.3.3 Joint Packet-Level and Call-Level Resource Allocation . . . 20

2.3.4 Movable Boundary Channel Allocation . . . . . . . . . . . . 21

2.4 Summary of Objectives . . . . . . . . . . . . . . . . . . . . . . . . . 23

**3 Effects of Mobility on QoS Provisioning 25**

3.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . 27

3.2 Performance Analysis . . . . . . . . . . . . . . . . . . . . . . . . . 36

3.3 Effect of Mobility on System Performance . . . . . . . . . . . . . . 43

3.3.1 Utilization Degradation . . . . . . . . . . . . . . . . . . . . . 44

3.3.2 Excess Capacity Requirement . . . . . . . . . . . . . . . . . . 45

3.3.3 $E_b/I_0$ Requirement . . . . . . . . . . . . . . . . . . . . . . . 46

3.4 Numerical Results and Discussions . . . . . . . . . . . . . . . . . . 47

3.4.1 Utilization Degradation . . . . . . . . . . . . . . . . . . . . . 47

3.4.2 Excess Capacity Requirements . . . . . . . . . . . . . . . . . . 49

3.4.3 $E_b/I_0$ Requirements . . . . . . . . . . . . . . . . . . . . . . . 53

3.5 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

**4 Joint-Levels Resource Allocation 58**

4.1 The Resource Allocation (RA) Problem . . . . . . . . . . . . . . . . 59

4.1.1 Packet-Level RA . . . . . . . . . . . . . . . . . . . . . . . . . 60

4.1.2 Call-level RA . . . . . . . . . . . . . . . . . . . . . . . . . . . 61

4.1.3 Joint Packet and Call Level Resource Allocation (RA) . . . 63

4.2 Joint-Levels RA in a Cellular Environment . . . . . . . . . . . . . . 66

4.3 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . 75

5 Channel Allocation for Two Traffic Types 76

5.1 System Model . . . . . . . . . . . . . . . . . . . . . . . . . . . 77

5.2 Performance Analysis . . . . . . . . . . . . . . . . . . . . . . . 79

5.2.1 Case of $r = 1$ . . . . . . . . . . . . . . . . . . . . . . . 80

5.2.2 Case of $r \geq 1$ . . . . . . . . . . . . . . . . . . . . . . . 82

5.2.3 Delay, $D$ . . . . . . . . . . . . . . . . . . . . . . . . . . 85

5.3 Numerical Results . . . . . . . . . . . . . . . . . . . . . . . . . 86

5.3.1 Effect of $\sigma$ on $K_{NRT}$ . . . . . . . . . . . . . . . . . . . . 87

5.3.2 The Effect of User Mobility . . . . . . . . . . . . . . . . . 89

5.3.3 Utilization Degradation . . . . . . . . . . . . . . . . . . . 91

5.3.4 Excess Capacity . . . . . . . . . . . . . . . . . . . . . . . 95

5.4 Discussions . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98

5.5 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99

6 Call Admission Control 100

6.1 Performance of the Combined Scheme . . . . . . . . . . . . . . . . 101

6.2 Call Admission Control . . . . . . . . . . . . . . . . . . . . . . . 107

7 Conclusions and Future Works 110

7.1 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110

7.2 Future Works . . . . . . . . . . . . . . . . . . . . . . . . . . . 113

**A Forced Termination Probability 116**

**B Typical Values of Parameters 119**

**C Fluid Flow Analysis 121**

**Bibliography 125**

# List of Tables

4.1 Maximum value of $K_c$, $N_u$, and $\rho_c$ when PLR$^* = 10^{-6}$ and $P_n^* = 10^{-3}$ 65

4.2 Performance in CDMA environment $P_c^{(1)}$ with PLR$^* = 10^{-6}$, $P_n^* = 10^{-3}$, and $P_f^* = 10^{-6}$, from $\lambda_h = 0$ to $\lambda_h = 10$ . . . . . . . . . . . . . 73

4.3 Performance in CDMA environment $P_c^{(2)}$ with PLR$^* = 10^{-6}$, $P_n^* = 10^{-3}$, and $P_f^* = 10^{-6}$, from $\lambda_h = 0$ to $\lambda_h = 10$ . . . . . . . . . . . . . 73

4.4 Performance in CDMA environment $P_c^{(3)}$ with PLR$^* = 10^{-6}$, $P_n^* = 10^{-3}$, and $P_f^* = 10^{-6}$, from $\lambda_h = 0$ to $\lambda_h = 10$ . . . . . . . . . . . . . 73

5.1 Parameters for the two considered cases . . . . . . . . . . . . . . . . . 87

5.2 Utilization performance of the MB channel allocation scheme in terms of utilization with $P_n^* = P_f^* = 10^{-3}$, from $\lambda_h = 0$ to $\lambda_h = 10$ . 94

5.3 Excess capacity performance of the MB channel allocation scheme in terms of utilization with $P_n^* = P_f^* = 10^{-3}$, from $\lambda_h = 0$ to $\lambda_h = 10$ 98

6.1 Performance of the combined scheme for the RT service in terms of utilization with $P_n^* = P_f^* = 10^{-3}$. from $\lambda_h = 0$ to $\lambda_h = 8$ . . . . . . . 104

6.2 Performance of the combined scheme for the NRT service in terms of utilization with $P_n^* = P_f^* = 10^{-3}$. from $\lambda_h = 0$ to $\lambda_h = 8$ . . . . . 107

B.1 Typical call duration and $\mu$. . . . . . . . . . . . . . . . . . . . . . . . 120

B.2 Typical user speed and handoff rate $\lambda_h$. . . . . . . . . . . . . . . . . . 120

# List of Figures

1.1 System Model . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2 System Model . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.3 A wireless subnet . . . . . . . . . . . . . . . . . . . . . . . 5

2.1 Capacity region for $M = 2$ . . . . . . . . . . . . . . . . . . . . . . 15

3.1 Accuracy of the exponential inter-handoff time assumption with $K_c = 35$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33

3.2 Accuracy of the exponential inter-handoff time assumption with $K_c = 50$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

3.3 Queueing Model for Homogeneous Traffic . . . . . . . . . . . . . . 36

3.4 Utilization and Capacity Requirement as a function of mobility ($K_c = 139$, $\mu^{-1} = 2$ mins, $P_n^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . 50

3.5 Utilization degradation and excess capacity requirement as a function of mobility ($K_c = 139$, $\mu^{-1} = 2$ mins, $P_n^* = 10^{-3}$) . . . . . . . . 50

3.6 Percentage utilization degradation and excess capacity requirement as a function of mobility ($K_c = 139$, $\mu^{-1} = 2$ mins, $P_n^* = 10^{-3}$) . . . 51

3.7 Utilization and capacity requirement as a function of mobility ($\mu^{-1} = 2$ mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . 52

3.8 Utilization degradation and excess capacity requirement as a function of mobility ($\mu^{-1} = 2$ mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$) . . . . . . . 53

3.9 Percentage utilization degradation and excess capacity requirement as a function of mobility ($\mu^{-1} = 2$ mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$) . . 54

3.10 $E_b/I_0$ requirement in dB as a function of mobility ($\mu^{-1} = 2$ mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . 56

3.11 $E_b/I_0$ requirement in dB as a function of mobility ($K_c = 139$, $\mu^{-1} = 2$ mins, $P_n^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . . . . 56

3.12 $E_b/I_0$ and capacity tradeoff ($K_c = 139$, $\mu^{-1} = 2$ mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . . . . . . 57

4.1 Two-state on-off voice model . . . . . . . . . . . . . . . . . . . . . 61

4.2 Illustration of the joint-levels RA as a queueing model . . . . . . . . 64

4.3 Achievable PLR using joint-levels RA ($P_n^* = 10^{-3}$) . . . . . . . . . . 66

4.4 Achievable PLR using joint-levels RA in CDMA environment ($P_n^* = 10^{-3}$, $\lambda_h = 0$) . . . . . . . . . . . . . . . . . . . . . . . . . . . 70

4.5 Call-level utilization factor in CDMA environment $P_c^{(1)}$ (PLR$^*$ = $10^{-6}$, $P_n^* = 10^{-3}$, and $P_f^* = 10^{-6}$) . . . . . . . . . . . . . . . . . . . . 71

4.6 Call-level utilization factor in CDMA environment $P_c^{(2)}$ (PLR$^*$ = $10^{-6}$, $P_n^* = 10^{-3}$ and $P_f^* = 10^{-6}$) . . . . . . . . . . . . . . . . . . . . . 71

4.7 Call-level utilization factor in CDMA environment $P_c^{(3)}$ (PLR$^*$ = $10^{-6}$, $P_n^* = 10^{-3}$ and $P_f^* = 10^{-6}$) . . . . . . . . . . . . . . . . . . . . . 72

5.1 RT and NRT channel allocation with a movable boundary . . . . . 79

5.2 NRT Birth-Death Process . . . . . . . . . . . . . . . . . . . . . . 81

5.3 NRT Birth-Death Process for $r \geq 1$ . . . . . . . . . . . . . . . . . . 84

5.4 Channel Requirement as a function of Utilization for different $\sigma$ with $r = 6$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89

5.5 Preemption Probabilities as a function of Utilization for different $\sigma$ with $r = 6$ . . . . . . . . . . . . . . . . . . . . . . . . . . 90

5.6 NRT channel requirement as a function of utilization for different mobility and $r$ ($P_n^* = P_f^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . 91

5.7 NRT utilization as a function of mobility ($N_{u_{NRT}} = 80$, $r = 3$, $P_n^* = P_f^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . . . 93

5.8 NRT utilization degradation as a function of mobility ($N_{u_{NRT}} = 80$, $r = 3$, $P_n^* = P_f^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . . 93

5.9 NRT % utilization degradation as a function of mobility ($N_{u_{NRT}} = 80$, $r = 3$, $P_n^* = P_f^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . . . . 94

5.10 NRT capacity requirement as a function of mobility ($N_{u_{NRT}} = 80$, $r = 3$, $P_n^* = P_f^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . . . . 96

5.11 NRT excess capacity requirement as a function of mobility ($N_{u_{NRT}} = 80$, $r = 3$, $P_n^* = P_f^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . 97

5.12 NRT % excess capacity requirement as a function of mobility ($N_{u_{NRT}} = 80$, $r = 3$, $P_n^* = P_f^* = 10^{-3}$) . . . . . . . . . . . . . . . . . . . . . . 97

6.1 Utilization as a function of mobility for RT service ($c = 1.5$ Mbps, $\mu^{-1} = 2$ mins, $P_n^* = P_f^* = 10^{-3}$, PLR$^* = 10^{-6}$) . . . . . . . . . . . . 102

6.2 Utilization degradation as a function of mobility for RT service ($c =$ 1.5 Mbps, $\mu^{-1} = 2$ mins, $P_n^* = P_f^* = 10^{-3}$, PLR$^* = 10^{-6}$) . . . . . . 103

6.3 % utilization degradation as a function of mobility for RT service ($c = 1.5$ Mbps, $\mu^{-1} = 2$ mins, $P_n^* = P_f^* = 10^{-3}$, PLR$^* = 10^{-6}$) . . . 103

6.4 Utilization as a function of mobility for NRT service ($c = 1.5$ Mbps, $\mu^{-1} = 2$ mins, $P_n^* = P_f^* = 10^{-3}$, PLR$^* = 10^{-6}$) . . . . . . . . . . . . 105

6.5 Utilization degradation as a function of mobility for NRT service ($c = 1.5$ Mbps, $\mu^{-1} = 2$ mins, $P_n^* = P_f^* = 10^{-3}$, PLR$^* = 10^{-6}$) . . . 106

6.6 % utilization degradation as a function of mobility for NRT service ($c = 1.5$ Mbps, $\mu^{-1} = 2$ mins, $P_n^* = P_f^* = 10^{-3}$, PLR$^* = 10^{-6}$) . . . 106

# List of Abbreviations

| | |
|---|---|
| 2G | second generation |
| 3G | third generation |
| AP | access point |
| ATM | Asynchronous Transfer Mode |
| BER | bit error rate |
| bps | bits per second |
| BS | base station |
| CAC | call admission control |
| CDMA | code division multiple access |
| EDGE | Enhanced Data Rate for GSM Evolution |
| FDMA | frequency division multiple access |
| GPRS | General Packet Radio Service |
| GSM | Global System for Mobile Communications |
| IMT | International Mobile Telecommunications |
| IP | Internet Protocol |
| JDC | Japan Digital Cellular |
| MAC | media access control |
| MB | movable boundary |

| | |
|---|---|
| MMPP | markov modulated poisson process |
| MT | mobile terminal |
| NRT | non-real-time |
| QoS | quality of service |
| RA | resource allocation |
| RT | real-time |
| TDMA | time division multiple access |
| TDM | time division multiplexing |
| SIR | signal-to-interference ratio |
| VCT | virtual connection tree |
| W-CDMA | wideband CDMA |

# Glossary of Symbols

| | |
|---|---|
| $a$ | number of attempts before a successful connection re-establishment |
| $\alpha$ | carried load |
| $\alpha_s$ | on-to-off rate of an on-off source |
| $B$ | buffer size |
| $\beta$ | service rate of a cell |
| $\beta_s$ | source activity factor |
| $c$ | link rate in bits per second |
| $c_{\mathrm{NRT}}$ | link rate allocated to NRT traffic in bits per second |
| $c_{\mathrm{RT}}$ | link rate allocated to RT traffic in bits per second |
| $C$ | physical capacity of the system in number of channels |
| $C_j$ | the $j^{\mathrm{th}}$ possible capacity of the system due to channel variation |
| $D$ | delay experienced by NRT traffic |
| $\Delta D$ | excess delay due to the MB scheme |
| $E_b$ | bit energy |
| $\varepsilon_i$ | marginal packet loss rate when there are $i$ ongoing traffic flows |
| $\varepsilon_{ij}$ | marginal packet loss rate when there are $i$ ongoing traffic flows and the capacity is $C_j$ |
| $\eta$ | noise density to interference density ratio |

| | |
|---|---|
| $\eta_s$ | off-to-on rate of an on-off source |
| $f$ | ratio of other interference-to-own cell interference |
| $\Gamma_{m_j}$ | traffic demand of the $j^{\text{th}}$ type $m$ traffic |
| $I_0$ | maximum total acceptable interference density |
| $I_{t_{m_j}}$ | max. tolerable total interference experienced by the $j^{\text{th}}$ type $m$ traffic |
| $J$ | number of values that the capacity can take due to channel variation |
| $K$ | number of users that can be handled by the BS |
| $K_c$ | number of user channels |
| $K_{c_m}$ | number of user channels for type $m$ traffic |
| $K_m$ | channel requirement when all users are in motion |
| $K_s$ | channel requirement when all users are stationary |
| $K_g$ | number of guard channels |
| $K_{\mathrm{NRT}}$ | number of NRT channels |
| $K_{\mathrm{RT}}$ | number of RT channels |
| $\Delta K_c$ | excess channel requirement due to user mobility |
| $\Delta K_{\mathrm{NRT}}$ | NRT excess channel requirement |
| $\Delta K_{\mathrm{RT}}$ | RT excess channel requirement |
| $l$ | buffer occupancy |
| $\lambda$ | call arrival rate |

| | |
|---|---|
| $\Lambda$ | aggregate arrival rate |
| $\lambda_h$ | call handoff rate |
| $\lambda_{h_{cr}}$ | critical handoff rate |
| $\lambda_{\mathrm{NRT}}$ | NRT call arrival rate |
| $\lambda_{\mathrm{RT}}$ | RT call arrival rate |
| $M$ | number of traffic classes |
| $\mu$ | call termination rate |
| $\mu_{\mathrm{NRT}}$ | call termination rate for a NRT call |
| $\mu_{\mathrm{RT}}$ | call termination rate for a RT call |
| $n_{u_m}$ | number of ongoing calls of type $m$ |
| $N$ | number of users |
| $N_0$ | thermal (background) noise density |
| $N_m$ | utilization when all mobiles are in motion |
| $N_p$ | number of preemptions before call completion |
| $N_{u_m}$ | maximum number of ongoing calls of type $m$ |
| $N_u$ | utilization (number of users) |
| $N_s$ | utilization when all users are stationary |
| $N_{u_{\mathrm{NRT}}}$ | utilization of NRT calls |
| $N_{u_{\mathrm{RT}}}$ | utilization of RT calls |

| | |
|---|---|
| $\Delta N_u$ | utilization degradation due to user mobility |
| $\Delta N_{u_{\mathrm{NRT}}}$ | utilization degradation of NRT calls |
| $\Delta N_{u_{\mathrm{RT}}}$ | utilization degradation of RT calls |
| $p_{\mathrm{NRT}}$ | probability that the arrival is a NRT call |
| PLR | packet loss rate |
| PLR$^*$ | packet loss rate requirement |
| $P$ | generic call blocking probability at the cell |
| $P^*$ | generic call blocking probability requirement at the cell |
| $P_{C_j}$ | probability that the capacity is $C_j$ |
| $P_f$ | forced termination probability |
| $P_f^*$ | forced termination probability requirement |
| $P_m$ | generic blocking probability when all mobiles are in motion |
| $P_m^*$ | generic blocking requirement when all mobiles are in motion |
| $P_p$ | call preemption probability |
| $P_{\mathrm{NRT}}$ | NRT call blocking probability |
| $P_{\mathrm{NRT}}^*$ | NRT call blocking probability requirement |
| $P_{\mathrm{RT}}$ | RT call blocking probability |
| $P_{\mathrm{RT}}^*$ | RT call blocking probability requirement |
| $P_n$ | new call blocking probability |

| | |
|---|---|
| $P_n^*$ | new call blocking probability requirement |
| $P_s$ | generic blocking blocking when all mobiles are stationary |
| $P_s^*$ | generic blocking requirement when all mobiles are stationary |
| $o$ | effective bandwidth weighting factor |
| $r$ | number of channels that a NRT source requires |
| $R$ | data rate |
| $R_m$ | data rate of type $m$ traffic |
| $R_{mj}$ | data rate of the $j^{th}$ type $m$ traffic |
| $\rho_c$ | call-level utilization factor |
| $S$ | number of slots in a TDM frame allocated to voice user |
| $S_{mj}$ | transmit signal strength of the $j^{th}$ type $m$ traffic |
| $\theta$ | mobility parameter, the probability that the handoff is due to handoff |
| $\sigma$ | number of RT channels that NRT calls may borrow |
| $\tau$ | time between connection re-establishment attempts |
| $W$ | spread spectrum bandwidth |
| $\zeta$ | amount of RT capacity which may be borrowed by NRT traffic |

# Chapter 1

# Introduction

Wireless technology provides tetherless communication access. The prevailing penetration of the second generation wireless systems has proven the wide acceptance of wireless communications. The second generation systems, such as GSM, IS-136 and IS-95, are digital narrowband systems. They can support only digitized voice, and limited low bit rate data services. As the popularity of the internet and the world wide web grows at an exponential rate, the extension of these services from the fixed wireline domain to the wireless domain becomes inevitable. The need of multimedia support manifests the inadequacy of the second generation wireless systems. The drive toward the third generation (3G) is now on a full swing. Figure 1.1 gives the currently envisioned evolution path of wireless systems from the second generation (2G) to 3G.

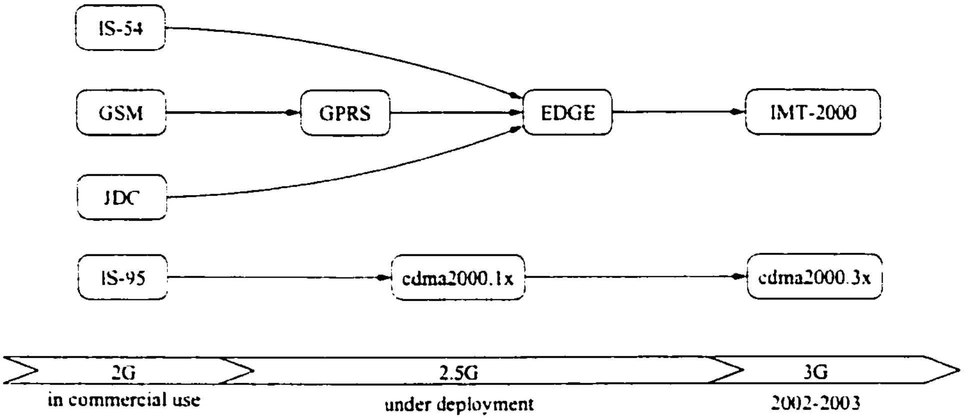

Figure 1.1: System Model

There are four major 2G standards: IS-136 (IS-54). GSM (The Global System for Mobile Communications). JDC (Japan Digital Cellular). and IS-95 (cdmaone). They are all TDMA-based (time division multiple access). except for IS-95. which is CDMA-based (code division multiple access). IS-136. GSM. and JDC are mainly used in North America. Europe. and Japan. respectively. IS-95 has its major market in North America. The 2G systems are circuit switched. making it inefficient for digitized information transfer. In addition. the bit rate of the 2G systems are limited to 9.6 kbps to 14.4 kbps. This makes 2G insufficient for multimedia traffic.

Most of the air interface proposals for the 3G are CDMA-based. e.g.. cdma2000.3x from North America and W-CDMA in IMT-2000 (International Mobile Telecommunications) from Europe. In fact. its ability of universal frequency reuse. soft handoff and multipath resistance. etc.. make CDMA the de-facto standard for 3G wireless systems. The 3G system is expected to support information rate of up to

2 Mbps for stationary wireless access and about 500 kbps with mobility.

Except for IS-95. evolution toward the 3G involves major changes to existing infrastructure due to the inherent difference in the multiple access technology. To meet current demand of higher bit rates. modification to the current 2G systems is underway so that minimal changes are involved. allowing a cost-effective migration toward the 3G. The evolved 2G systems are called the 2.5G systems. This includes the GPRS (General Packet Radio Service) systems and the EDGE (Enhanced Data Rate for GSM Evolution). The most distinctive difference between the 2G systems and 2.5G systems is the support of packet switching. With the support of multiple time slot assignment and more efficient modulation scheme. information rate of up to 384 kbps is expected. EDGE and cdma2000.1x represent the major stepping stone toward 3G.

The network under consideration is shown in Figure 1.2. It consists of a collection of wireless subnets interconnected through a wireline backbone network. The wireline backbone is an IP/ATM based network which supports multimedia traffic and guarantees quality of service (QoS). Each wireless subnet. commonly called a radio cell. consists of a base station (BS) or access point (AP). serving a number of mobile terminals (MT's) within its coverage area. The BS's are connected to the backbone wireline network through a mobile switching centre (MSC). As its name implies. the MSC has the responsibility of collecting information and making control decisions. e.g.. handoff decisions. The MSC also serves as the wireline attachment point for a wireless subnet.

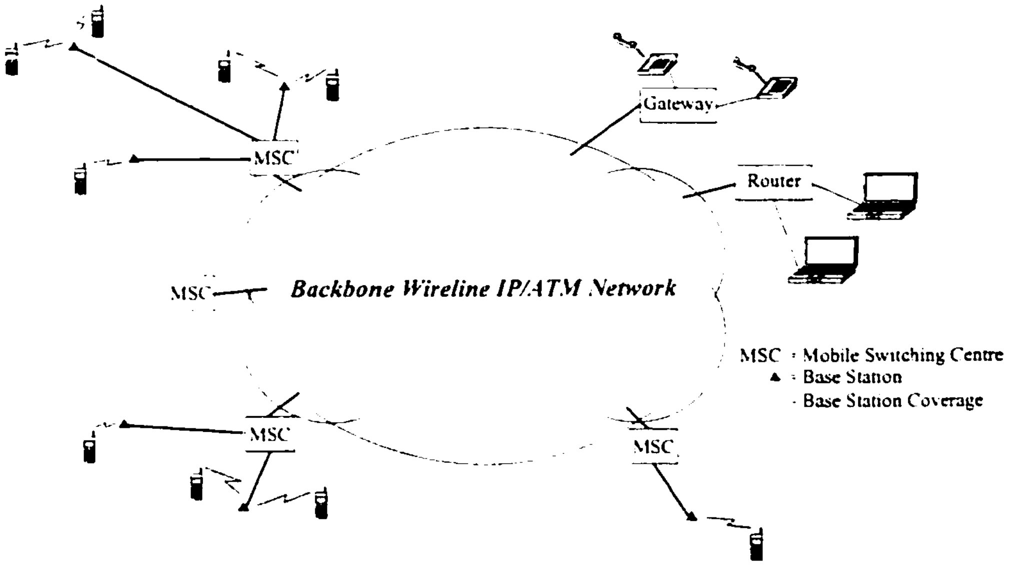

Figure 1.2: System Model

Resource management is essential in QoS provisioning [1. 2. 3. 4. 5. 6]. It includes flow control. resource allocation (RA). congestion control. call admission control (CAC). etc. Resource management in the wireline ATM network is already very challenging due to the complexity in the traffic behaviour of the multiplexed multimedia traffic. In a wireless network. the presence of user mobility poses another level of complexity. The main focus of this thesis is in resource management and call admission control in a wireless subnet. shown in Figure 1.2.

This thesis has four main objectives:

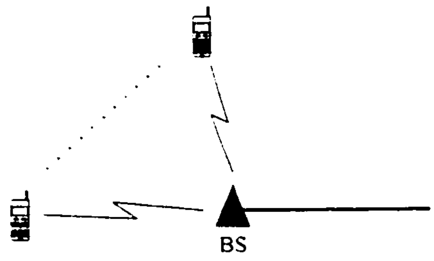

Figure 1.3: A wireless subnet

1. to study the effect of mobility on QoS provisioning, and to provide a linkage between the link layer specification and network layer performance.
2. to propose a new resource allocation scheme based on simultaneous satisfaction of the packet-level and call-level QoS factors.
3. to propose a novel channel allocation scheme to support two QoS classes based on a movable boundary approach, and
4. to structure a CAC mechanism, using the understanding of the effect of mobility, and incorporating the proposed resource allocation scheme and channel allocation scheme.

The rest of this thesis is organized as follows. In Chapter 2, an overview of resource management and call admission control is presented. Recent work in the literature is reviewed. The motivation for pursuing the research is then given.

Chapter 3 studies the effect of mobility on QoS provisioning. The analytical queueing model is first presented, then the effect of mobility on utilization performance and capacity requirement is discussed.

In Chapter 4, the proposed joint-levels resource allocation (RA) scheme is described. It is based on simultaneous satisfaction of both the packet-level and call-level QoS factors. Numerical examples show that it allows more multiplexing and increases the system utilization.

Chapter 5 describes the proposed movable boundary (MB) channel allocation scheme. It supports two QoS classes, real-time (RT) and non-real-time (NRT). In this scheme, NRT traffic may borrow idle RT channels on the basis that it may be preempted if the borrowed channels are required by the RT traffic. This scheme also increases the system utilization significantly.

In Chapter 6, the joint-levels RA scheme and MB channel allocation scheme are integrated. The combined scheme also takes into consideration the mobility effects obtained in Chapter 3. Based on the combined scheme, a CAC mechanism is structured. It improves the system utilization significantly.

Chapter 7 gives the concluding remarks. The contributions of this thesis are summarized. Some interesting further research topics are also described.

# Chapter 2

# Background and Motivation

Call admission control (CAC) is to ensure the network integrity by restricting access to the network to avoid overload and congestion. at the same time, to ensure the quality of service (QoS) requirements of all on-going connections are satisfied. The CAC problem is the following. Suppose there are $N_u - 1$ ongoing calls. When the $N_u^{\text{th}}$ request arrives, the network calculates the available resources. If there is enough resource to admit the $N_u^{\text{th}}$ call such that its QoS requirements, as well as those of the ongoing calls, are satisfied, then admission is granted. Otherwise, the connection request is denied.

Resource allocation is a vital component of CAC. The performance of CAC depends on the effectiveness of resource allocation. Given a knowledge of the amount of physical resources, resource allocation determines the maximum number of calls of each of the $M$ classes that can be supported while satisfying all the QoS constraints.

To set the stage for our discussion of cellular mobile communications and networking, the communications problem is divided into three conceptual levels: (i) the physical level which involves modulation, coding, receiver design, etc.; (ii) the link level where quality is characterized by the error rate or signal-to-interference ratio (SIR); and (iii) the network level that is characterized by the protocol layers above the media access control (MAC) protocol in the protocol stack.

Resource allocation is a network level problem. In a wireless mobile network, a mobile user can freely travel from one place to another during a connection. This gives rise to the necessity for handoff, in order to ensure service continuity. Handoff is unique to wireless mobile networking. Therefore, the effect of user mobility needs to be taken into consideration in resource allocation. The rest of this chapter provides a literature review of resource allocation and call admission control in wireless communications. Their shortcomings are identified and the motivation behind this thesis is then presented.

## 2.1 Review on Resource Allocation

Resource allocation are often considered at either the packet level or the call level. The packet-level QoS factors include the packet loss rate, PLR, the mean packet delay, and the packet delay jitter. The delay and delay variation requirements can often be met by choosing an appropriate buffer size. Therefore, the primary packet-level QoS factor of interest is PLR. The call-level QoS factors of interest include the new call blocking probability, $P_n$, the forced termination probability, $P_f$. In resource allocation, the network resources are assigned to the ongoing calls on the ground that their QoS requirements, at both the packet and call level, are satisfied.

### 2.1.1 Packet-Level Resource Allocation

The packet-level resource allocation is concerned with the assignment of the packet transmission capacity of the link. It attempts to maximize the system utilization through statistical multiplexing. Resource at the packet level is the link transmission capacity, often measured in bits per second (bps) or packets per second. When a link capacity is specified, the resource allocation problem in a wireless system is very similar to its wireline counterpart.

There are three major approaches to resource allocation: pure analytical-based, pure measurement-based, and hybrid analytical/measurement-based. In analytical-based approaches (e.g., [7, 8, 9, 10, 11, 12, 13, 14, 15]), sources are assumed to conform to some mathematical models. The resource allocation problem is then solved analytically or numerically. For example, effective bandwidth [10] is one of the most popular analytical-based resource allocation approaches ([7] to [15]), where the packet arrival stream is modeled by an on/off process with exponential on and off periods. Service at the packet level is modeled using a single server queue with infinite waiting room. The PLR can then be estimated by the tail distribution of the buffer occupancy, described by a set of differential equations. Analytic tractability is the main advantage of the effective bandwidth approach and other analytical-based schemes.

In measurement-based approaches (e.g., [16, 17, 18]), the aggregate traffic is characterized by some measured statistics. Measurements are mapped into system parameters to yield the current resource usage. With the given traffic parameters and the prescribed QoS factors, the resource requirement of the new call is computed. Admission decision is then generated by determining whether there is enough spare resource to support the new call. For example, Saito and Shimoto [18] attempt to construct the distribution of the arrival process from traffic measurements. While this approach is completely adaptive to any traffic stream, the overhead to construct the traffic distribution is enormous.

In hybrid analytical/measurement-based approaches [19, 20, 21, 22, 23, 24, 25], there is an underlying assumption of traffic model, either analytical [19, 20] or empirical [24]. Traffic measurements are used to fine tune the parameters of the traffic model to yield the resource allocation results and admission decisions. The hybrid approaches try to strike the best tradeoff between adaptability and implementation overhead.

### 2.1.2 Call-Level Resource Allocation

Call-level resource allocation is concerned with the call handling capacity of the link. It attempts to maximize the channel occupancy. Resource at the call level is the number of available transmission channels, $K_c$. In FDMA, $K_c$ is the number of frequency bands, while in TDMA, $K_c$ is the number of slots during one use of the system. In CDMA, the users all share a common spectral allocation over the time they are active. The capacity requires qualification because it is soft and interference-limited.

From Viterbi and Viterbi [26], the number of users that can be supported per cell satisfies the inequality

$$N \le \frac{W/R}{E_b/I_0} \cdot \frac{1-\eta}{1+f} \tag{2.1}$$

where $\eta = N_0/I_0$, and

$W$ = spread spectrum bandwidth,
$R$ = data rate,
$E_b$ = bit energy,
$I_0$ = maximum total acceptable interference density,
$N_0$ = thermal (background) noise density, and
$f$ = ratio of other interference-to-own cell interference.

If the number of users exceeds $N$, then the total interference will increase beyond the acceptable level $I_0$ and the bit error rate will not be satisfied. We define $K_c = N$ as given by (2.1), in the case of CDMA. The aim of the call-level resource allocation is to utilize the $K_c$ channels as many as possible. Due to user mobility, handoff, a call-level phenomenon, is inevitable. Resource allocation at the call level thus needs to consider the effect of handoff.

A simple, yet effective approach to perform resource allocation at the call level and take into consideration handoff from neighbouring cells is guard channels reservation, e.g., [27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37]. $K_g$ channels are reserved solely for the purpose of handoff. Out of the $K_c$ channels, if more than $K_g$ are unoccupied, both new calls and handoff calls will be accepted. When less than $K_g$ channels are free, only handoff calls will be accepted. New calls will be blocked. Guard channels reservation reduces the forced termination probability significantly by compromising with the new call blocking probability slightly.

Virtual connection tree (VCT) is another approach [38]. A VCT is a cluster, which is a collection of cells in which a mobile user is most likely to be found for the duration of the call. When a mobile user is admitted into the network. resource is reserved in each base station in the VCT to ensure fast handoff by simplifying the handoff admission procedures while keeping handoff blocking probability below some desired level. Application of VCT in multimedia traffic and the optimal cluster size are discussed in [39]. Utilization of VCT is low due to excessive resource reservation. Shadow cluster is thus introduced [40]. It takes into account the direction of the mobile movement and associates a probability of visit for each cell in the cluster. The amount of resource reserved at each base station in the shadow cluster is according to the probability of which the mobile terminal will visit.

A channel carrying technique is proposed in [41] where a set of channels in each cell. sharing common frequency slots as the neighbouring BSs. is dedicated for handoff calls. The transmission channel is thus allowed to *carry* over and it facilitates handoff. at the expense of increased frequency reuse factor and reduced capacity (number of users).

## 2.2 Review on Call Admission Control

Call admission control in wireless mobile communications networks can be performed from either the network layer or from the link layer perspective.

### 2.2.1 CAC in Network Layer

From the network layer perspective, the QoS of interest are PLR, $P_n$ and $P_f$. When there is a new connection request, resource allocation determines the amount of resource required to guarantee the prescribed QoS constraints of both the requesting call and the ongoing calls. The packet-level resource allocation schemes are concerned only with the packet-level QoS factors, namely PLR. The effectiveness of the resource allocation scheme is reflected in the utilization. This utilization can be translated into the capacity region defined by an $M$-dimensional space with intercept points at $X_{u_m}$, $m = 1, 2, \cdots, M$, for the purpose of CAC, where $M$ is the total number of traffic classes. Figure 2.1 shows a capacity region for $M = 2$. After admitting the new call request, if the number of calls, $n_u = (n_{u_1}, n_{u_2}, \cdots, n_{u_M})$, is within the capacity region, then the new call can be admitted. Otherwise, the call request will be rejected. The packet-level resource allocation schemes discussed in Section 2.1.1 can be used for CAC in this fashion.

The call-level resource allocation schemes are concerned only with the call-level QoS factors, namely $P_n$ and $P_f$. They are discussed in Section 2.1.2. The results are already expressed as the number of admissible users and can be used for CAC decisions directly.

There are also CAC approaches which do not use the result of resource allo-

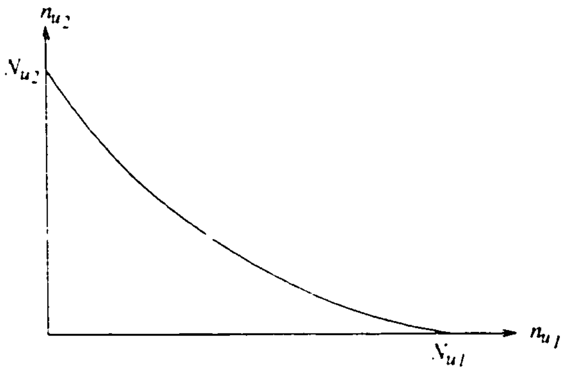

Figure 2.1: Capacity region for $M = 2$

cation, e.g., [42, 43, 44, 45, 46, 47]. For a given number of transmission channels, these approaches aim to satisfy the $P_n$ and $P_f$ requirements. In [42], handoff characteristics are estimated by exchanging state information (number of handoff calls) among neighbouring cells. On the other hand, a forced termination probability index is computed in [43, 44]. The forced termination probability index reflects the likelihood that the new call will be terminated prematurely if admitted. Admission is granted if the index is below a preset threshold value. Given the number of channels, [45, 46, 47] study the admission cutoff threshold. Integrated voice and data traffic is studied in [47] and data traffic may use idle voice channels on a reservation basis, to enhance the utilization. These papers give case studies of the blocking performance for different system settings.

### 2.2.2 CAC in Link Layer

There are CAC schemes aiming at satisfying the link-level QoS constraints, e.g. [48, 49, 50]. As mentioned at the beginning of this chapter, the link-level quality is characterized by the signal-to-interference ratio (SIR). A desired bit error rate performance will give a specification of the required SIR value. This approach is quite common when studying CDMA systems because the capacity is interference-limited. In [48, 50, 49], the interference and the achievable SIR are analyzed. If the anticipated SIR value is below the threshold, then the new call request will be rejected. The admission threshold in [50] is adaptive through traffic measurements. CAC is used in conjunction with power control in [49]. The admissible region is defined by the power allocation requirement of the $M$ traffic classes. If the power allocation vector is within the admissible region, the call request is accepted. These schemes are mainly targeted for homogeneous voice traffic.

Multimedia traffic is considered in [51, 52, 53, 54, 55, 56, 57, 58, 59]. Different power allocations for different traffic classes are considered in [51, 53, 52, 58]. The achieved interference is obtained and compared to the interference threshold to make admission decision. Integrated data and voice are considered in [54, 56, 57, 59]. In [54, 56], data traffic use the left-over capacity of voice. Data traffic can be queued up in [54] while access to the idle voice channels is achieved on a slot-by-slot reservation basis. Admission is also based on the interference analysis. Wu *et al.* allows soft-QoS requirements to enhance the system utilization [59]. In [55], the effective bandwidth from the packet-level resource allocation is used in obtaining the processing gain. The resultant interference is then determined and served as the basis of the admission decision.

## 2.3 Motivation and Goals

### 2.3.1 Effect of Mobility on QoS Provisioning

Network layer CAC needs to satisfy both the packet-level and call-level QoS factors. Packet-level resource allocation approaches guarantee the packet-level QoS factors, namely PLR. The techniques are borrowed heavily from wireline ATM resource allocation. The call-level resource allocation schemes guarantee the call-level QoS factors, namely $P_n$ and $P_f$. The need to satisfy $P_f$ is due to presence of user mobility. All the guard channel techniques improve the $P_f$ performance, by considering certain handoff characteristics. The question of how mobility affects the performance has to date received little attention except in [60, 61, 62].

The effect of user mobility, using a two-state MMPP model, is studied in [60, 61]. The two states are referred to as normal and hot, where the latter causes user aggregation. These two papers are mainly concerned with the effect of the duration of the hot state, which is characterized by relatively low user speed and hence long cell crossing time. In [62], the effect of mobility is represented in a more general way using a mobility parameter. This way of representing mobility information offers some degree of flexibility in terms of studying the impact of user mobility on network-level QoS provisioning undertaken in the present paper. For this reason, in what follows we also employ a general user mobility parameter to incorporate the effects of handoff and call blocking. This thesis first studies the effect of user mobility on the cell utilization and the cell capacity required to maintain the prescribed network-level QoS. This problem is of great interest because it provides a fundamental understanding of the effect of mobility on QoS provisioning.

### 2.3.2 Coupling of the Link and Network Layer Requirements

CAC can be studied from the link layer or network layer point of view. Although the main focus of this thesis is on the network layer CAC, an understanding of the coupling effect of the two layers is important. From the network layer perspective, the capacity is the number of transmission channels, $K_c$. $K_c$ is limited by the tolerable interference. This is particularly crucial in CDMA-based system since the capacity is soft and interference-limited. The capacity of CDMA systems has been studied extensively, e.g., [63, 26, 64, 65, 66, 61, 67, 68]. From the link layer point of view, these papers consider the effect of transmission channel characteristics, such as fading, and imperfect power control, on the bit error performance and the achievable SIR value, $E_b/I_0$. $E_b/I_0$ determines the system capacity, $K_c$. The consideration is in a bottom-up manner. Network layer quality of service (QoS) performance such as new call blocking probability, $P_n$, and forced termination probability, $P_f$, are not studied.

From the network layer point of view in a top-down manner, the effect of mobility on the channel requirement is considered in [69, 62]. The number of channels available in the system, $K_c$, is the measure of capacity in these studies. For the second generation TDMA-based systems such as GSM and IS-136, the capacity (number of channels) is fixed. However, for the 3G wideband CDMA (either IMT-2000 or cdma2000) systems, the number of channels becomes a variable due to its soft capacity characteristics. More importantly, the capacity studies in the link layer and the network layer have been considered in isolation. Nonetheless, the capacity available at the network layer is the same capacity provided by the link layer. Therefore, the capacity can be viewed of a CDMA system as a varying parameter $K_c$, expressed as a function of $E_b/I_0$. In this regard, a linkage between the link layer and the network layer can be provided through $K_c$. This coupling provides the relationship between the link layer specification $E_b/I_0$ and the network layer QoS performance and requirements, in the presence of user mobility. To the best of our knowledge, this coupling has not been studied in the literature.

### 2.3.3 Joint Packet-Level and Call-Level Resource Allocation

Resource allocation either focuses on satisfying the packet-level or call-level QoS factors. Few studies are available for joint consideration of the QoS factors of the two levels, [70, 71, 72, 73]. However, the two levels are considered in isolation, except in [73]. As also noted in [73], the non-interaction between the two levels results in excessive allocation of resource at the packet level. Therefore, it is important to consider a joint satisfaction of the packet-level and call-level QoS factors in resource allocation. A joint-levels resource allocation scheme is thus introduced. It 1) couples the packet-level and call-level QoS parameters for CAC, and 2) captures the non-persistent nature of the traffic flows at the packet level, in order to achieve further statistical multiplexing gain. This piece of work is in a similar spirit as [73] in which the source effective bandwidth developed in [20] is weighted by a factor $\phi \in (0, 1]$, in order to reduce the resource requirement at the packet level. However, no attempt is made to optimize or to determine the best value of $\phi$ that would maximize utilization. The proposed joint-levels resource allocation scheme makes use of statistical multiplexing by directly linking the packet-level and call-level constraints, and taking advantage of the non-persistent nature of the traffic flows, to perform resource allocation to maximize the utilization.

### 2.3.4 Movable Boundary Channel Allocation

To consider heterogeneous traffic types, we grossly divide the traffic into two types, real-time (RT) and non-real-time (NRT). RT traffic will be of an interactive type, e.g., voice, which is delay-sensitive. NRT traffic will be of a distributive (non-interactive) type, e.g., data, which is delay-insensitive, but is loss-sensitive. With a specified cell capacity $K_c$, an efficient scheme to divide the $K_c$ channels among the RT and NRT users is required. A movable boundary channel allocation scheme is proposed. The scheme is a form of virtual partition in the sense that NRT traffic can make use of the unused capacity nominally allocated to RT traffic without affecting the RT service performance. This is achieved by allowing RT users to preempt NRT users when necessary.

The idea of movable boundary for integrated voice/data services in computer communication networks was first introduced by Fischer et al. [74]. In the intervening time, there have been numerous ensuing studies reported in the literature, e.g., [75, 76, 77]. In general, a TDM (Time Division Multiplexing) structure is assumed. In a frame of $K$ slots, $S$ slots are allocated to voice users, thus supporting $S$ voice users. The remaining $K - S$ slots are dedicated to data users. In addition, the $S'$ idle voice slots are also utilized by data users to increase data transmission rate. Data packets are buffered and transmitted via $K - S + S'$ slots in each frame. These work are concerned with resource sharing at the packet level. While suitable for wireline communications such as ATM, they are not completely applicable in the wireless mobile environment, in particular, when we are interested in the satisfaction of call-level QoS, i.e. $P_n$ and $P_f$.

In a wireless communications system supporting voice and data, the capacity, $K$, ($K$ slots for TDMA, or $K$ spreading codes for CDMA) are allocated to voice and data users. The above works use the $S'$ idle slots to increase the data transmission rate. In practice, assigning these idle slots dynamically on a frame-by-frame basis is difficult. Instead, the proposed scheme uses the idle channels to support more data users and thus enhance the system utilization, $N_u$. In other words, the above works deal with resource sharing at the packet level while the proposed scheme is concerned with resource sharing at the call level. To the best of our knowledge, using movable boundary for resource sharing at the call level has not been considered.

In the literature, a number of capacity allocation schemes are reported, e.g., [78, 79, 80, 81]. Virtual partitioning is performed in [78, 79, 80]. Incoming packet priority varies dynamically according to the instantaneous capacity usage and buffer occupancy of each class. Again, these schemes are mainly concerned with multiplexing traffic types in a wireline network. Also, these papers are concerned with enhancing QoS performance. The question of how to engineer the network, such that the set of prescribed QoS parameters are satisfied, is not addressed. In [81], a dynamic channel assignment algorithm by carrier switching is proposed. Data users may obtain their required slots through different carriers. If necessary, a data user may be assigned fewer slots than required during handoff to reduce handoff blocking. This scheme increases resource utilization and reduces blocking. However, carrier switching is difficult to implement in practice. The reason of adopting the movable boundary approach as opposed to virtual partitioning and carrier switching is that it is simpler to implement.

## 2.4 Summary of Objectives

The four objectives of this thesis are summarized as follows.

1. To study the effect of user mobility on QoS provisioning. The cell capacity and cell utilization as a function of the network-level QoS factors and user mobility are quantified. The excess capacity requirement and utilization degradation as a function of handoff rate are given. Using the cell capacity as the common

thread, the linkage between the link-level QoS ($E_b/I_0$) and the network-level QoS ($P_n$ and $P_f$) as a function of handoff rate is obtained. These are studied in Chapter 3.

2. To propose a joint packet-level and call-level resource allocation (RA) scheme. It incorporates the non-persistent flow characteristics in the packet-level analysis. The joint-levels RA problem is formulated as an optimization problem to find the maximum $K_c$, subject to simultaneous satisfaction of the packet-level and call-level QoS constraints. This scheme can be used in wireline networks as well as wireless networks including CDMA-based systems. The joint-levels RA scheme is discussed in Chapter 4.

3. To introduce a movable boundary (MB) channel allocation scheme to support two traffic types: RT and NRT. It assigns the idle voice channels to data traffic on the basis that it may be preempted if voice traffic needs the channels at a later time. The amount of data traffic can be increased. The preemption process gives priority to voice traffic so that data traffic will not affect its performance. The MB channel allocation scheme is studied in Chapter 5

4. To combine the joint-levels RA scheme and the MB channel allocation scheme, and incorporate the findings of the effect of mobility, to structure a mechanism for CAC. This is discussed in Chapter 6.

# Chapter 3

# Effects of Mobility on QoS Provisioning

The objectives of this chapter are two folds. First, the effect of user mobility on system utilization and excess capacity required to maintain a prescribed level of QoS is investigated. Second, the link-level QoS in terms of $E_b/I_0$ and the network-level QoS in terms of $P_n$ and $P_f$ are coupled through the available system capacity.

This chapter offers the following major contributions:

- The cell capacity and cell utilization as a function of the network-level QoS factors and the user mobility parameter are quantified. This knowledge can be used to enforce call admission control.

- Using the cell capacity as the common thread, the linkage between the link-level QoS ($E_b/I_0$) and the network-level QoS ($P_n$ and $P_f$) as a function of user mobility parameter is obtained. For the prescribed network-level QoS requirements, this information quantifies the value of $E_b/I_0$ required as a function of handoff rate. If the $E_b/I_0$ value turns out to be too small so that the bit error rate (BER) at the link-level is not acceptable, then the user and the network provider can use this information to renegotiate network-level QoS specification.
- The degradation in cell utilization as a function of the handoff rate is established. To maintain the zero-handoff-rate utilization as the handoff rate increases, a larger cell capacity is needed. This information can be used to request the radio transmission level to choose a better channel code and/or antenna diversity.

The rest of this chapter is organized as follows. Section 3.1 describes the problem associated with QoS provisioning in a CDMA setting, and the representation of user mobility using a mobility parameter. The behaviour of a radio cell is characterized using a queueing model in Section 3.2. The mathematical characterization of the effect of user mobility and the coupling between the network-level and link-level QoS are provided in Section 3.3. Section 3.4 presents numerical results to demonstrate the effect of user mobility on resource management. Concluding remarks are given in Section 3.5.

## 3.1 Problem Statement

Let $K_c$ be the system capacity measured in terms of the number of channels per cell. In the case of FDMA, $K_c$ is the number of frequency bands, while in the case of TDMA, $K_c$ is the number of slots during one use of the system. In either case, $K_c$ is well-defined, giving rise to a hard capacity. In CDMA, the users all share a common spectral allocation over the time they are active. The quality of service is specified by a prescribed level of signal-to-interference ratio, or $E_b/I_0$. In this respect, the capacity of a cell is soft and requires qualification.

CDMA is interference-limited. The link-level QoS in terms of BER can be specified by a prescribed level of $E_b/I_0$. For example, the minimum required $E_b/I_0$ for BER $< 10^{-3}$ is 7 dB, as specified in [63]. To support multimedia traffic, it is assumed that there are $M$ traffic types, with different traffic rates $R_m$ and QoS requirements $(E_b/I_0)_m$. The minimum required transmit power of the $j^{\text{th}}$ type $m$ source is given by

$$S_{m_j} \geq \frac{R_{m_j}}{W} \cdot \left(\frac{E_b}{I_0}\right)_{m_j} \cdot I_{t_{m_j}} = \Gamma_{m_j} \cdot I_{t_{m_j}}. \tag{3.1}$$

where $I_{t_{m_j}}$ is the maximum tolerable total interference consisting of intracell and intercell interference and background thermal noise, experienced by the $j^{\text{th}}$ type $m$ source. $\Gamma_{m_j}$ and $R_{m_j}$ are the traffic demand and the data rate of the $j^{\text{th}}$ type $m$ source, respectively. This treatment is adopted from [82]. For $K_{c_m}$ users of type $m$,

$$S_m = \sum_{j=1}^{K_{c_m}} S_{m_j} \geq \sum_{j=1}^{K_{c_m}} \Gamma_{m_j} \cdot I_{t_{m_j}}. \tag{3.2}$$

If $\boldsymbol{S} = [S_1, S_2, \cdots, S_M]$ can be found such that inequality (3.2) is met, the required QoS of all users is satisfied. Due to finite transmit power, the values of $K_{c_m}$ are limited. The $M$-dimension hyperplane defined by inequality (3.2) describes the capacity, $K_c = [K_{c_1}, K_{c_2}, \cdots, K_{c_M}]$, of the CDMA system. A detailed treatment of power allocation for wideband CDMA systems is given in [82].

In this thesis, the primary interests are in investigating the effect of mobility on utilization and capacity requirement, and establishing the coupling between the link-level and the network-level QoS. In this regard, a simpler case is treated in this chapter and Chapter 4, where all users are of the same type, i.e., $M = 1$. Two traffic classes, i.e., $M = 2$, is considered in Chapter 5. In the sequel, the subscript $m$ is dropped. For $M = 1$, the maximum number of users that can be supported per cell is as obtained in [26], and it satisfies the inequality (2.1). Imposing equality and setting $K_c = N$, (2.1) becomes

$$K_c \leq \frac{W/R}{E_b/I_0} \cdot \frac{1-\eta}{1+f}. \tag{3.3}$$

Let $N_u$ be the system utilization, defined as the average number of users that can be supported per cell. If each user occupies only one channel, as for cellular voice services, then the system can support at most $K_c$ users at any given time. Due to the random nature of the call arrival process, allowing $N_u = K_c$ users in the system would drive the call blocking probability toward one. Although the blocking probability is always less than unity because the system is finite, if we wish to maintain it at any reasonable level of practical interest, say $10^{-2}$, $N_u$ must be less than $K_c$. Intuitively, when users roam from cell to cell, to maintain the same QoS performance, either the system utilization will decrease or additional capacity is required.

Let $N_s$ and $N_m$ denote, respectively, the system utilization when all mobile users are stationary and when mobile users are in motion. Our first problem is to determine the degradation in utilization, $\Delta N_u = N_s - N_m$, due to user mobility, subject to a prescribed set of QoS specifications. For the problem under consideration, the QoS requirements are the new call blocking probability, $P_n^*$, and the forced termination probability, $P_f^*$.

Our second problem is to determine the additional capacity, $\Delta K_c$, required to maintain the zero-handoff-rate utilization when all users are in motion, i.e., $X_m = X_s$. In other words, it is necessary to increase the system capacity from $K_c$ to $K_c + \Delta K_c$ in order to maintain the same level of QoS. We are interested in assessing the value of the excess capacity, $\Delta K_c$, needed to maintain a given utilization and to guarantee the prescribed QoS factors when users are on the move.

In practice, the number of available channels subject to a prescribed link-level QoS specification is often fixed by the radio transmission method used. Therefore, the first problem is of practical interest while the second provides a motivation to find methods to increase the system capacity, e.g., antenna design and/or channel coding. This is of particular interest to CDMA systems where the capacity is soft. Eq. (2.1) illustrates the soft capacity feature of CDMA. Consider homogeneous traffic, for example, digitized voice of rate $R$. If we assume that both $\eta$ and $f$ in (2.1) can be bounded and treat them as constants, then the capacity, $K_c$, is just a function of $E_b/I_0$. The smaller is the value of $E_b/I_0$, the higher will be the capacity. $E_b/I_0$ specifies the SIR requirement in order to satisfy a certain bit error rate (BER). For example, as mentioned earlier, the minimum required $E_b/I_0$ for BER $< 10^{-3}$ is specified as 7dB in [63]. This takes into account the coding performance and antenna design efficiency. Therefore, with better coding and/or antenna design, the $E_b/I_0$ requirement can be reduced to increase the system capacity.

Note that in defining the capacity in terms of the link-level QoS, (2.1) does not capture mobility information. However, user mobility directly affects the network-level QoS, i.e., $P_n^*$ and $P_f^*$, which are in turn a function of the system utilization $X_u$ and capacity $K_c$. It is noted that new call initiation by different users and handoff call initiation by the system (the mobile switching center, assisted by the mobile and the base station) from other cells are independent random events. In this thesis, new calls and handoff calls are treated alike, and it is assumed that call arrivals obey a Poisson distribution and that the mean service time and mean inter-handoff time are exponentially distributed.

Inter-handoff call arrivals may not obey a Poisson distribution, since handoff initiation depends on both the direction of movement and the speed of the mobile [83, 84]. Gamma distribution and the mixed Erlang distribution have been proposed to model general inter-handoff time. However, in certain situations, the exponential distribution gives good approximations. Figures 3.1 and 3.2 show the simulation results demonstrating the accuracy of the exponential approximation for $K_c = 35$ channels and $K_c = 50$ channels, respectively. Mobile units are traveling inside a $10 \times 10$ square grid with wrap-around edges. Different handoff rates are achieved by choosing different user speed and cell size. For example, the handoff rate for a user traveling at 4 km/h in a picocell of $20 \times 20$ m$^2$ is 3.33 per min., while that for a user traveling at 40 km/h in a macrocell of $1000 \times 1000$ m$^2$ is 0.667 per min.

Four movement directions are possible: *north*, *south*, *east* and *west*. Call duration is assumed exponentially distributed with a mean of 2 minutes. Four different scenarios are simulated and are described as follows.

1. *Constant speed with fixed direction* All mobile units have the same speed, constant throughout the call lifetime. The direction is randomly chosen at call initiation and is fixed until the call ends.

2. *Constant speed with direction change* - All mobile units have the same speed, constant throughout the call lifetime. The direction is randomly chosen at call initiation and it changes every 10 seconds with

   $$P(\text{go straight, turn left, turn right,U-turn}) = (0.7, 0.13, 0.13, 0.04).$$

3. *Uniform speed with fixed direction* - The mobile speed is uniformly distributed (between (40, 60) km/h for cell edge greater than 200 m in length and between (1, 4) km/h for cell edge less than 200 m in length.) It is determined at call initiation and remains constant throughout the call lifetime. The direction is randomly chosen at call initiation and is fixed until the call ends.

4. *Uniform speed with direction change* - The mobile speed is uniformly distributed, as described above. It is determined at call initiation and remains

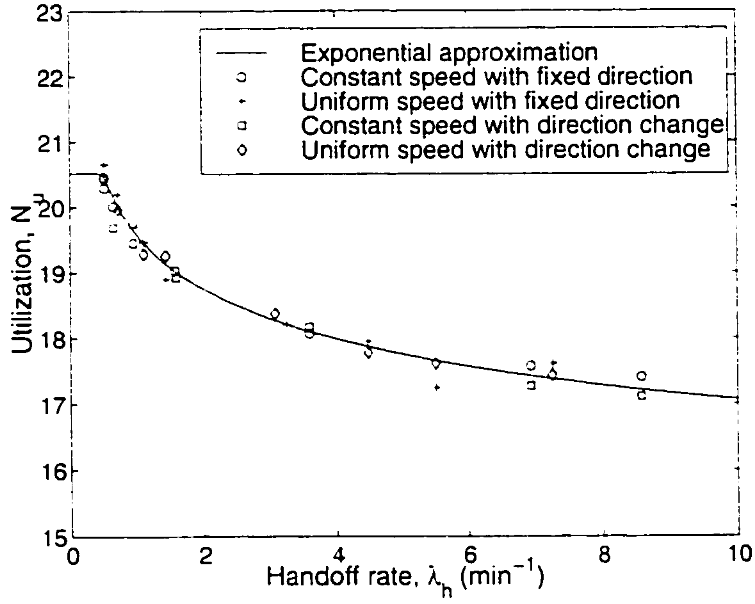

Figure 3.1: Accuracy of the exponential inter-handoff time assumption with $K_c^* = 35$

constant throughout the call lifetime. The direction is randomly chosen at call initiation and it changes every 10 seconds with

$$P(\text{go straight. turn left. turn right.U-turn}) = (0.7. 0.13. 0.13. 0.04).$$

Figures 3.1 and 3.2 show the achievable cell utilization when $P_n^* = P_f^* = 10^{-3}$ are satisfied simultaneously for increasing handoff rates. It is seen that the exponential inter-handoff time gives good approximation to the achievable utilization. Therefore, the exponential approximation is used in this thesis. The analytical results obtained will provide a meaningful relationship between the link-level and

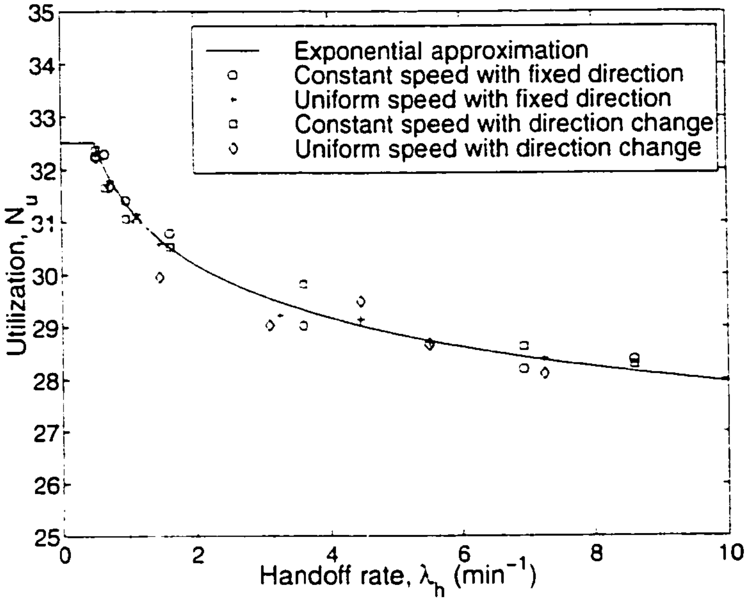

Figure 3.2: Accuracy of the exponential inter-handoff time assumption with $K_c = 50$

network-level QoS.

With the above assumptions, we can model the radio cell as an $M/M/K_c/K_c$ queueing system. In this case, the system blocking probability is given by the Erlang Loss Formula, denoted by

$$P = Er(N_u, K_c) \tag{3.4}$$

where $P$ is a generic call blocking probability at the radio cell, related to $P_n$ and $P_f$.

From the above discussion, we see that $E_b/I_0$ gives the link-level specification in order to support the required BER. At the same time, $K_c$ specifies the requirement to support network-level QoS requirements ($P_n^*$ and $P_f^*$) in the presence of user mobility. $K_c$ thus provides the vertical coupling between the link-level $E_b/I_0$ specification and the network-level QoS performance through (2.1) and (3.4).

Our third problem is to quantify the $E_b/I_0$ requirement to provide the excess capacity in order to support a prescribed level of user movement. $K_c$ is the coupling parameter between the network-level QoS performance and link-level BER requirement.

Note that the excess capacity requirement is referred to as the channel cost of mobility in [62]. We believe that the utilization degradation is a more appropriate quantification of the cost of mobility because this is the price for supporting mobility. An increase in capacity is not always feasible. The work in this chapter of the thesis differs from that in [62] in the following ways. First, the QoS factor is $P_f$ while the QoS factor in [62] is $P_d$, where $P_d = P_n + P_f$ is the call dropping probability, the probability that a call will ever be dropped from the time of connection request to the expected end of connection. $P_n$ and $P_f$ are considered separately here because they are the performance seen directly from the users' perspective. Second, the handoff rate, $\lambda_h$, is used as the mobility parameter rather than $\theta \in (0,1]$ as in [62], which is mapped from $\lambda_h \in (0,\infty)$. This will give more insightful results. Third, the utilization degradation, which is of practical interest, as mentioned be-

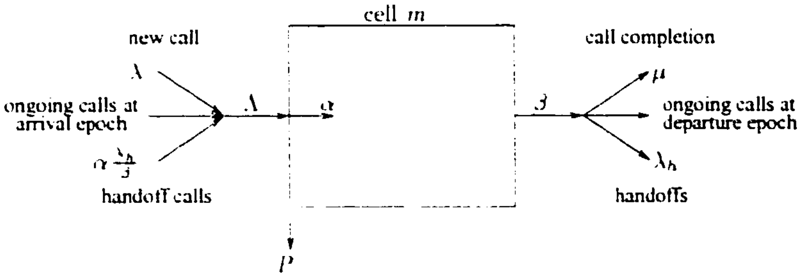

Figure 3.3: Queueing Model for Homogeneous Traffic

fore, is also considered. In addition, the coupling between the network-level QoS performance and the link-level $E_b/I_0$ requirement through the common parameter $K_c$ is provided. To our knowledge, this has not been addressed in the literature.

## 3.2 Performance Analysis

In the preceding section, we alluded to the fact that the behaviour of a radio cell can be modeled using an $M/M/K_c/K_c$ queueing system. The functional block diagram of this queueing model is shown in Figure 3.3. A similar modeling approach was used in [85]. This model encapsulates the parameters that characterize the effects due to user mobility. In a cellular network, it is assumed that every cell exhibits identical behaviour, so that Figure 3.3 models the behaviour of a generic cell.

In the queueing model of Figure 3.3, ongoing calls cut through the system; only new calls and handoff calls may experience blocking. From the blocking probability point of view, only the new calls and handoff calls need to be considered. The parameters that govern the queueing model are as follows.

- $\lambda$ - new call arrival rate.
- $1/\mu$ - call holding time.
- $\lambda_h$ - handoff rate.
- $P$ - generic call blocking probability at the cell.
- $\Lambda$ - aggregate arrival rate of new and handoff calls (offered load).
- $\alpha$ - arrival rate of accepted traffic after blocking (carried load), and
- $\beta$ - service rate of the cell.

The carried load $\alpha$ is related to the offered load $\Lambda$ by

$$\alpha = (1 - P) \cdot \Lambda. \tag{3.5}$$

where the offered load is given by

$$\Lambda = \lambda + \frac{(1 - P)\Lambda}{\beta}\lambda_h.$$

The service rate is

$$\beta = \mu + \lambda_h. \tag{3.6}$$

Solving for $\Lambda$, we obtain

$$\Lambda = \frac{\lambda}{1 - (1 - P)\frac{\lambda_h}{\beta}}. \tag{3.7}$$

From a modeling point of view, we are interested in assessing the performance trend as the handoff rate increases. To this end, a homogeneous traffic environment is considered and handoff and new calls are treated equally likely, i.e., handoff calls are not given higher priority. It is further assumed that each user occupies one channel, as in the cellular voice systems. Let $P$ = Pr(no channel is available), i.e., the probability of blocking in the queueing system. Using (3.5), (3.6) and (3.7), the utilization (or the average number of users), $N_u$, is given by

$$N_u = \frac{\alpha}{\beta} = \frac{(1 - P)\lambda}{\mu + \lambda_h P} \approx \frac{\lambda}{\mu}. \tag{3.8}$$

The approximation in (3.8) for $P \leq 10^{-2}$ is reasonable.

Define the mobility parameter $\theta \in [0, 1)$ as the probability that a departure is due to handoff [62]. Thus,

$$\begin{aligned}\theta &\triangleq \Pr(\text{the departure is due to handoff}) \\ &= \tfrac{\lambda_h}{\mu+\lambda_h}.\end{aligned} \tag{3.9}$$

The closer $\theta$ is to 1, the higher is the mobility. $1-\theta$ is the probability the departure is due to call completion. Recall that the two considered QoS factors are $P_n$ and $P_f$. A new call is blocked when there is no free channel. Therefore, $P_n = P$. The forced termination probability, $P_f$, is the probability that a call is forced to terminate prematurely due to an unsuccessful handoff at some instant during the lifetime of the call. A handoff is unsuccessful if it cannot find a free channel during any handoff attempt in the lifetime of a connection. Therefore, taking into consideration the number of handoffs during the call's lifetime, $P_f$ is given by

$$\begin{aligned}P_f &= (1-P)\theta P + (1-P)^2\theta^2 P + \\ &\quad (1-P)^3\theta^3 P + \cdots \qquad (3.10)\\ &= \frac{\theta P(1-P)}{1-(1-P)\theta}. \qquad (3.11)\end{aligned}$$

Eq. (3.11) is a quadratic function in $P$. Solving for $P$, the admissible root is expressible as a function of $P_f$:

$$P = \frac{\theta(1-P_f) - \sqrt{\theta^2(1-P_f)^2 - 4\theta P_f(1-\theta)}}{2\theta}. \tag{3.12}$$

Therefore, to satisfy the prescribed $P_f^*$,

$$P^* = \frac{\theta(1-P_f^*) - \sqrt{\theta^2(1-P_f^*)^2 - 4\theta P_f^*(1-\theta)}}{2\theta}. \tag{3.13}$$

Eq. (3.13) gives the required $P$ in order to support the prescribed $P_f$. Notice that there are two solutions when solving for $P$ in (3.11). Only (3.13) is picked because the other solution results in a situation which is not of practical interest. The derivation of (3.13) is given in Appendix A. To satisfy the prescribed $P_n^*$,

$$P^* = P_n^*. \tag{3.14}$$

Equations (3.13) and (3.14) impose the requirement of $P^*$ of the cell, in order to satisfy $P_n^*$ and $P_f^*$. From (3.11), it is observed that with $P_n$ being satisfied, there is a certain degree of $P_f$ guarantee. In fact,

$$\frac{\theta P_n^*(1-P_n^*)}{1-(1-P_n^*)\theta} < P_f^*.$$

is possible for certain values of $\theta$. When $\theta \leq \theta_{cr}$, where

$$\theta_{cr} = \frac{P_f^*}{(1 - P_n^*)(P_n^* - P_f^*)}. \tag{3.15}$$

Satisfying $P_n^*$ implies satisfaction of $P_f^*$ as well. Using (3.9), the above condition is equivalent to $\lambda_h \leq \lambda_{h_{cr}}$, where

$$\lambda_{h_{cr}} = \frac{\mu P_f^*}{P_n^*(1 - P_n^* - P_f^*)}. \tag{3.16}$$

Therefore, as far as $\lambda_h$ is concerned, $P^*$ is divided into two regions. When $\lambda_h \leq \lambda_{h_{cr}}$, $P_n^*$ is dominant and the requirement of $P^*$ is given by (3.14). When $\lambda_h > \lambda_{h_{cr}}$, $P_f^*$ is dominant and the requirement of $P^*$ is given by (3.13). The requirement of $P^*$ to satisfy $P_n^*$ and $P_f^*$ can be expressed as

$$P^* = \begin{cases} P_n^* & \text{if } \lambda_h \leq \lambda_{h_{cr}} \\ \frac{\theta(1-P_f^*) - \sqrt{\theta^2(1-P_f^*)^2 - 4\theta P_f^*(1-\theta)}}{2\theta} & \text{if } \lambda_h > \lambda_{h_{cr}}. \end{cases} \tag{3.17}$$

The need to satisfy $P_f^*$ is due to the presence of user roaming. As a result, (3.17) captures the effect of user mobility through $P_f^*$ and $\lambda_h$ and encapsulates them into the $P^*$ requirement of the cell.

Assume that the arrival stream is Poisson distributed with mean rate $\lambda$, and that the mean service time $\mu^{-1}$ and mean handoff time $\lambda_h^{-1}$ are exponentially distributed. We can model the radio cell as an $M/M/K_c/K_c$ queueing system. The system blocking probability is given by the Erlang Loss Formula [86]:

$$\begin{aligned} P = Er(N_u, K_c) &= \frac{N_u^{K_c}/K_c!}{\sum_{j=0}^{K_c} N_u^j/j!} \\ &\approx \frac{N_u^{K_c} e^{-N_u}}{K_c^{K_c} e^{-K_c}\sqrt{2\pi K_c}}. \end{aligned} \tag{3.18}$$

i.e., the probability that all $K_c$ channels are occupied for an average number of users $N_u$. The approximation in (3.18) is based on the Stirling approximation:

$$x! \approx e^{-x} x^x \sqrt{2\pi x},$$

and that the denominator in (3.18) is approximated by the exponential function. Assuming equality and taking the natural logarithm on both sides of (3.18), it reduces to

$$K_c(\ln N_u - \ln K_c) - \frac{1}{2}\ln(2\pi K_c) + K_c - N_u - \ln P = 0. \tag{3.19}$$

Here, the system capacity $K_c$ is expressed in terms of the network-level QoS parameters. Eqs. (3.13), (3.14) and (3.19) give the relationship among mobility profile $\lambda_h$, utilization $N_u$, capacity $K_c$, new call blocking probability $P_n$ and forced termination probability $P_f$. These equations can be solved numerically to obtain the performance characteristics. Here, the Newton-Raphson method is used to obtain numerical results for utilization degradation $\Delta N_u = N_s - N_m$, excess capacity requirement $\Delta K_c = K_m - K_s$, and the $E_b/I_0$ requirement, where

$$E_b/I_0 \triangleq \frac{W/R}{K_c} \cdot \frac{1-\eta}{1+f}. \tag{3.20}$$

is obtained from (3.3). The subscripts $m$ and $s$ are used to represent the parameters when all mobiles are in motion, and stationary, respectively.

## 3.3 Effect of Mobility on System Performance

The presence of user mobility degrades system performance. Specifically, if the network-level QoS were to be preserved, then the number of connections that can be supported would decrease. Otherwise it would be necessary to increase the system capacity by whatever means. In this section, the impact on utilization degradation is first examined. Then, the additional capacity required to maintain the same level of network QoS as user mobility increases is investigated. One way to increase system capacity is to lower the $E_b/I_0$ specification. A relationship between the link-level QoS, the network-level QoS and handoff rate is derived. In this way, specifications of blocking probability, forced termination probability and handoff rate yields the value of $E_b/I_0$ necessary for supporting the prescribed network-level QoS and user mobility information.

### 3.3.1 Utilization Degradation

To study the impact on utilization degradation due to user mobility, the number of transmission channels, $K_c$, is assumed fixed. To accommodate a new call blocking probability of $P_n^* < 1$, it is required that $N_u < K_c$. In the case where all users are immobile, the system utilization is $N_u = N_s$. When users are in motion, the utilization, $N_u$, reduces from $N_s$ to $N_m$ in order to maintain the QoS requirement. Clearly, $K_c > N_s > N_m$.

Let $P_s^*$ and $P_m^*$ denote the system blocking probabilities when all users are stationary and in motion, respectively, so that the user prescribed QoS is satisfied. When there is no mobility, $P_s^* = P_n^*$. In the presence of mobility, to guarantee a forced termination probability of $P_f^*$, the utilization degradation is obtained by substituting $N_u = K_c - \Delta N_u$ in (3.19):

$$K_c[\ln(K_c - \Delta N_u) - \ln K_c] - \frac{1}{2}\ln(2\pi K_c) + \Delta N_u - \ln P_m^* = 0, \qquad (3.21)$$

where $P_m^*$ is given by (3.13). $\Delta N_u$ in (3.21) can be solved using the Newton-Raphson method.

### 3.3.2 Excess Capacity Requirement

When considering the excess capacity requirement, the utilization is held fixed. Let $K_s$ denote the number of channels required to guarantee $P_n^*$ when all users are stationary. When mobility is introduced, in order to maintain the same utilization (number of users), $N_u$, additional channels (capacity) are required.

Notice from (3.8), since the value of $P$ can vary, $\lambda$ and $\mu$ need to be varied appropriately to maintain a constant $N_u$. However, from the network's perspective, $\lambda$ and $\mu$ are the user traffic parameters that should not be changed. Consequently, as $P$ changes the value of $N_u$ will *not* be preserved. However, since $P$ is small, the value of $N_u$ remains approximately the same even with user mobility. In other words, $N_m \approx N_s \approx \lambda/\mu$. The excess capacity requirement is then obtained by substituting $K_c = K_m = K_s + \Delta K_c$ in (3.19):

$$(K_s + \Delta K_c)[\ln N_u - \ln(K_s + \Delta K_c)] - \frac{1}{2}\ln[2\pi(K_s + \Delta K_c)] + \\ (K_s + \Delta K_c) - N_u - \ln P_m^* = 0. \quad (3.22)$$

$\Delta K_c$ can then be solved using the Newton-Raphson method.

### 3.3.3 $E_b/I_0$ Requirement

$E_b/I_0$ is a link-level QoS and the system capacity as a function of $E_b/I_0$ is given by (2.1). Consider the capacity of an isolated cell. In this case, $f = 0$. Assuming equality, (2.1) becomes

$$K_c = \frac{k}{E_b/I_0}. \tag{3.23}$$

where we have substituted $K_c$ for $N$ and let $k = (W/R)(1 - \eta)$.

**Remark:** $K_c$ in (3.19) and (3.23) represents the same system capacity of a single cell.

Substituting $K_c$ as given by (3.23) into (3.19) yields

$$\frac{k}{E_b/I_0}[1 + \ln N_u - \ln k + \ln(E_b/I_0)] + \frac{1}{2}\ln(E_b/I_0) - \frac{1}{2}\ln(2\pi k) - N_u - \ln P^* = 0. \tag{3.24}$$

Eq. (3.24) gives the relationship among the link-level SIR specification $E_b/I_0$, the network-level QoS parameters $P_n^*$ and $P_f^*$ (through (3.13) and (3.14)), the mobility information $\lambda_h$, and the system utilization $N_u$. Given the network-level specifications, the user handoff rate, the system bandwidth, and the user transmission rate, Eq. (3.24) can be solved for $E_b/I_0$ using the Newton-Raphson method.

## 3.4 Numerical Results and Discussions

This section computes and presents the results of utilization degradation, excess capacity required to maintain the same network-level QoS, and the $E_b/I_0$ requirement as user mobility changes. The values of the parameters used in the numerical evaluations are:

- call duration $\mu^{-1} = 2$ mins..
- System bandwidth $W = 12.5$ MHz.
- Data rate $R = 16.2$ kbps.
- New call blocking probability $P_n^* = 10^{-3}$.
- Forced termination probability $P_f^* = 10^{-6}$.

With $\eta = 0.1$ and assuming a nominal $E_b/I_0$ value of 7 dB [63], the number of channels, $K_c$, is 139, by imposing equality in (2.1). The handoff rate $\lambda_h$ is varied between 0 and 10 per min. This covers basically all the scenarios from pedestrian users to users traveling on a highway (see Appendix B). $\lambda_h = 0$ is the case when all users are stationary.

### 3.4.1 Utilization Degradation

Figures 3.4 to 3.6 (the solid lines corresponding to the left vertical axes) show respectively the system utilization $N_u$, the utilization degradation ($\Delta N_u = N_s - N_u$)

and the percentage degradation ($\Delta N_u/N_s \times 100\%$) as user mobility $\lambda_h$ increases for a given number of transmission channels, $K_c$, subject to satisfaction of the $P_f^*$ specification.

As expected, with no increase in system capacity, the utilization reduces as mobility increases. Utilization also decreases for more stringent $P_f^*$ requirements. Inspection of the figures shows that with the indicated parameter values, the utilization, $N_u$, is 110 out of the 139 channels, in order to support a $P_n^*$ of $10^{-3}$, even when all users are stationary ($\lambda_h = 0$). When mobility is introduced, the utilization is further reduced. This shows that it is necessary to provide *standby resources* to support a $P_n^* < 1$. Supporting user mobility requires excess standby resources.

Note that when $P_f^* = 10^{-3}$, $N_u$ and $\Delta N_u$ remain constant until $\lambda_h$ exceeds a certain threshold (at the left end of the graphs). This is the result of (3.17), where the $P^*$ requirement of the cell is divided at $\lambda_h = \lambda_{h_{cr}}$ into two regions. Note that when $\lambda_h \leq \lambda_{h_{cr}}$, satisfaction of $P_n^*$ imples the satisfaction of $P_f^*$. Therefore, no excess capacity is required to maintain the utilization. Since $K_c$ is fixed, utilization is not degrading.

It is also observed that a small change in mobility when mobility is low (small $\lambda_h$) causes a larger utilization degradation than when mobility is high. This is manifested in Figure 3.5 where the $\Delta N_u$ curves are the steepest for small $\lambda_h$ and level off with increasing $\lambda_h$. This suggests that a change in mobility has a greater effect in a macrocell than in a microcell or a picocell. It is because user speed in a macrocell is larger and handoff rates tend to be larger. This effect is even more pronounced for more stringent $P_f^*$ requirements.

Figures 3.7 to 3.9 show the utilization for different number of available channels, $K_c$, as a function of the mobility parameter $\lambda_h$, with $P_f^*$ as a parameter. It is observed that utilization degradation increases with increasing system resources, $K_c$ (Figure 3.8). However, the percentage utilization degradation actually *decreases* with increasing system resources, $K_c$ (Figure 3.9). Therefore, for a cell with *more* resources, i.e., larger number of transmission channels, it requires a smaller percentage of standby resources to support mobility. This suggests a benefit for macrocellular environment. For picocellular environment, $K_c$ is typically much smaller. Therefore, it requires more overhead to support mobility.

### 3.4.2 Excess Capacity Requirements

The excess capacity requirement results are shown in Figures 3.4 to 3.9 (the dashed lines corresponding to the right vertical axes). The observations are similar to those of utilization degradation in the preceding subsection. From Figures 3.4 to 3.6, it can be seen that excess capacity requirement increases for increasing mobility, $\lambda_h$. It is also observed that the excess capacity requirement is larger for more stringent

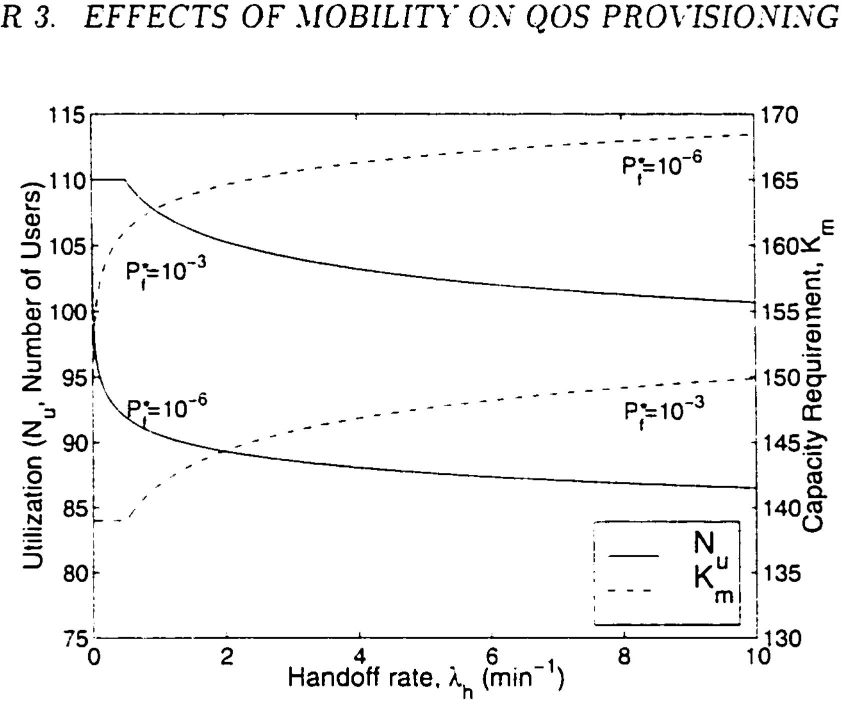

Figure 3.4: Utilization and Capacity Requirement as a function of mobility ($K_c$ = 139, $\mu^{-1}$ = 2 mins, $P_n^*$ = $10^{-3}$)

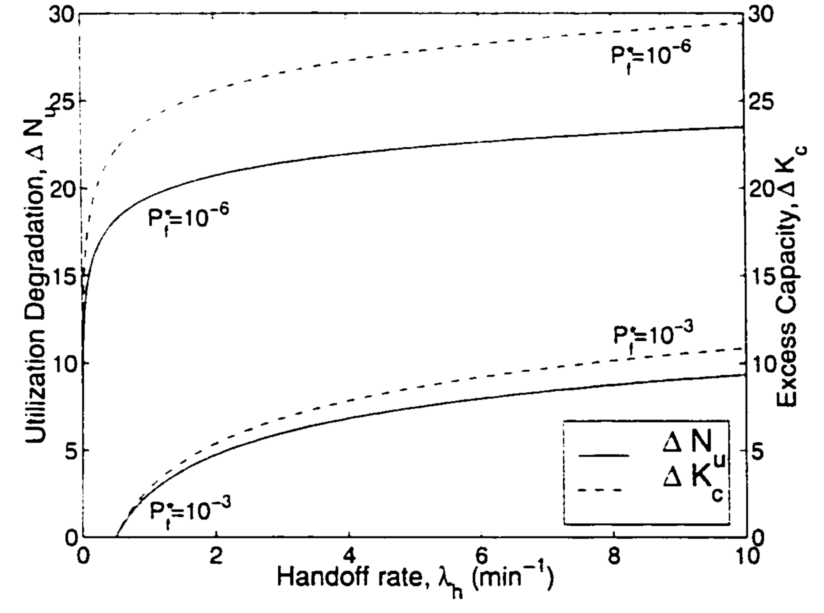

Figure 3.5: Utilization degradation and excess capacity requirement as a function of mobility ($K_c$ = 139, $\mu^{-1}$ = 2 mins, $P_n^*$ = $10^{-3}$)

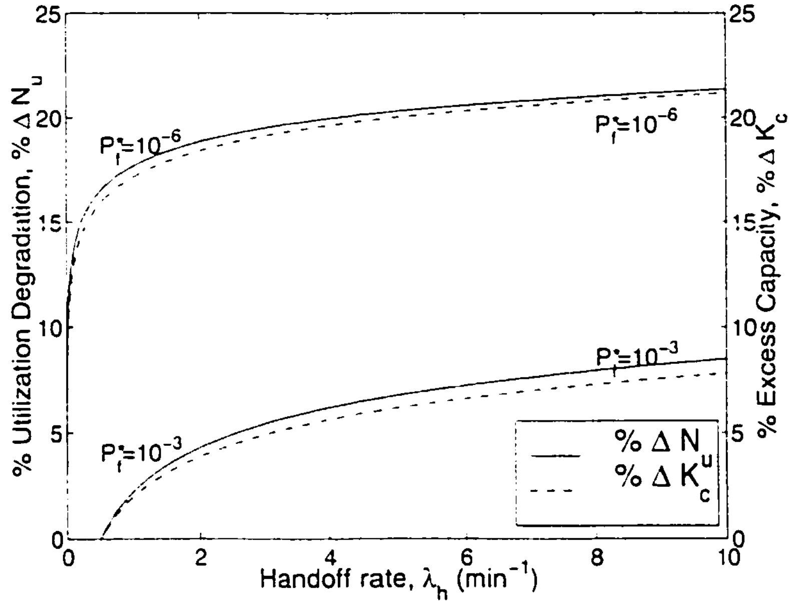

Figure 3.6: Percentage utilization degradation and excess capacity requirement as a function of mobility ( $K_c^{*}$ = 139, $\mu^{-1}$ = 2 mins, $P_n^{*}$ = $10^{-3}$)

$P_f^{*}$, and that, when $\lambda_h < \lambda_{h_{cr}}$, where $\lambda_{h_{cr}}$ is given by (3.16), no excess capacity is required.

From Figure 3.5, it is seen that a small change in mobility when $\lambda_h$ is small results in a larger excess capacity requirement than when $\lambda_h$ is large. In fact, the $\Delta K_c^{*}$ curves are the steepest for small $\lambda_h$ and level off with increasing $\lambda_h$. As with utilization degradation, this suggests that a change in user mobility has a greater effect on excess capacity requirement in a macrocell than in a microcell or picocell. These observations are not made in [62] because of the use of $\theta$ as the mobility parameter by Foschini *et al.*

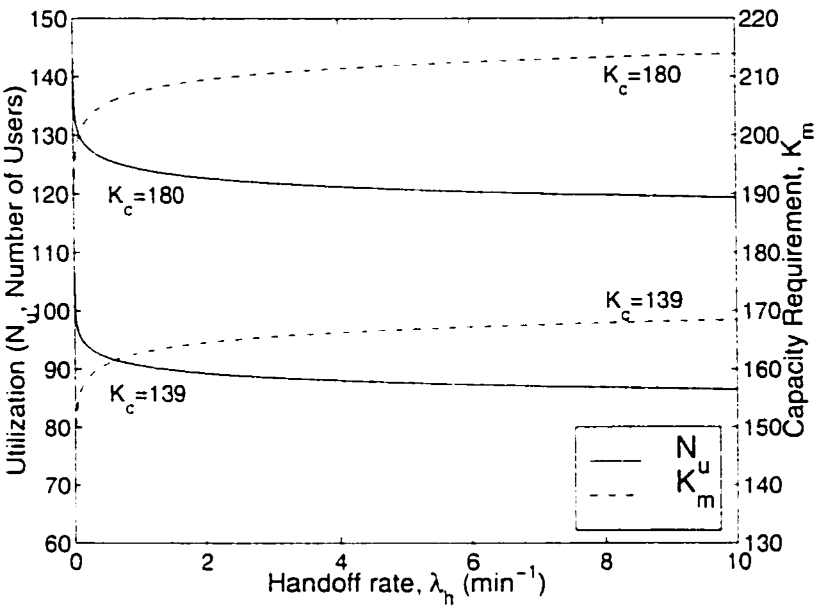

Figure 3.7: Utilization and capacity requirement as a function of mobility ($\mu^{-1} = 2$ mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$)

Figures 3.7 to 3.9 show the excess capacity requirement for different number of available channels, $K_c$. It is observed that excess capacity requirement increases with increasing system resources, $K_c$. However, the percentage of excess capacity required actually *decreases* with increasing $K_c$. This is in agreement with the observations in utilization degradation. This suggests that it requires more overhead to support mobility in a cell with a smaller amount of resources. Since a macrocell usually has more transmission channels than a microcell or picocell, this means a macrocell requires a smaller percentage of excess capacity to support mobility.

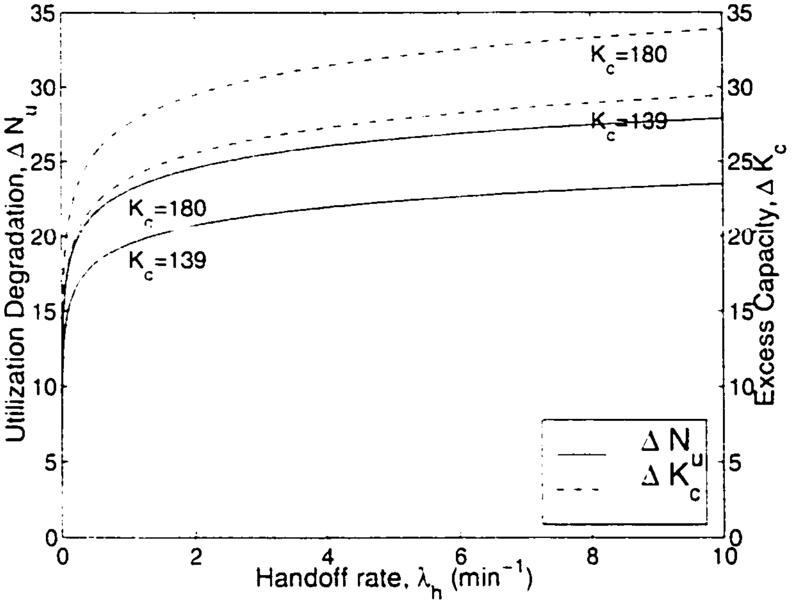

Figure 3.8: Utilization degradation and excess capacity requirement as a function of mobility ($\mu^{-1}$ = 2 mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$)

### 3.4.3 $E_b/I_0$ Requirements

Figures 3.10 and 3.11 show the $E_b/I_0$ requirement as mobility increases. From (2.1), it is seen that $E_b/I_0$ and $K_c$ are inversely proportional. If more users are to be supported (by providing more channels, i.e., by increasing $K_c$), then it requires a smaller $E_b/I_0$. In the studies here, if mobility is introduced, to maintain the same utilization, $K_c$ needs to be increased by decreasing the $E_b/I_0$ specification. This is illustrated in Figure 3.10, where 139 and 180 channels are initially provided when the users are all stationary, in order to support $P_n^* \leq 10^{-3}$. With user mobility, to maintain the same utilization, smaller $E_b/I_0$ values are required. From Figure 3.11,

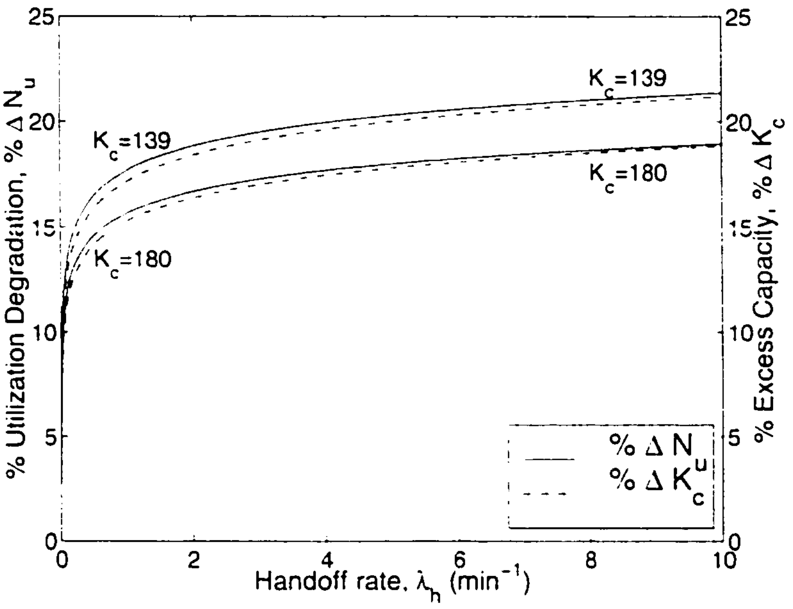

Figure 3.9: Percentage utilization degradation and excess capacity requirement as a function of mobility ($\mu^{-1}$ = 2 mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$)

it is observed that small $P_f^*$ specifications will require the link level to operate at small $E_b/I_0$ specifications. In general, the larger is the capacity requirement, the lower will be the $E_b/I_0$ specification. This provides a benchmark for manipulation of the link-level functions such as channel coding and power allocation.

From Figure 3.4, the utilization when all users are stationary is 110. Consider the case when $\lambda_h$ = 2 per min. In order to maintain $N_u$ = 110, the capacity needs to be increased from 139 to 168. This corresponds to decreasing the $E_b/I_0$ specification from 7 dB to 6.15 dB. In practice, this may not be achievable, e.g., when both channel coding and antenna diversity have been frozen. Figure 3.12

illustrates such a tradeoff between $E_b/I_0$ and $N_u$ for different values of $\lambda_h$. The curves exhibit a convex behaviour. Note that when $\lambda_h \to 0$, $N_u \to 110$ at 7 dB. This is the case when all users are stationary and thus no excess capacity is required, so that the nominal value of $E_b/I_0 = 7$ dB is sufficient. With improved channel coding such that $E_b/I_0 < 7$ dB is implemented, Figure 3.12 gives the expected utilization, subject to satisfaction of network-level QoS, namely $P_n^*$ and $P_f^*$. For example, for $\lambda_h = 2$ per min., if $E_b/I_0 = 6.15$ dB in order to maintain $N_u = 110$ is too costly, perhaps choosing a coding scheme which gives $E_b/I_0 = 6.47$ dB to allow $N_u = 100$, i.e., a utilization degradation of 10, is more desirable. The vertical coupling between the network-level performance ($P_n^*$ and $P_f^*$) and the link-level requirement $E_b/I_0$ through $K_c$ provides a valuable linkage between the network-level and link-level design.

## 3.5 Conclusions

A queueing model for analyzing and assessing the performance of a radio cell is described. Numerical results which show the impact of mobility on (i) utilization degradation, and (ii) excess capacity requirement in a wireless network to maintain a prescribed level of QoS are presented. The link-level $E_b/I_0$ specification needed to support user mobility is also given. This investigation provides a good

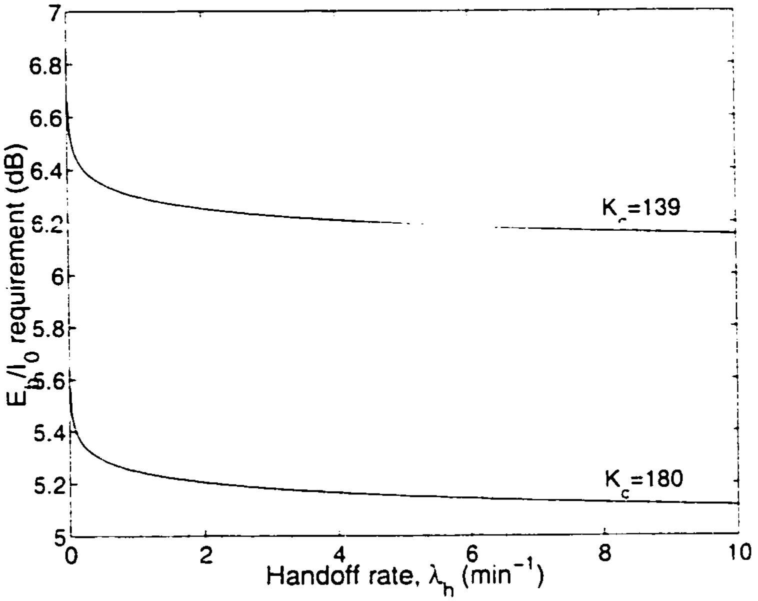

Figure 3.10: $E_b/I_0$ requirement in dB as a function of mobility ($\mu^{-1}$ = 2 mins. $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$)

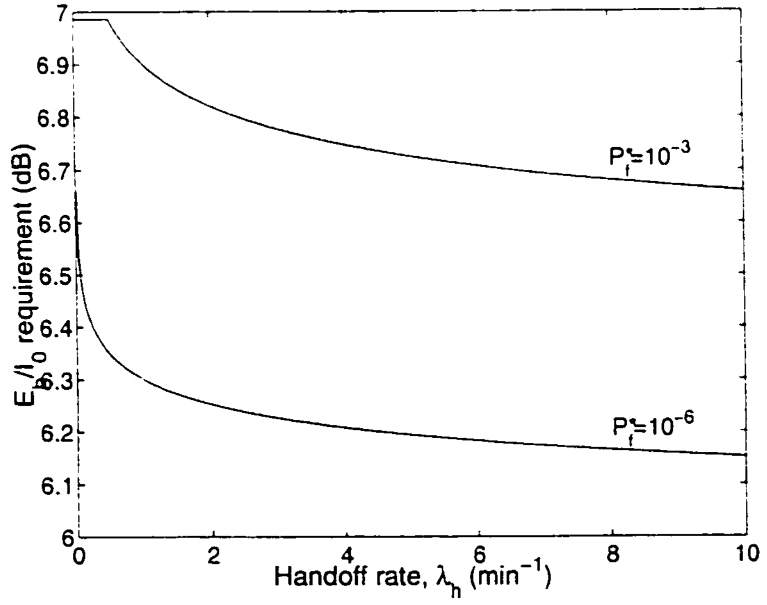

Figure 3.11: $E_b/I_0$ requirement in dB as a function of mobility ($K_c = 139$. $\mu^{-1} = 2$ mins. $P_n^* = 10^{-3}$)

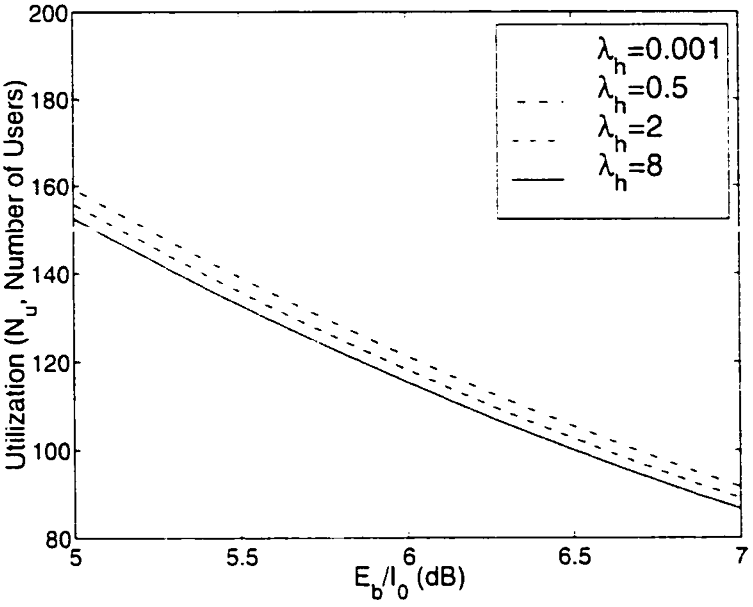

Figure 3.12: $E_b/I_0$ and capacity tradeoff ($K_c = 139$, $\mu^{-1} = 2$ mins, $P_f^* = 10^{-6}$, $P_n^* = 10^{-3}$)

understanding of the impact of mobility on QoS provisioning and hence resource management. An interesting result is the coupling made between the network-level QoS performance and the link-level $E_b/I_0$ requirement. The tradeoff between $E_b/I_0$ specification and the system utilization offers a unique linkage between the link and network-level QoS provisioning.

# Chapter 4

# Joint-Levels Resource Allocation

This chapter studies the proposed joint packet-level and call-level resource allocation method. referred to as the joint-levels RA scheme. It 1) couples the packet-level and call-level QoS parameters for CAC. and 2) captures the non-persistent nature of the traffic flows at the packet level. in order to achieve further statistical multiplexing gain.

The rest of this chapter is organized as follows. Section 4.1 discusses the resource allocation problem. In particular. subsection 4.1.1 considers the RA problem at the packet level: subsection 4.1.2 treats the RA problem at the call level; and subsection 4.1.3 couples the packet-level and call-level QoS factors and studies the joint-levels RA. The joint packet-level and call-level QoS-based RA problem in a cellular CDMA environment is investigated in Section 4.2. Finally. concluding remarks are given in Section 4.3.

## 4.1 The Resource Allocation (RA) Problem

In this section, the packet-level and call-level resource allocation problems are examined in turn. Then, the combined call-level and packet-level RA problem is studied. The following notations are used.

$C$ = the physical capacity of the system defined as the number of channels available (corresponding to peak rate allocation with no statistical multiplexing).

$K$ = the maximum number of users that can be handled by the BS taking into consideration the achievable statistical multiplexing and *only* the packet-level QoS constraint.

$K_c$ = the nominal capacity defined as the number of virtual channels.

$N_u$ = the channel occupancy (or utilization) in Erlangs subject to satisfaction of the constraints $P_n \leq P_n^*$. $P_f \leq P_f^*$ and PLR $\leq$ PLR$^*$. The superscript * denotes QoS specification.

### 4.1.1 Packet-Level RA

The packet-level resource allocation attempts to maximize the system utilization through statistical multiplexing of the bursty traffic sources. This is equivalent to finding the maximum number of traffic flows $K^*$ that the system can support, subject to satisfaction of the QoS requirements. In order to demonstrate the viability of the joint-levels RA scheme, voice is used as the test traffic. Only PLR $\leq$ PLR$^*$ is considered. The delay and delay variation requirements are assumed satisfied by limiting the buffer size.

A voice source is modeled by a two-state on-off process, as shown in Figure 4.1. The mean on period is $\alpha_s^{-1} = 1$ sec and the mean off period is $\eta_s^{-1} = 1.5$ sec, and they are assumed exponentially distributed. The voice activity factor is given by $\beta_s = \eta_s/(\eta_s + \alpha_s) = 0.4$. Suppose the outgoing link from the BS is a T1 link of 1.5 Mbps. Voice is sampled at a rate of 8000 samples per second. With 8 bits per sample, a voice source is generating 170 ATM packets[1] per second. With peak rate allocation, the T1 link can support $C = 20.8$ voice sources. Because the source activity factor is less than unity, statistical multiplexing gain will allow sharing of channels by the ongoing calls. The packet-level RA takes the source activity factor to perform statistical multiplexing to find the actual capacity $K^*$, subject to

[1] The 53-bytes ATM cells are referred to as ATM packets to avoid confusion with radio cells.

satisfaction of PLR $\leq$ PLR$^*$.

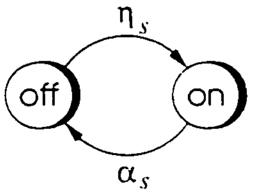

Figure 4.1: Two-state on-off voice model

To assess the PLR, the service at the packet level is modeled by a single-server queue with infinite waiting room. The PLR can then be estimated by the tail distribution of the buffer occupancy $l$ when it exceeds a prescribed buffer size $B$, i.e., $P(l > B)$. The equilibrium buffer occupancy is described by a set of differential equations, as described in [15, 87]. It is found that for $B = 1000$ packets, $K^* = 35$ in order to satisfy PLR $\leq$ PLR$^* = 10^{-6}$.

### 4.1.2 Call-level RA

The call-level resource allocation determines the maximum utilization $N_u$ that the system with $K_c^*$ channels can achieve, with satisfaction of $P_n \leq P_n^*$. Assume that the call arrival process is Poisson, and the call interarrival time and call holding time are exponentially distributed with mean $\lambda^{-1}$ and $\mu^{-1}$, respectively. Then, the BS can be modeled by an $M/M/K_c^*/K_c^*$ queueing system. Given that there are $K_c^*$ channels, the average channel occupancy $N_u$ subject to $P_n \leq P_n^*$ is governed by the

Erlang Loss Formula from (3.18), which is repeated here:

$$P = Er(\Lambda_u, K_c) \leq P_n^*.$$

Note that $\lambda^{-1}$ and $\mu^{-1}$ characterize the non-persistent behaviour of the traffic flows. $\Lambda_u = \frac{\lambda}{\mu}(1 - P_n^*) \approx \frac{\lambda}{\mu}$. The approximation is reasonable for practical values of $P_n^*$ which is usually less than $10^{-2}$. When packet-level and call-level resource allocation are performed in isolation, the flow non-persistent nature cannot be taken into consideration. We can only use $K$ obtained from the packet-level analysis to approximate $K_c$. With our voice test sources, $K_c = K = 35$. $\Lambda_u = 20.52$ Erlangs by solving (3.18) numerically. Let $\rho_c$ be the call-level utilization factor defined by

$$\rho_c \triangleq \frac{\Lambda_u}{K_c}. \tag{4.1}$$

In this case, $\rho_c = 58.62\%$. This, together with $\Lambda_u = 20.52$ Erlangs, will serve as the call-level performance benchmark.

Note that only $P_n$ is considered here. When mobility is taken into consideration, the probability of forced termination, $P_f$, is another QoS parameter. To facilitate the illustration of the joint packet-level and call-level RA, the treatment of mobility is deferred until Section 4.2.

### 4.1.3 Joint Packet and Call Level Resource Allocation (RA)

The joint packet and call levels RA problem is to maximize the channel occupancy (or utilization) $N_u$ subject to satisfaction of $P_n \leq P_n^*$ and PLR $\leq$ PLR$^*$ by finding the maximum allowable $K_c$. It is noted that if $K_c$ is maximized, then $N_u$ will be maximized. This can be formulated as the following constraint optimization problem:

$$\begin{aligned} &\max K_c \\ &\text{s.t. } \mathrm{PLR} \leq \mathrm{PLR}^*. \end{aligned} \tag{4.2}$$

where

$$\mathrm{PLR} = \sum_{i=0}^{K_c} P_i \cdot \varepsilon_i. \tag{4.3}$$

The $P_i$'s are the steady state probabilities of the M/M/$K_c$/$K_c$ queueing model at the call-level. They are obtained by solving the birth-death process, with state transition diagram shown in Figure 4.2. The $\varepsilon_i$'s are the marginal packet loss rate when there are $i$ ongoing traffic flows. They are obtained using fluid flow analysis as described in [15, 87]. The strategy is to express PLR as a sum of the marginal packet loss rates, $\varepsilon_i$'s, weighted by their corresponding state probabilities, $P_i$'s. This weighted sum incorporates the non-persistent nature of the traffic flows in the packet-level analysis. From (4.3). it is observed that the packet-level QoS performance is governed by the $\varepsilon_i$'s. and is coupled with the non-persistent behaviour of the traffic flows described by the $P_i$'s and $K_c$. $K_c$ then restricts the allowable $N_u$ and $\rho_c$. Figure 4.2 depicts the idea of joint-levels RA.

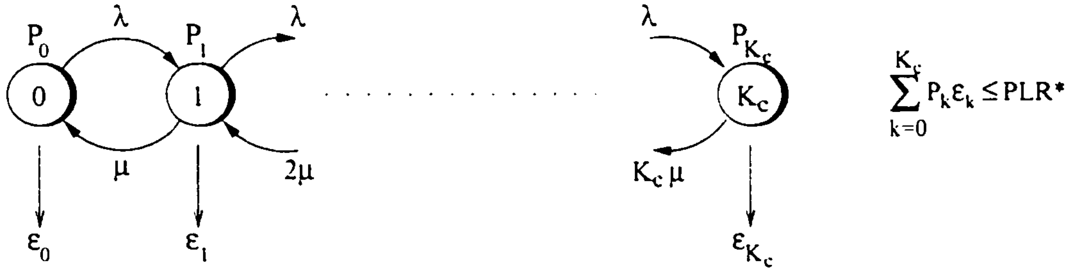

Figure 4.2: Illustration of the joint-levels RA as a queueing model

Equations (4.2) and (4.3) together constitute a nonlinear optimization problem which we solve by enumeration. That is. the problem is solved by fixing $P_n^*$ and PLR$^*$ and obtain a graph of the achievable packet loss rate as a function of $K_c$. From this graph. the largest value of $K_c$ such that PLR $\leq$ PLR$^*$ is determined. The corresponding value of $N_u$ is then obtained by solving (3.18) numerically or by solving

$$K_c(\ln N_u - \ln K_c) - \frac{1}{2}\ln(2\pi K_c) + K_c - N_u - \ln P_n^* = 0 \qquad (4.4)$$

numerically. Eq. (4.4) is an approximation of (3.18) [69].

Without joint-levels RA. $K_c$ is approximated by $K = 35$. It is found from (4.3)

Table 4.1: Maximum value of $K_c$, $N_u$, and $\rho_c$ when PLR$^*$ = $10^{-6}$ and $P_n^* = 10^{-3}$

| | $K_c$ | $N_u$ (Erlangs) | $\rho_c$ |
|---|---|---|---|
| Without Joint-Levels RA | 35.00 | 20.52 | 58.62% |
| With Joint-Levels RA | 40.54 | 24.88 | 61.35% |

that the actually achieved PLR is $7.41 \times 10^{-10} \ll 10^{-6}$. Therefore, further statistical multiplexing is possible to re-gain the loss in utilization due to overallocation of resource at the packet level.

Figure 4.3 shows the achievable packet loss rate as a function of $K_c$ when $P_n^* = 10^{-3}$, obtained by solving (4.2) and (4.3). It is compared to the case when joint-levels RA is not used (using only the packet-level analysis and approximate $K_c$ by $K$). From Figure 4.3 we can deduce that the value of $K_c$ can be increased from 35 to 40.544 while satisfying $P_n^* = 10^{-3}$ and PLR$^*$ = $10^{-6}$ simultaneously[2]. $N_u$ is increased from 20.52 Erlangs to 24.88 Erlangs and $\rho_c$ is increased from 58.62% to 61.35%. The results are summarized in Table 4.1. The improvement in $\rho_c$ is modest. However, the capacity that is re-claimed via further statistical multiplexing at the packet-level results in an increase in $N_u$ by 4.36 Erlangs, a 21% improvement. This translates into higher amount of multiplexing and higher revenues.

So far, the effect of user mobility has not been considered. A salient feature of

[2] By enumeration, $K_c$ is an integer. Non-integral value of $K_c$ is obtained by interpolation assuming the curve is a straight line around the optimal value of $K_c$.

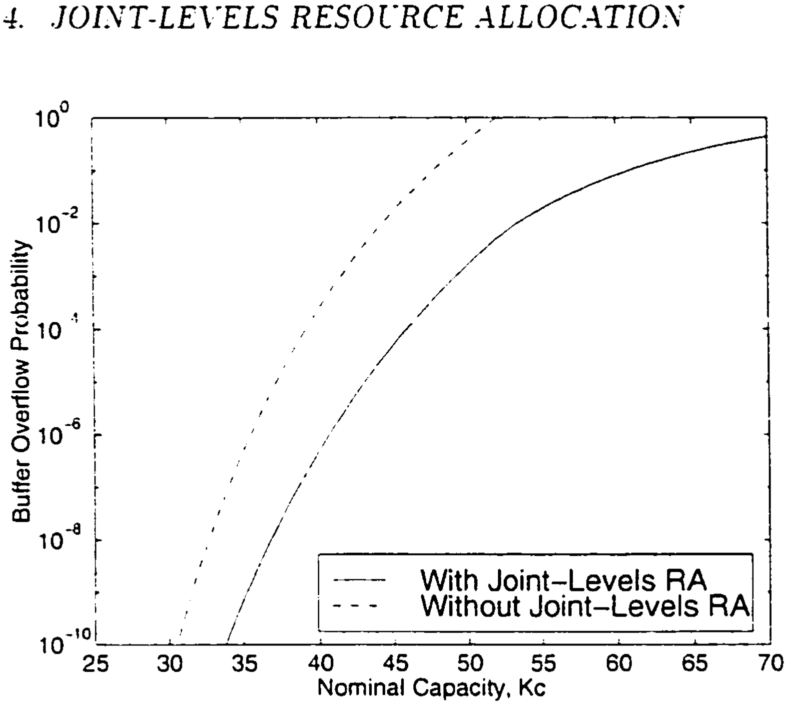

Figure 4.3: Achievable PLR using joint-levels RA ($P_n^* = 10^{-3}$)

cellular communications is the ability to support user roaming. Joint-levels RA in the presence of user mobility is considered in the next section.

## 4.2 Joint-Levels RA in a Cellular Environment

It is shown in Chapter 3 that the presence of mobility gives rise to the requirement of $P_f^*$. To satisfy both $P_n^*$ and $P_f^*$, the required $P^*$ at the BS is given by (3.17), repeated here:

$$P^* = \begin{cases} P_n^* & \text{if } \lambda_h \le \lambda_{h_{cr}} \\ \frac{\theta(1 \;\; P_f^*) \;\; \sqrt{\theta^2(1 \;\; P_f^*)^2 \;\; 4\theta P_f^*(1 \;\; \theta)}}{2\theta} & \text{if } \lambda_h > \lambda_{h_{cr}}. \end{cases}$$

In this chapter, $\mu^{-1} = 2$ mins, $P_n^* = 10^{-3}$ and $P_f^* = 10^{-6}$ are considered. For the cellular environment considered in this section, $P^*$ replaces $P_n^*$ in (4.4) to obtain $N_u$. Notice that $P^* \le P_n^*$ with equality when $\lambda_h \le \lambda_{h_{cr}}$. Therefore, the effect of mobility is encapsulated in the more stringent $P^*$. $P^*$ in turn affects the capacity requirement and the achievable call-level utilization. In general, from Chapter 3, it is found that in the presence of mobility,

1. if the value of $N_u$ at zero-mobility is to be maintained, excess physical transmission capacity is required, or
2. if the physical transmission capacity is kept constant, the utilization $N_u$ will decrease.

The transmission capacity of CDMA systems is interference-limited. In the wireless transmission media, the time-varying interference results in the variation of physical transmission capacity, $C$. $K_c$ will then be varied, causing a variation in $N_u$. For simplicity, assume that $C$ takes on $J$ values: $(C_1, C_2, \cdots, C_J)$. $P_C =$

$(P_{C_1}, P_{C_2}, \cdots, P_{C_J})$ is the capacity distribution characterizing the wireless channel, where $P_{C_j}$ is the probability that the capacity is $C_j$, for $j = 1, 2, \cdots, J$. The optimization problem for the CDMA environment is given by:

$$\begin{aligned} &\max K_c \\ &\text{s.t. PLR} \leq \text{PLR}^*, \end{aligned} \tag{4.5}$$

where

$$\text{PLR} = \sum_{j=1}^{J} P_{C_j} \sum_{i=0}^{K_c} P_i \cdot \varepsilon_{ij}. \tag{4.6}$$

$P_i$'s are the probabilities that there are $i$ admitted connections and $\varepsilon_{ij}$'s are the marginal packet loss rates when there are $i$ admitted connections and $C = C_j$. They are obtained as described in subsections 4.1.1 and 4.1.2. Notice that the objective function (4.5) for the varying capacity environment is the same as (4.2) for the wireline case, except that PLR is modified to (4.6). PLR is the sum of the marginal $\text{PLR}_j$ when the physical capacity is $C_j$, weighted by its probability of occurrence $P_{C_j}$, i.e.,

$$\text{PLR}_j = \sum_{i=0}^{K_c} P_i \cdot \varepsilon_{ij}. \tag{4.7}$$

The effect of $C_j$ is captured in $\varepsilon_{ij}$. The effect of the time varying interference is embraced in the $P_{C_j}$'s and the call-level traffic behaviour is encapsulated in the $P_i$'s in (4.6). Note that it is crucial to obtain the $P_{C_j}$'s, the characterization of the varying channel capacity. In what follows, we assume the existence of the capacity distribution, but make no attempt to derive it.

For illustration, $J = 4$ is assumed. For $c = 1.5$ Mbps, the physical capacity, without taking the voice activity factor into account, is $C = 20.8$. In the considered CDMA environment, it is assumed that $J = 4$, i.e., $C = (17.8, 18.8, 19.8, 20.8)$ (equivalent to $c = (1.28, 1.36, 1.41, 1.50)$ in Mbps). $C_4 = 20.8$ corresponds to the case when interference is minimal and there is no capacity degradation. $C_1 = 17.8$ corresponds to the case when interference is large, resulting in a large capacity degradation. Three different characterization of the wireless channel are considered:

$$\begin{aligned} P_C^{(1)} &= (0, 0, 0, 1) \\ P_C^{(2)} &= (0.02, 0.05, 0.13, 0.8) \\ P_C^{(3)} &= (0.25, 0.25, 0.25, 0.25). \end{aligned} \tag{4.8}$$

$P_C^{(1)}$ represents the case when there is no physical capacity variation, which corresponds to a TDMA environment. $P_C^{(3)}$ corresponds to the case when the physical capacity variation is severe. Figure 4.4 shows the achievable packet loss rate as a

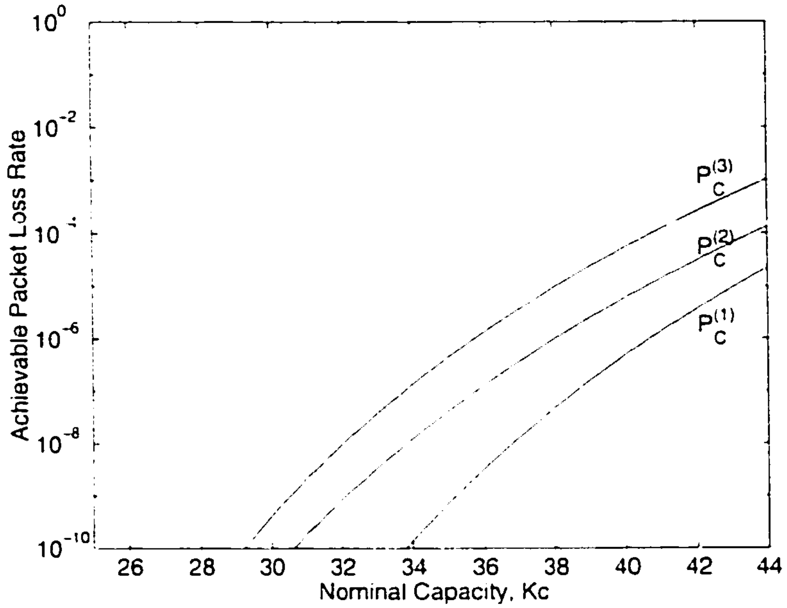

Figure 4.4: Achievable PLR using joint-levels RA in CDMA environment ($P_n^* = 10^{-3}$, $\lambda_h = 0$)

function of nominal capacity $K_c$ for these three wireless environments when $\lambda_h = 0$, i.e., when the users are immobile. It is observed that for a given PLR*, the physical capacity variation causes a reduction in $K_c$.

Figures 4.5 to 4.7 show the achievable $\rho_c$ in the three wireless environments, as a function of $\lambda_h$. It is observed that the utilization factor $\rho_c$ degrades for increasing mobility. With the use of joint-levels RA, $\rho_c$ is increased due to the increase in the allowable $K_c$ and achievable $N_u$. Since larger capacity variation reduces the allowable $K_c$, $\rho_c$ for $P_C^{(3)}$ is the lowest among the three wireless environments.

The performance of the joint-levels RA for the three wireless environments ($P_C^{(1)}$

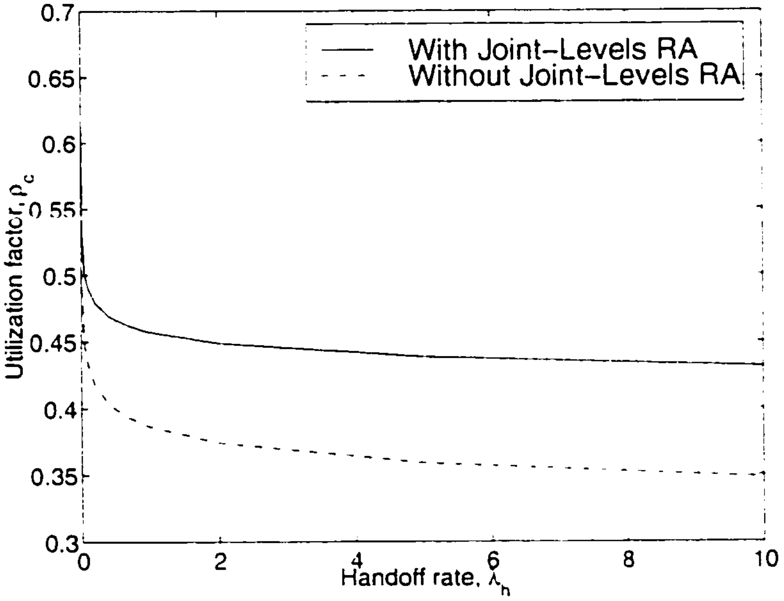

Figure 4.5: Call-level utilization factor in CDMA environment $P_C^{(1)}$ (PLR* = $10^{-6}$, $P_n^* = 10^{-3}$, and $P_f^* = 10^{-6}$)

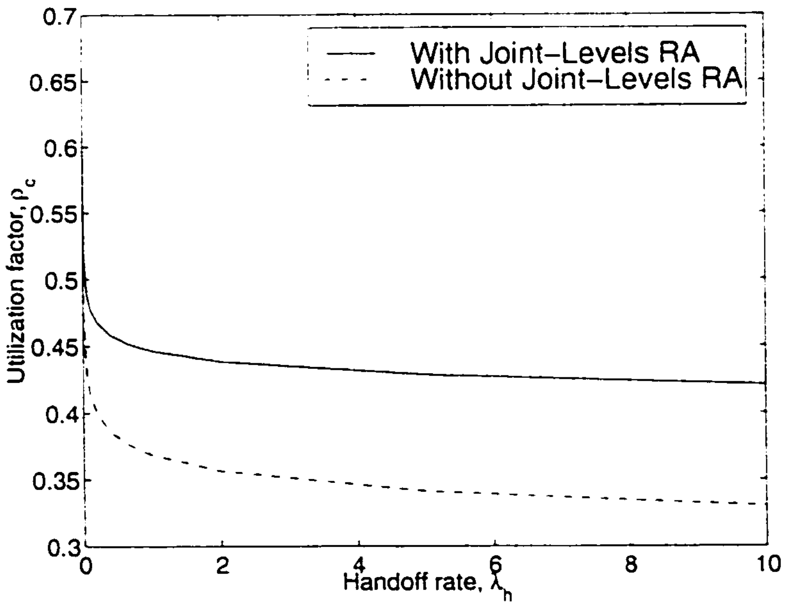

Figure 4.6: Call-level utilization factor in CDMA environment $P_C^{(2)}$ (PLR* = $10^{-6}$, $P_n^* = 10^{-3}$ and $P_f^* = 10^{-6}$)

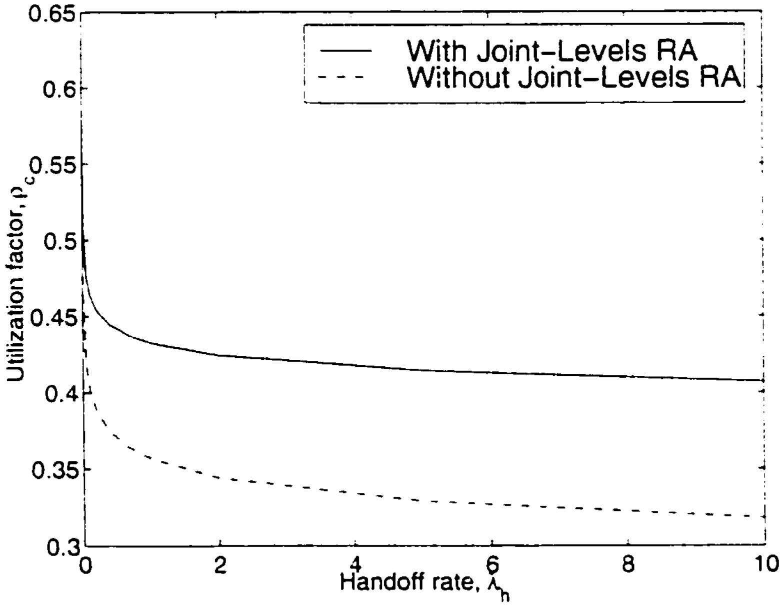

Figure 4.7: Call-level utilization factor in CDMA environment $P_C^{(3)}$ (PLR$^*$ = $10^{-6}$, $P_n^* = 10^{-3}$ and $P_f^* = 10^{-6}$)

Table 4.2: Performance in CDMA environment $P_C^{(1)}$ with PLR$^* = 10^{-6}$, $P_n^* = 10^{-3}$, and $P_f^* = 10^{-6}$, from $\lambda_h = 0$ to $\lambda_h = 10$

| J-RA | $K_c$ | $N_u$ (Erlangs) | $\rho_c$ |
|---|---|---|---|
| Without | 35.31 → 35.31 | 20.75 → 12.19 | 58.76% → 34.52% |
| With | 40.54 → 51.15 | 24.88 → 22.06 | 61.37% → 43.13% |

Table 4.3: Performance in CDMA environment $P_C^{(2)}$ with PLR$^* = 10^{-6}$, $P_n^* = 10^{-3}$, and $P_f^* = 10^{-6}$, from $\lambda_h = 0$ to $\lambda_h = 10$

| J-RA | $K_c$ | $N_u$ (Erlangs) | $\rho_c$ |
|---|---|---|---|
| Without | 32.31 → 32.31 | 18.44 → 10.67 | 57.07% → 33.02% |
| With | 37.99 → 48.73 | 22.85 → 20.51 | 60.15% → 42.09% |

to $P_C^{(3)}$) are summaried in Tables 4.2 to 4.4. The degradation in performance as $\lambda_h$ is increased from 0 to 10 (from immobile case to high mobility case) is shown. The improvement by using joint-levels RA is evident.

By comparing Tables 4.2 to 4.4, it is seen that the effect of physical capacity variation is a reduction of $N_u$ from $P_C^{(1)}$ to $P_C^{(3)}$, at any $\lambda_h$, due to the decrease in the allowable $K_c$. With joint-levels RA, the allowable $K_c$ is increased. As a result,

Table 4.4: Performance in CDMA environment $P_C^{(3)}$ with PLR$^* = 10^{-6}$, $P_n^* = 10^{-3}$, and $P_f^* = 10^{-6}$, from $\lambda_h = 0$ to $\lambda_h = 10$

| J-RA | $K_c$ | $N_u$ (Erlangs) | $\rho_c$ |
|---|---|---|---|
| Without | 30.61 → 30.61 | 17.14 → 9.74 | 56.00% → 31.82% |
| With | 35.61 → 45.68 | 20.99 → 18.59 | 58.94% → 40.69% |

$N_u$ is also increased.

At $\lambda_h = 0$, $N_u$ is increased from 20.75 Erlangs to 24.88 Erlangs in the $P_C^{(1)}$ environment. This corresponds to an increase of 4.13 Erlangs, or a 19.90% improvement. The percentage improvement is also increased with capacity variation. In $P_C^{(2)}$, there is a 23.92% increase (4.41 Erlangs) in $N_u$, while in $P_C^{(3)}$, there is a 22.46% increase (3.85 Erlangs).

The improvement by the joint-levels RA is even more prominent at higher user mobility. At $\lambda_h = 10$, $N_u$ is increased from 12.19 Erlangs to 22.06 Erlangs in the $P_C^{(1)}$ environment. This corresponds to an increase of 9.87 Erlangs, or a 80.96% improvement. The percentage improvement is also increased with capacity variation. In $P_C^{(2)}$, there is a 92.22% increase (9.84 Erlangs) in $N_u$, while in $P_C^{(3)}$, there is a 90.86% increase (8.85 Erlangs).

Significant gain is achieved. The improvement is more noticeable when there is physical capacity variation. Notice that the joint-levels RA increases the allowable nominal capacity $K_c$ without invoking excess transmission capacity, e.g., by improved antenna design and channel coding.

Joint-levels RA is especially beneficial in CDMA. In TDMA ($P_C^{(1)}$), the increase in $K_c$ due to joint-levels RA entails simultaneous time slot assignment to multiple users. Implementation can be difficult due to the complexity in time synchronization of different users. The implementation complexity of joint-level RA in TDMA may outweigh its benefits. In CDMA, joint-levels RA is readily applicable because simultaneous access of the wireless media of multiple users is implied. The complex time synchronization is not required. As a result, joint-levels RA is a great asset to CDMA systems.

## 4.3 Conclusions

In this chapter, the proposed resource allocation method based on joint packet/call-level QoS constraints is studied. The method capitalizes on the extra statistical multiplexing gain achievable at the packet level when incorporating the non-persistent flow behaviour in the packet-level analysis. Numerical results indicate that the proposed method yields a significant gain in utilization in terms of Erlangs at the call level. The joint packet/call-level RA approach is applicable in both the wire-line environment, such as ATM networks, and the wireless environment with user mobility. In TDMA, the complexity in the implementation of time slot assignment to multiple users may overshadow the benefits of the joint-levels RA. However, in CDMA-based systems, the improvement in utilization in the presence of physical capacity variation and its readiness in application make the joint packet/call-level RA method especially beneficial.

# Chapter 5

# Channel Allocation for Two Traffic Types

In this chapter, the proposed movable boundary resource (channel) allocation scheme for handling two traffic types, RT and NRT, at the base station, is studied. The scheme is a form of virtual partition in the sense that NRT traffic can make use of the unused capacity (channels) nominally allocated to RT traffic without affecting the RT service performance. This is achieved by allowing RT users to preempt NRT users when necessary. The QoS parameters of interest are the new call blocking probability, $P_n^*$ and the forced termination probability, $P_f^*$. The goal is to increase the system utilization, defined as the number of users, $N_u$, that the system can support subject to satisfaction of $P_n^*$ and $P_f^*$.

## 5.1 System Model

The total system capacity is partitioned into equal rate transmission channels. A transmission channel can be considered as the basic capacity unit. capable of supporting one RT user. which is presumably voice of 64 kbps. An RT user occupies one channel while a NRT user takes up $r \geq 1$ channels. $r$ is constant among all NRT users. Consider a total of $K_c$ channels where $K_{\mathrm{RT}}$ are allocated to RT traffic and $K_{\mathrm{NRT}} = K_c - K_{\mathrm{RT}}$ are allocated to NRT traffic. The aggregate arrival rate $\lambda$ consists of the RT arrival. $\lambda_{\mathrm{RT}}$. and NRT arrival. $\lambda_{\mathrm{NRT}}$. Let $p_{\mathrm{NRT}}$ be the probability that the arrival is NRT. Then. $\lambda_{\mathrm{RT}} = \lambda(1 - p_{\mathrm{NRT}})$. and $\lambda_{\mathrm{NRT}} = \lambda \cdot p_{\mathrm{NRT}} \cdot r$. They are assumed independent. The call holding times are $\mu_{\mathrm{NRT}}^{-1} = \mu_{\mathrm{RT}}^{-1} = \mu^{-1}$. To be consistent with the previous chapters. it is assumed that the arrival rate is Poisson distributed with mean $\lambda$ and the mean call holding time $\mu^{-1}$ and mean handoff time $\lambda_h^{-1}$ are exponentially distributed.

If the partition between the RT allocation and the NRT allocation is hard. it in effect consists of two independent queueing systems and the performance analysis of each can be treated as in Chapter 3. This is simply complete partitioning and the RT and NRT performance are guaranteed. However. it is inefficient because utilization is low. and is thus undersirable. in particular in the wireless environment where the spectral resource is limited. The other extreme is complete sharing where utilization is high by compromising on performance guarantee.

The movable boundary channel allocation (MB) scheme is proposed to take advantage of the delay-insensitive nature of NRT traffic to increase the system efficiency. The underlying assumption of the MB scheme is that NRT traffic is non-delay sensitive and if being preempted, upper layer protocol will take care of the re-transmissions, so that the loss requirement is met.

Figure 5.1 depicts the model of the proposed MB scheme. When the NRT allocation is fully occupied and the RT allocation is underutilized, a new NRT call can *borrow* an idle RT channel. That is, the NRT traffic is allowed to be *overutilized* when the RT traffic is *underutilized*. The NRT users using the borrowed channels will be preempted whenever the RT traffic requires the borrowed channels back. The MB scheme is illustrated in Figure 5.1. The RT/NRT allocation boundary can move above but not below the nominal allocation boundary (the dotted line).

The performance metric consists of the RT blocking probability, $P_{\mathrm{RT}}$, NRT blocking probability, $P_{\mathrm{NRT}}$, NRT preemption probability $P_p$, and NRT delay $D$. Blocking occurs when a newly arrived call finds there are no free channels. The QoS parameters of interest are the new call blocking probability $P_n^*$ and forced termination probability $P_f^*$. They are assumed to be the same for both the RT and NRT traffic.

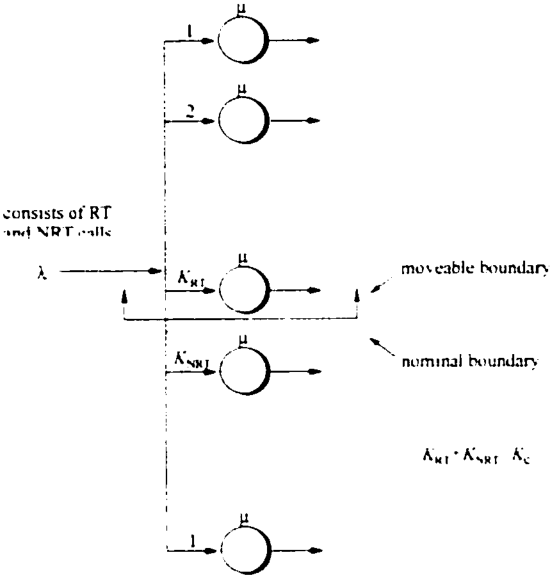

Figure 5.1: RT and NRT channel allocation with a movable boundary

## 5.2 Performance Analysis

As far as the RT traffic is concerned, it is simply an $M/M/K_{RT}/K_{RT}$ system. The performance analysis of such a system is well-known and can be found in, e.g., [86]. The blocking performance of the RT partition is given by the Erlang Loss Formula:

$$P_{RT} = Er(N_{u_{RT}}, K_{RT}) = \frac{N_{u_{RT}}^{K_{RT}}/K_{RT}!}{\sum_{j=0}^{K_{RT}} N_{u_{RT}}^{j}/j!}. \tag{5.1}$$

$N_{u_{RT}}$ is the utilization, i.e., the average number of occupied channels, of the RT partition. Strictly speaking, the first argument of the Erlang Loss Formula in (5.1) should be the traffic intensity $\rho_{\mathrm{RT}}$, given by

$$\rho_{\mathrm{RT}} = \lambda_{u_{\mathrm{RT}}}(1 - P_{\mathrm{RT}}). \tag{5.2}$$

On the basis that $P_{\mathrm{RT}} \ll 1$ in order to satisfy $P_n^* = P_f^* = 10^{-3}$, $\rho_{\mathrm{RT}} \approx \lambda_{u_{\mathrm{RT}}}$ is reasonable. From Chapter 3, it is found that the presence of user mobility entails a more stringent blocking requirement $P^*$ at the base station. In this case, it is required that $P_{\mathrm{RT}} \leq P_{\mathrm{RT}}^*$, where $P_{\mathrm{RT}}^*$ is given by (3.17). The rest of this chapter is devoted to the queueing behaviour of the NRT traffic. The case of $r = 1$ will be considered first, followed by the more general case of $r \geq 1$.

### 5.2.1 Case of $r = 1$

For $r = 1$, a simple one-dimension birth-death process can be used to describe the queueing behaviour of the NRT traffic. It can be described by the birth-death process as shown in Figure 5.2. This birth-death process is dependent on the behaviour of the RT process. Note that $P(RT \leq x)$ denotes the probability that the number of on-going RT traffic flows is less than or equal to $x$.

From flow balancing in the state transition diagram of Figure 5.2, the state

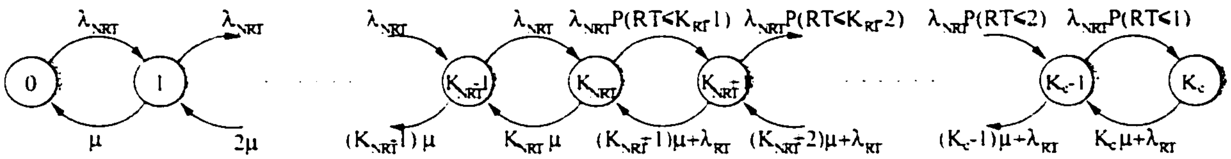

Figure 5.2: NRT Birth-Death Process

probabilities are given by

$$P_j = \frac{\lambda_{NRT}^j}{j!\mu_{NRT}^j} P_0 \qquad j = 1, 2, \cdots, K_{NRT}. \tag{5.3}$$

and

$$P_{K_{NRT}+i} = \frac{\lambda_{NRT}^{K_{NRT}+i} \prod_{j=1}^{i} P(RT \le K_{RT} - 1)}{\prod_{j=1}^{i} [(K_{NRT} + i)\mu_{NRT} + \lambda_{RT}]\, \mu_{NRT}^{K_{NRT}} K_{NRT}!} P_0 \qquad i = 1, 2, \cdots, \sigma. \tag{5.4}$$

where $\sigma$ is the maximum number of channels that can be borrowed from the RT partition. It is obvious that $\sigma \le K_{NRT}$. The performance will be measured in terms of the call blocking probability, $P_{NRT}$, call preemption probability, $P_p$, and delay $D$. $P_{NRT}$ is defined as the fraction of incoming call requests that are rejected due to insufficient channels. $P_p$ is the fraction of the admitted NRT calls that are being preempted to make room for RT traffic. $P_{\mathrm{NRT}}$ and $P_p$ are given as follows:

$$P_{\mathrm{NRT}} = \sum_{j=0}^{\sigma-1} P_{K_{\mathrm{NRT}}+j} \frac{P(\mathrm{RT} = K_{\mathrm{NRT}} - j)}{P(\mathrm{RT} \leq K_{\mathrm{NRT}} - j)}. \tag{5.5}$$

$$P_p = \frac{\sum_{i=K_{\mathrm{NRT}}+1}^{K_{\mathrm{NRT}}+\sigma} P_i \cdot \lambda_{\mathrm{RT}} \cdot \frac{P(\mathrm{RT}=K_c-i)}{P(\mathrm{RT} \leq K_c-i)}}{\sum_{i=0}^{K_{\mathrm{NRT}}+\sigma} P_i \cdot \left[ i\mu_{\mathrm{NRT}} + \frac{P(\mathrm{RT}=K_c-i)}{P(\mathrm{RT} \leq K_c-i)} \lambda_{\mathrm{RT}} \right]}. \tag{5.6}$$

Note that $K_c = K_{\mathrm{RT}} + K_{\mathrm{NRT}}$. Therefore, $P_{\mathrm{NRT}}$ and $P_p$ depend on the RT queueing system and $K_{\mathrm{RT}}$. It is required that $P_{\mathrm{NRT}} \leq P^*_{\mathrm{NRT}}$, where $P^*_{\mathrm{NRT}}$ is given by (3.17).

### 5.2.2 Case of $r \geq 1$

For the general case of $r \geq 1$, a two dimensional state transition diagram is considered, as shown in Figure 5.3. The state is a 2-tuple $(n_{\mathrm{RT}}, n_{\mathrm{NRT}})$, the number of on-going RT and NRT calls, respectively. The (0,0) state is the case where the system is empty. The state migrates to the right when there is an RT arrival, and migrates down when there is an NRT arrival. Conversely, the state moves to the left when an RT call is completed and moves up when an NRT call is completed. The RT allocation can accommodate at most $n_1 \leq N_1 = K_{\mathrm{RT}}$ calls simultaneously. $N_1$ thus constitutes the right limit in Figure 5.3. Similarly, the NRT allocation can accommodate at most $n_2 \leq N_2 = \lfloor K_{\mathrm{NRT}}/r \rfloor$ calls, where $\lfloor x \rfloor$ is defined as the largest integer less than or equal to $x$ (the floor function of $x$). Since NRT traffic may borrow up to $\sigma$ idle RT channels, $n_2$ may exceed $N_2$. In particular, $0 \leq n_2 \leq \lfloor (K_{\mathrm{NRT}} + \sigma)/r \rfloor$. The states in the shaded region occur when NRT traffic borrows RT channels. Figure 5.3 assumes $N_2$ and $\sigma$ are integral multiples of $r$.

$(N_1, N_2)$ corresponds to a fully utilized system when there is no borrowing. When NRT traffic borrows RT channels, a fully utilized system consists of the right-most state in each row of the shaded region. In this case, when there is an RT arrival, it will randomly preempt an NRT user currently borrowing RT channels. This corresponds to the diagonal transition where one NRT user is given up to make room for a newly arrived RT user. Therefore, preemption occurs in the right most state in each row in the shaded region. Blocking of the RT traffic occurs at the right-most states in the unshaded region. Blocking of the NRT traffic occurs at the last $r$ states in each row in the shaded region and the last row in the unshaded region (the row immediate above the shaded region), where there are no free channels to accommodate the incoming NRT call.

The state probabilities can be solved, and thus $P_{\mathrm{NRT}}$ and $P_p$ can be obtained.

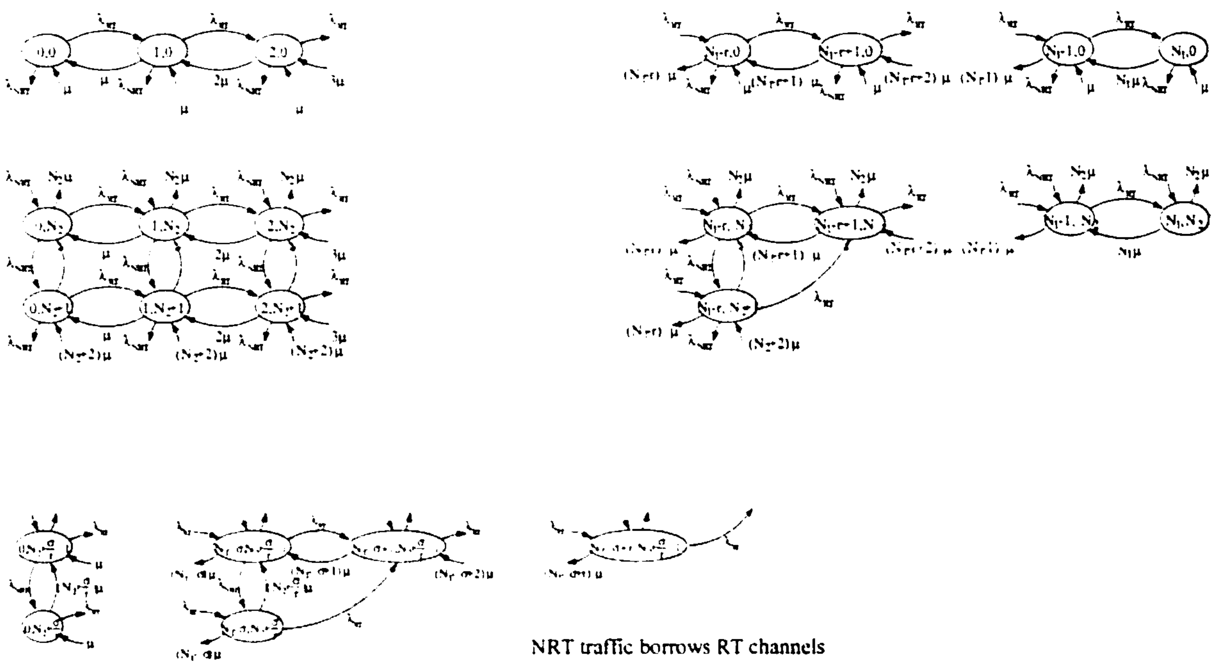

Figure 5.3: NRT Birth-Death Process for $r \geq 1$

In particular,

$$P_{\mathrm{RT}} = \sum_{i=1}^{N_2} P(N_1, i), \tag{5.7}$$

$$P_{\mathrm{NRT}} = \sum_{k=0}^{r-1} \sum_{j=0}^{\lfloor (K_{\mathrm{NRT}}+\sigma)/r \rfloor} P(N_1 - j \cdot r - k, N_2 + j), \tag{5.8}$$

$$P_p = \sum_{j=0}^{\lfloor (K_{\mathrm{NRT}}+\sigma)/r \rfloor} P(N_1 - j \cdot r, N_2 + j). \tag{5.9}$$

Due to state explosion problem, analytic values are only feasible for small systems, e.g., $K_c = K_{\mathrm{RT}} + K_{\mathrm{NRT}} < 100$. Therefore, computer simulations are used to obtain numerical results.

### 5.2.3 Delay, $D$

The NRT delay, $D$, is given by

$$D = \mu_{\mathrm{NRT}}^{-1} + E(N_p)E(a)E(\tau). \tag{5.10}$$

where $\mu_{\mathrm{NRT}}^{-1}$ is the call holding time. $E(N_p)$ is the expected number of preemptions before the call is completed. $E(a)$ is the expected number of attempts before a successful connection re-establishment after being preempted. $E(\tau)$ is the mean time between re-establishment retry attempts which depends on the particular retry algorithm. Note that the excess delay due to the MB scheme, $\Delta D = E(N_p)E(a)E(\tau)$. $E(N_p)$ and $E(a)$ are given by the following.

$$E(N_p) = \frac{P_p}{(1 - P_p)^2}. \tag{5.11}$$

$$E(a) = \frac{1}{1 - P_{\mathrm{NRT}}}. \tag{5.12}$$

Substituting (5.11) and (5.12) in (5.10) gives

$$D = \mu_{\mathrm{NRT}}^{-1} + \frac{P_p}{(1-P_p)^2} \cdot \frac{1}{1-P_{\mathrm{NRT}}} \cdot E(\tau). \tag{5.13}$$

It can be seen that $\Delta D$ is a function of $P_{\mathrm{NRT}}$ and $P_p$. Since NRT is mostly data, it is very likely that it requires $r > 1$. The analysis of $r = 1$ is included here because it gives closed-form expression for $P_{\mathrm{NRT}}$, $P_p$ and thus $D$.

## 5.3 Numerical Results

Numerical results for the MB scheme are obtained and presented in this section. Each RT user occupies one channel, capable of supporting one voice user of 64 kbps. All NRT users have the same channel requirement, $r$. $r = 3$ and $r = 6$ are considered, corresponding to data rates of 192 kbps and 384 kbps, respectively. Denote $\lambda$ the aggregate arrival rate and $p_{\mathrm{NRT}}$ the probability that the arrival is NRT. The individual arrival rates are $\lambda_{\mathrm{RT}} = \lambda \cdot (1 - p_{\mathrm{NRT}})$ and $\lambda_{\mathrm{NRT}} = \lambda \cdot p_{\mathrm{NRT}} \cdot r$ for RT and NRT, respectively.

In the proposed MB scheme, RT traffic is not affected by the NRT traffic. Therefore, the focus is on the performance of the MB scheme for NRT traffic only. $K_{\mathrm{RT}}$ is chosen such that $P^*$ is satisfied for a given $\lambda$. Note that $\lambda$ governs $N_{u_{\mathrm{RT}}}$.

Table 5.1: Parameters for the two considered cases

| | Case 1 | Case 2 |
|---|---|---|
| $r$ | 3 | 6 |
| $p_{\mathrm{NRT}}$ | 0.2 | 0.1 |

as given in (3.18). For the NRT traffic, to satisfy $P_{\mathrm{NRT}} < P^{*}_{\mathrm{NRT}}$, the following two problems are studied:

1. The effect of $\sigma$ on NRT channel requirement, $K_{\mathrm{NRT}}$.

2. The reduction of utilization degradation $\Delta N_{u_{\mathrm{NRT}}}$ and excess capacity requirement $\Delta K_{\mathrm{NRT}}$ to support mobility with the use of MB scheme.

### 5.3.1 Effect of $\sigma$ on $K_{\mathrm{NRT}}$

To study the effect of $\sigma$ on $K_{\mathrm{NRT}}$, the ratio of RT to NRT traffic is held fixed, and user mobility is not considered, i.e., $P^{*}_{\mathrm{RT}} = P^{*}_{\mathrm{NRT}} = P^{*}_{n}$. For $r = 3$ and $r = 6$, $p_{\mathrm{NRT}}$ are chosen as given in Table 5.1. Although data traffic is now growing, in the foreseeable future, RT voice is still the dominant traffic [88]. The choices of $p_{\mathrm{NRT}}$ in Table 5.1 are such that the amount of RT voice traffic exceeds NRT data traffic. The traffic ratio of RT to NRT are 4:3 and 3:2 for the two cases.

The results are obtained as follows. The arrival rate $\lambda$ is varied. $\lambda_{\mathrm{RT}}$ and $\lambda_{\mathrm{NRT}}$ are varied accordingly with $p_{\mathrm{NRT}}$. The purpose is to keep the traffic ratio constant.

The utilizations are

$$\begin{aligned} N_{u_{\mathrm{RT}}} &= \lambda_{\mathrm{RT}}/\mu, \\ N_{u_{\mathrm{NRT}}} &= \lambda_{\mathrm{NRT}}/\mu. \end{aligned}$$

$1/\mu = 2$ minutes is assumed. As a result, as $\lambda$ is varied, $N_{u_{\mathrm{RT}}}$ and $N_{u_{\mathrm{NRT}}}$ are also varied. To satisfy $P^{*}_{\mathrm{RT}} \leq P^{*}_{n}$, the RT channel requirement, $K_{\mathrm{RT}}$ can be obtained using (3.18). The NRT channel requirement, $K_{\mathrm{NRT}}$ are obtained from simulations, for different values of $\sigma$.

Figures 5.4 to 5.5 show the effect of $\sigma$ on channel requirement and preemption probabilities, respectively, as a function of utilization, for $r = 6$, and $P^{*}_{n} = 10^{-3}$. The results are obtained in the absence of user mobility. Similar observations and conclusions can be drawn for $r = 3$. Reduction in NRT channel requirement is achieved for increasing $\sigma$. However, $P_p$ increases at the same time for increasing $\sigma$. It may appear that it is necessary to set a upper bound on $\sigma$ such that $\sigma \leq \sigma_{max} < K_{\mathrm{RT}}$. Otherwise, NRT delay $D$ may be too large. However, from Figure 5.5, we observe that the increase in $P_p$ decelerates for increasing $\sigma$. If $\sigma = K_{\mathrm{RT}}$ is allowed, $P_p$ is still kept below $10^{-3}$. In this case, from (5.10) to (5.12), $E(N_p) \approx Pp$ and $E(r) \approx 1$ because $P_{\mathrm{NRT}} = P^{*}_{n} = 10^{-3}$. $E(\tau)$ is probably in the millisecond range. As a result, $D$ is dominated by $\mu^{-1}_{\mathrm{NRT}}$. In other words, the introduction of the MB

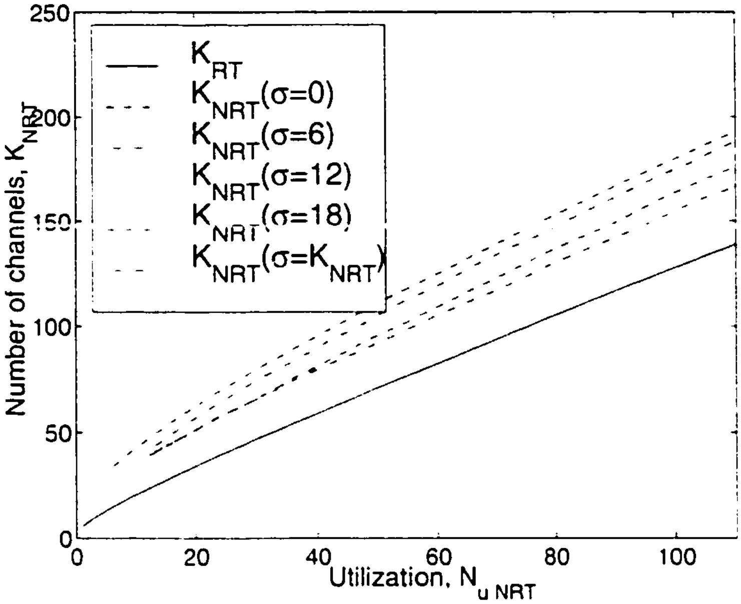

Figure 5.4: Channel Requirement as a function of Utilization for different $\sigma$ with $r = 6$

channel allocation scheme is not introducing significant excess delay. Therefore, it is concluded that NRT traffic should be allowed to borrow any number of idle RT channels, i.e., $\sigma = K_{RT}$. The full advantage of MB can thus be exploited. For the results in the sequel, $\sigma = K_{RT}$ is used.

### 5.3.2 The Effect of User Mobility

Figure 5.6 shows the NRT channel requirement in order to achieve a desired utilization for different user mobility $\lambda_h$ and two values of $r$, subject to satisfaction of $P_n^* = P_f^* = 10^{-3}$. It is observed that more channels are required for increasing

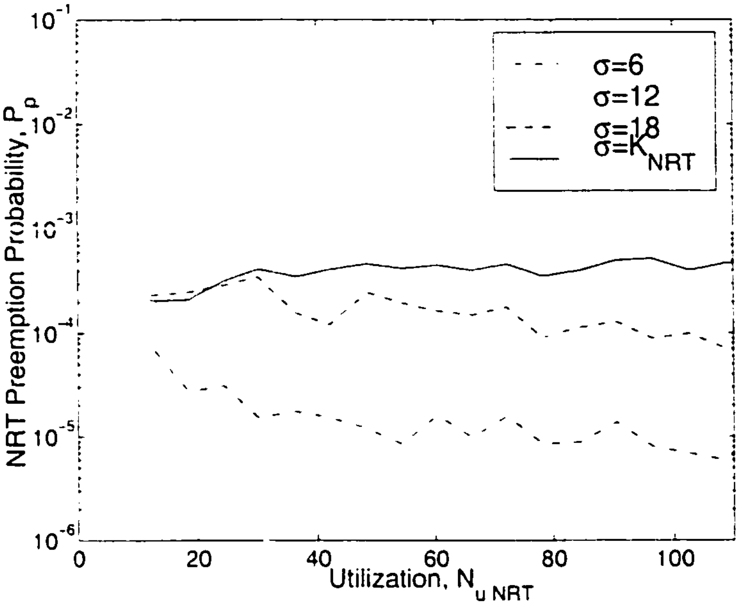

Figure 5.5: Preemption Probabilities as a function of Utilization for different $\sigma$ with $r = 6$

mobility. In addition, when each data user requires more channels (increasing $r$), a larger $K_{NRT}$ is necessary in order to maintain the same utilization. From this diagram, it re-confirms Chapter 3 with the effect of mobility:

1. To support higher mobility, utilization will decrease if the number of channel is fixed.

2. To support higher mobility, excess capacity is necessary if we want to maintain the utilization.

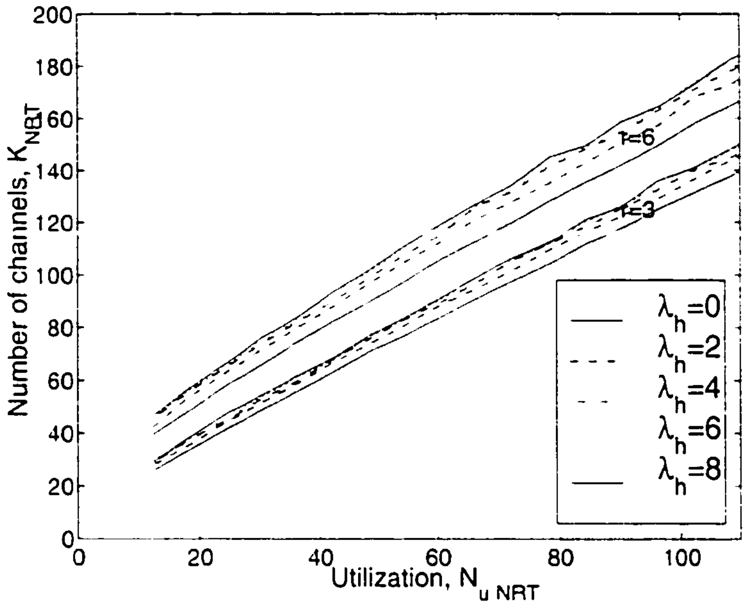

Figure 5.6: NRT channel requirement as a function of utilization for different mobility and $r$ ($P_n^* = P_f^* = 10^{-3}$)

### 5.3.3 Utilization Degradation

To study the effective of the MB channel allocation scheme, in terms of utilization performance in the presence of user mobility, arbitrary nominal allocations are assumed: $K_{RT} = 135$, $K_{NRT} = 106$, and they are held fixed. Using (3.18), $N_{u_{RT}} = 106.42$ in order to satisfy $P_n^* = 10^{-3}$. $K_{NRT} = 106$ is chosen so that $N_{u_{RT}} = 80$ at $\lambda_h = 0$. The performance of the RT traffic can be analyzed based on the $M/M/K_{RT}/K_{RT}$ model in Chapter 3 because it is not affected by the MB scheme.

For the NRT traffic, Figures 5.7 to 5.9 show the utilization, utilization degradation and percentage utilization degradation, respectively, as a function of mobility and subject to simultaneous satisfaction of $P_n^* = P_f^* = 10^{-3}$ for $r = 3$. The performance is summaried in Table 5.2. Comparison is made with the case when the MB scheme is not used. It is seen that the utilization decreases for increasing mobility. This decrease is also reflected in Figure 5.6. However, with the MB channel allocation scheme, a much higher utilization is achieved. The difference between the two curves is the gain due to the MB scheme. From Table 5.2, it can be seen that with the MB channel allocation scheme, the utilization of the NRT traffic is increased by 17.15 Erlangs (27.52% increase) at $\lambda_h = 0$ and by 20.62 Erlangs (39.73%) at $\lambda_h = 10$. In addition, the degradation in utilization, and more importantly, the percentage degradation, are much smaller when the MB scheme is used. The percentage degradation is almost halved. The gain in the MB channel allocation scheme comes from the fact that it is able to utilize the idle RT channels. As mobility is increased, the utilization of RT traffic is decreased. This is due to the requirement of standby resources to support user mobility. The MB scheme is able to utilize the idle standby capacity to increase the NRT performance. Note that the utilization remains constant until $\lambda_h = \lambda_{h_{cr}}$. When $\lambda_h > \lambda_{h_{cr}}$, the generic blocking requirement at the cell, $P_{\mathrm{NRT}}^*$ is dependent only on $P_f^*$. This phenomenon is studied in Chapter 3. Note also that $K_{\mathrm{NRT}}$ takes on non-integer values. The results are obtained from interpolation between integer points in order to give smoother curves (as opposed to staircase-like curves). In practice, $K_{\mathrm{NRT}}$ must be integer valued.

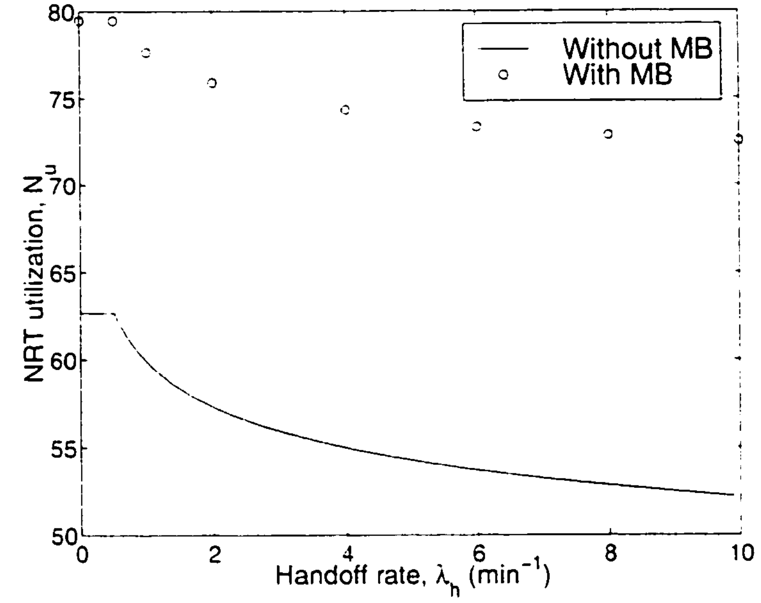

Figure 5.7: NRT utilization as a function of mobility ($N_{u_{NRT}}$ = 80, $r$ = 3, $P_n^* = P_f^* = 10^{-3}$)

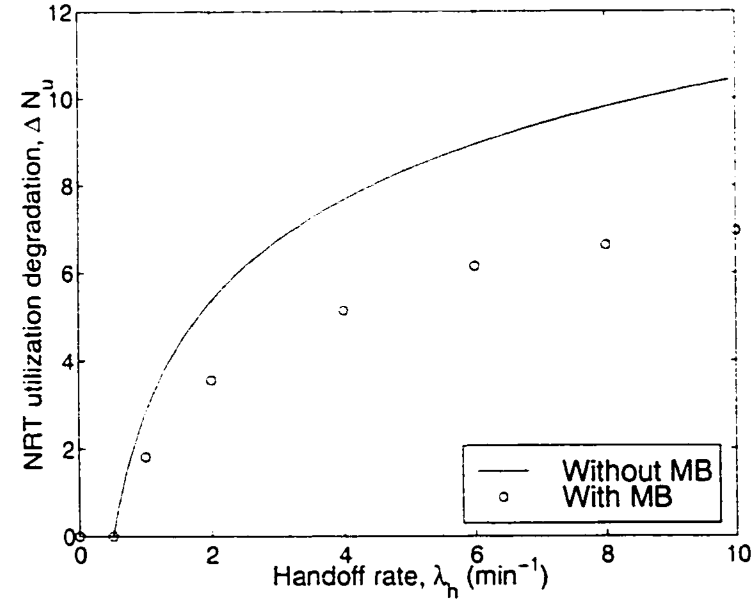

Figure 5.8: NRT utilization degradation as a function of mobility ($N_{u_{NRT}}$ = 80, $r$ = 3, $P_n^* = P_f^* = 10^{-3}$)

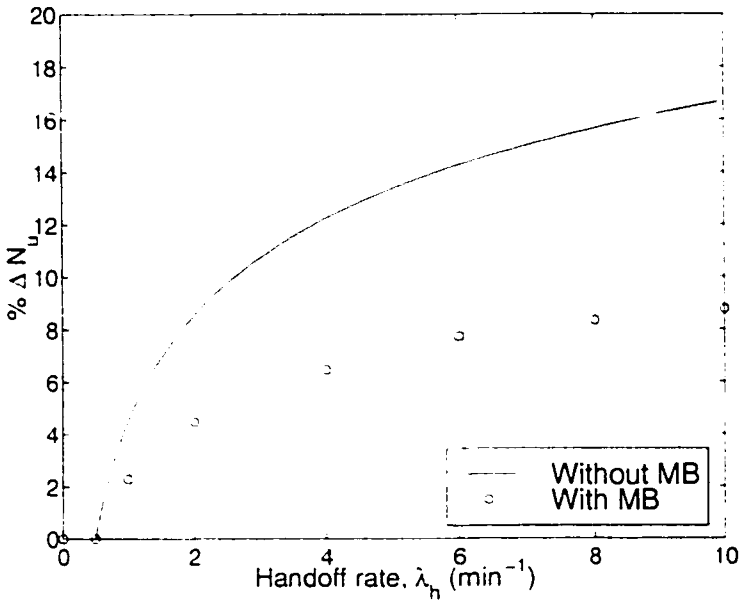

Figure 5.9: NRT % utilization degradation as a function of mobility ($N_{u_{NRT}} = 80$, $r = 3$, $P_n^* = P_f^* = 10^{-3}$)

Table 5.2: Utilization performance of the MB channel allocation scheme in terms of utilization with $P_n^* = P_f^* = 10^{-3}$, from $\lambda_h = 0$ to $\lambda_h = 10$

| MB scheme | $N_{u_{NRT}}$ | $\Delta N_{u_{NRT}}$ | % $\Delta N_{u_{NRT}}$ |
|---|---|---|---|
| Without | 62.32 → 51.89 | 10.43 | 16.74% |
| With | 79.47 → 72.51 | 6.96 | 8.76% |

### 5.3.4 Excess Capacity

To study effectiveness of the MB channel allocation scheme, in terms of excess capacity requirement in the presence of user mobility, utilizations are to be maintained constant at $\Lambda_{u_{\mathrm{RT}}} = 106.42$ and $\Lambda_{u_{\mathrm{NRT}}} = 80$, as chosen in the previous section. Note that as $\lambda_h$ is increased, it is necessary that $K_{\mathrm{RT}}$ is also increased, in order to maintain $\Lambda_{u_{\mathrm{RT}}} = 106.42$. Analysis from Chapter 3 can be used to study the RT traffic performance since it is unaffected by the NRT traffic.

For the NRT traffic, Figures 5.10 to 5.12 show the capacity requirement, excess capacity requirement and the percentage excess capacity requirement, respectively, as a function of mobility and subject to simultaneous satisfaction of $P_n^* = P_f^* = 10^{-3}$ for $r = 3$. The performance is also summaried in Table 5.3. Comparison of the performance with the case when the MB scheme is not used is made. It is seen that when the $\Lambda_{u_{\mathrm{NRT}}}$ is to be maintained, the capacity requirement is increased. This increase is also reflected in Figure 5.6. However, with the MB channel allocation, the capacity requirement to achieve $\Lambda_{u_{\mathrm{NRT}}} = 80$ is reduced. From Table 5.3, it can be seen that with the MB channel allocation scheme, the capacity requirement of the NRT traffic is decreased by 22.94 (17.09%) at $\lambda_h = 0$ and 28.50 (29.68%) at $\lambda_h = 10$. This improvement is, again, coming from the ability of the MB scheme to utilize the idle RT channels, putting aside standby to support mobility. The excess

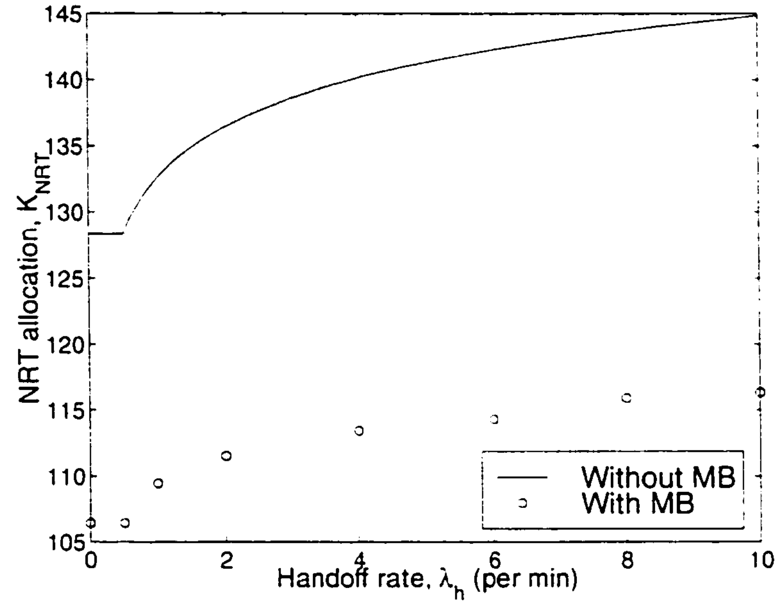

Figure 5.10: NRT capacity requirement as a function of mobility ($N_{u_{NRT}}$ = 80, $r = 3$, $P_n^* = P_f^* = 10^{-3}$)

capacity and percentage excess capacity requirements are also reduced significantly. This means that with the MB channel allocation scheme, less extra overhead is required to support increasing mobility.

Numerical results are also obtained for $r = 6$ and different target $N_{u_{NRT}}$. Results are similar to Figures 5.7 to 5.11 and are not shown here. Therefore, the MB scheme is very effective. It increases the radio cell utilization and reduces degradation of utilization when supporting higher degree of user mobility.

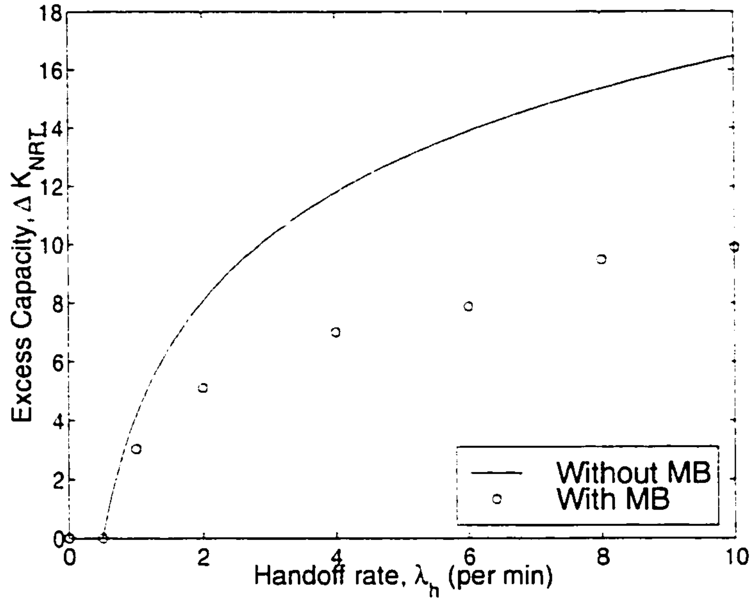

Figure 5.11: NRT excess capacity requirement as a function of mobility ($N_{u_{NRT}}$ = 80, $r = 3$, $P_n^* = P_f^* = 10^{-3}$)

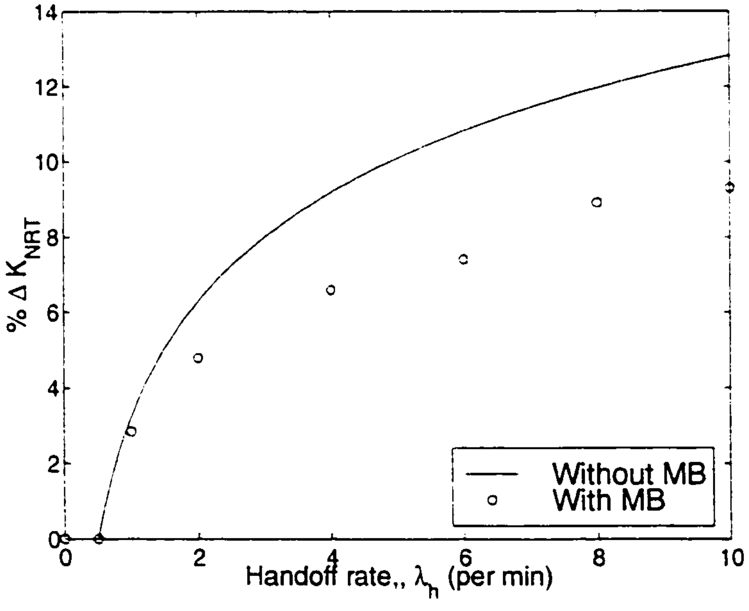

Figure 5.12: NRT % excess capacity requirement as a function of mobility ($N_{u_{NRT}}$ = 80, $r = 3$, $P_n^* = P_f^* = 10^{-3}$)

Table 5.3: Excess capacity performance of the MB channel allocation scheme in terms of utilization with $P_n^* = P_f^* = 10^{-3}$. from $\lambda_h = 0$ to $\lambda_h = 10$

| MB scheme | $K_{NRT}$ | $\Delta K_{NRT}$ | % $\Delta K_{NRT}$ |
|---|---|---|---|
| Without | 128.36 → 144.84 | 16.47 | 12.83% |
| With | 106.13 → 116.33 | 9.90 | 9.31% |

## 5.4 Discussions

The effectiveness of the MB channel allocation scheme comes from the fact that the NRT traffic can utilize the idle RT channels. The amount of RT capacity which may be borrowed by NRT traffic is given by

$$\zeta = 1 - N_{u_{RT}}. \tag{5.14}$$

As a result, when the capacity allocated to the RT and NRT traffic is $c$ Mbps each, the effective capacity for the NRT traffic is $c \cdot (1 + \zeta)$. As mentioned in Chapter 3, with an increased capacity, the utilization degradation and excess capacity requirement to support user mobility are reduced. At high user mobility, since more standby resource is necessary, $N_{u_{RT}}$ is decreased. Thus, $\zeta$ is increased and NRT traffic can borrow even more idle RT capacity. Therefore, the utilization degradation and excess capacity requirement in the presence of mobility, using the MB channel allocation scheme, is further reduced.

## 5.5 Conclusions

In this chapter. the proposed movable boundary chapter allocation scheme for handling RT and NRT traffic types. at a base station. is studied. This scheme allows NRT traffic to make use of unused RT capacity. For a given QoS requirement. it is demonstrated that RT service performance is unaffected while NRT resource requirement is reduced. This is achieved by allowing RT users to preempt NRT users who are borrowing RT channels. when necessary. The MB scheme is simple to implement. It increases the resource utilization and reduces the capacity requirement. The degradation in utilization and excess capacity requirement in order to support user mobility are significantly reduced. It makes the MB scheme especially applicable in the wireless mobile environment.

# Chapter 6

# Call Admission Control

This chapter integrates the joint-levels resource allocation scheme (in Chapter 4) and the movable boundary channel allocation scheme (in Chapter 5), and incorporates the findings of the effects of mobility (in Chapter 3), to devise a resource allocation strategy and provide a mechanism for structuring call admission control.

The effectiveness of the joint-levels resource allocation (RA) scheme and the movable boundary (MB) channel allocation scheme are demonstrated in chapters 4 and 5, respectively. With a fixed capacity, the utilization can be increased. In the presence of user mobility, the utilization degradation and the percentage degradation are reduced. Alternatively, with user mobility, excess capacity is necessary in order to maintain utilization at zero-mobility. With either of the two scheme, the amount of excess capacity required is reduced. Either scheme used in isolation already gives promising results. When they are used together, the performance can be further enhanced. In practice, the available capacity is fixed for a given antenna design and channel coding scheme. Therefore, in a practical situation, the joint-levels RA scheme and the MB channel allocation scheme can be used to increase utilization. The aim of this chapter is to study the effectiveness of the combined scheme in the utilization performance.

In the combined scheme, two traffic classes are considered: RT and NRT. The MB channel allocation scheme is used to increase the NRT utilization by utilizing idle RT channels, while satisfying the call-level QoS factors of $P_n^*$ and $P_f^*$. Within the individual RT and NRT allocation, joint-levels RA is performed in order to guarantee both the call-level QoS factors, $P_n^*$ and $P_f^*$, and the packet-level QoS factor, PLR$^*$.

The rest of this chapter is organized as follows. In Section 6.1, the performance of the combined scheme is studied. In Section 6.2, the mechanism for CAC using the combined scheme is described.

## 6.1 Performance of the Combined Scheme

In order to keep the analysis within a manageable size, $c = 1.5$ Mbps is assumed, as in Chapter 4. The mean call holding time, $\mu^{-1} = 2$ mins. The call-level QoS

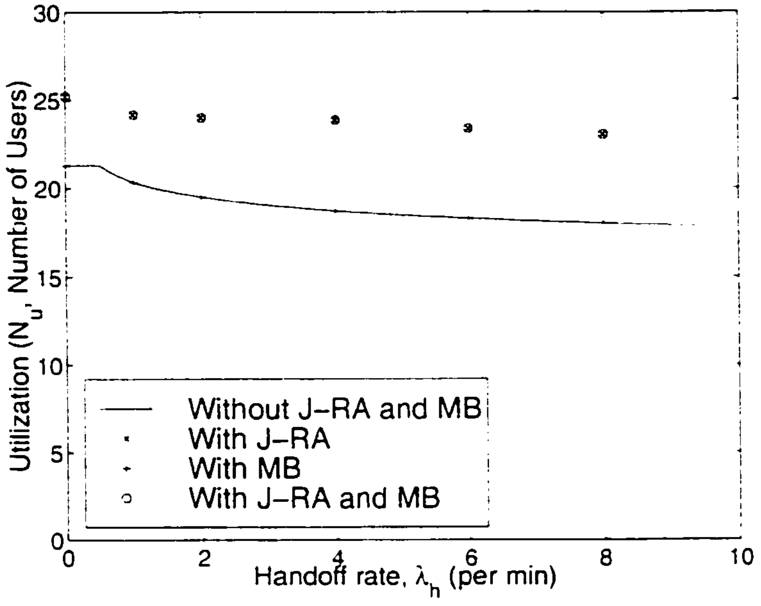

Figure 6.1: Utilization as a function of mobility for RT service ($c$ = 1.5 Mbps, $\mu^{-1} = 2$ mins, $P_n^* = P_f^* = 10^{-3}$, PLR* = $10^{-6}$)

factors are $P_n^* = P_f^* = 10^{-3}$. The packet-level QoS factor is PLR* = $10^{-6}$. The utilization performance of 1) the joint-levels RA scheme alone, 2) the MB channel allocation scheme alone, and 3) the combined scheme, are compared.

For the RT service, Figures 6.1 to 6.3 show the utilization, utilization degradation and percentage utilization degradation, respectively, as a function of mobility, subject to simultaneous satisfaction of the packet-level and call-level QoS factors. From the packet-level analysis, $c$ = 1.5 Mbps corresponds to 36 channels in order to satisfy PLR* = $10^{-6}$.

It is observed that the performance improvement only comes from the joint-

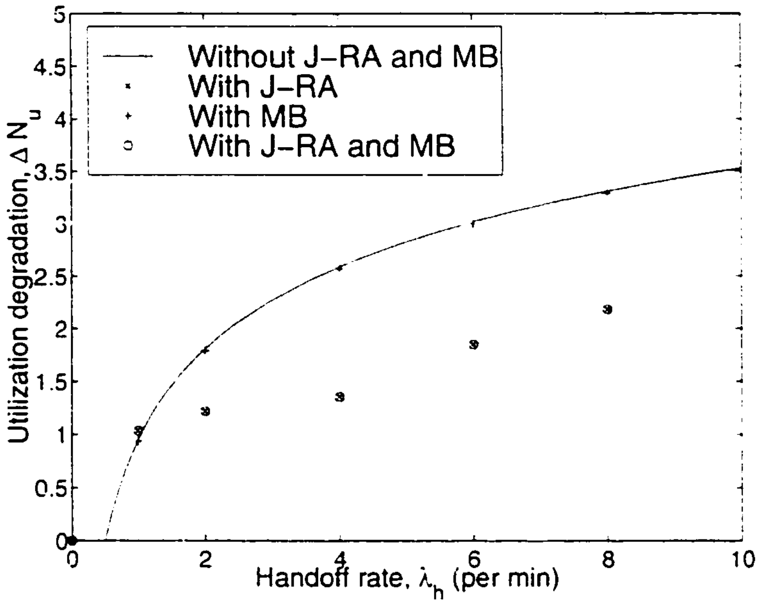

Figure 6.2: Utilization degradation as a function of mobility for RT service ($c$ = 1.5 Mbps, $\mu^{-1}$ = 2 mins, $P_n^* = P_f^* = 10^{-3}$, $\text{PLR}^* = 10^{-6}$)

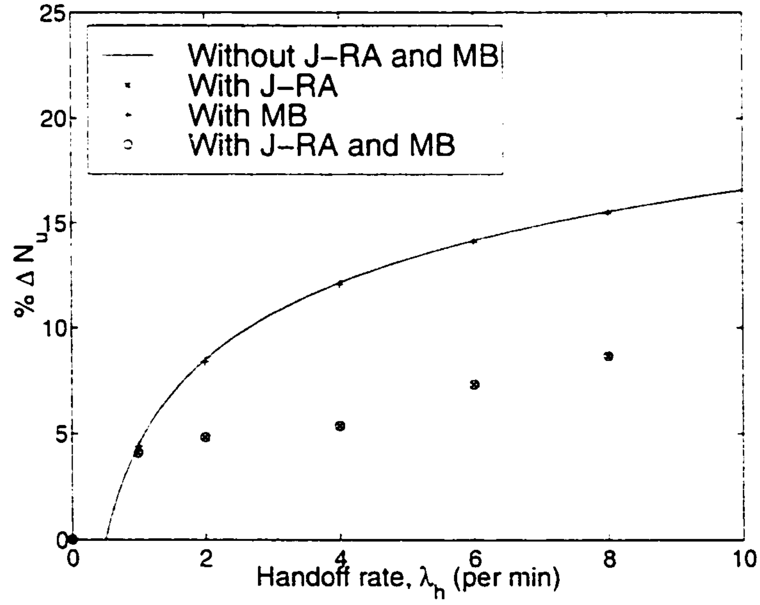

Figure 6.3: % utilization degradation as a function of mobility for RT service ($c = 1.5$ Mbps, $\mu^{-1} = 2$ mins, $P_n^* = P_f^* = 10^{-3}$, $\text{PLR}^* = 10^{-6}$)

Table 6.1: Performance of the combined scheme for the RT service in terms of utilization with $P_n^* = P_j^* = 10^{-3}$, from $\lambda_h = 0$ to $\lambda_h = 8$

| Scheme | $N_{u_{NRT}}$ | $\Delta N_{u_{NRT}}$ | $\% \ \Delta N_{u_{NRT}}$ |
|---|---|---|---|
| Without J-RA | 21.27 → 17.99 | 3.28 | 15.42% |
| With J-RA | 25.21 → 23.03 | 2.18 | 8.65% |

levels (RA) scheme. This is expected because the MB channel allocation scheme gives higher priority to RT traffic. The performance of NRT service improves without affecting the RT service performance. From Figures 6.2 and 6.3, it appears that the points obtained for using the joint-levels RA or the combined scheme are not following a smooth trend. This is due to discretization of $K_{NRT}$. In Chapter 5, the capacity requirement results $K_{NRT}$, e.g., in Table 5.3, are obtained by interpolation, giving rise to non-integer values. In practice, both $K_{RT}$ and $K_{NRT}$ must be integer-valued. For example, at $\lambda_h = 0$, the nominal capacity of $K_{RT}$ can be increased from 36 to 41.60 channels. However, 41 must be used in practice. Discretization causes the results to be slightly rough. Nonetheless, the general trend is still revealed clearly. Table 6.1 summarizes the utilization performance when $\lambda_h$ is increased from 0 to 8. At $\lambda_h = 0$, the utilization is increased by 3.94 Erlangs (18.52%) with either the joint-levels RA or the combined scheme. At $\lambda_h = 8$, the utilization is increased by 5.04 Erlangs (28.02%). At the same time, the utilization degradation and the percentage degradation due to user mobility are also reduced.

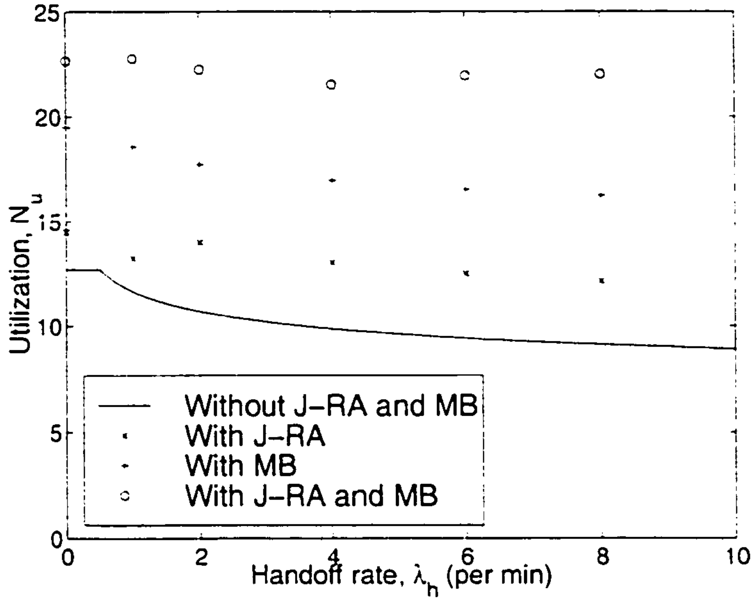

Figure 6.4: Utilization as a function of mobility for NRT service ($c$ = 1.5 Mbps, $\mu^{-1}$ = 2 mins, $P_n^* = P_f^* = 10^{-3}$, PLR$^* = 10^{-6}$)

For the NRT service, Figures 6.4 to 6.6 show the utilization, utilization degradation and percentage utilization degradation, respectively, as a function of mobility, subject to simultaneous satisfaction of the packet-level and call-level QoS factors. The NRT capacity is also $c$ = 1.5 Mbps, corresponding to 36 channels. Each NRT source requires $r$ = 3 channels.

Similar to the RT case, due to discretization, the results for the cases involving the joint-levels RA scheme do not yield perfectly smooth curves. However, the general trend is revealed clearly. Table 6.2 summarizes the utilization performance when $\lambda_h$ is increased from 0 to 8. Using the case when none of the two schemes

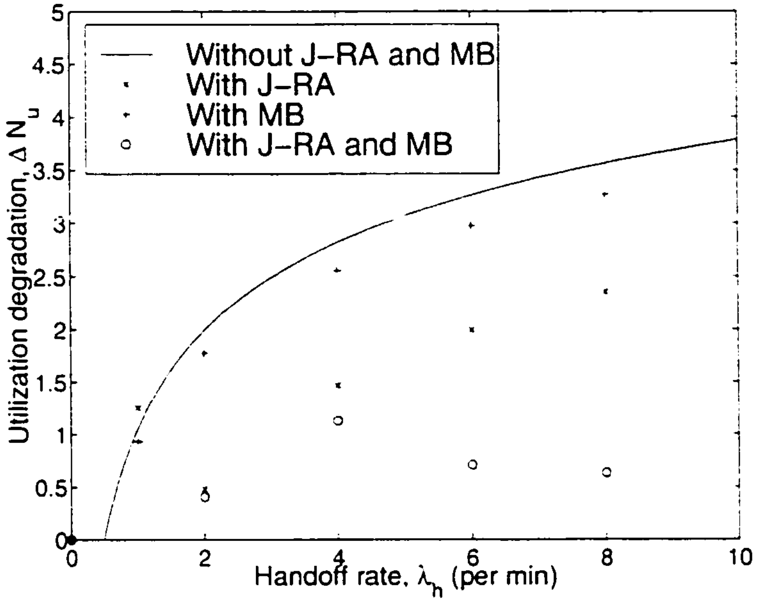

Figure 6.5: Utilization degradation as a function of mobility for NRT service ($c$ = 1.5 Mbps, $\mu^{-1}$ = 2 mins, $P_n^* = P_f^* = 10^{-3}$, $\text{PLR}^* = 10^{-6}$)

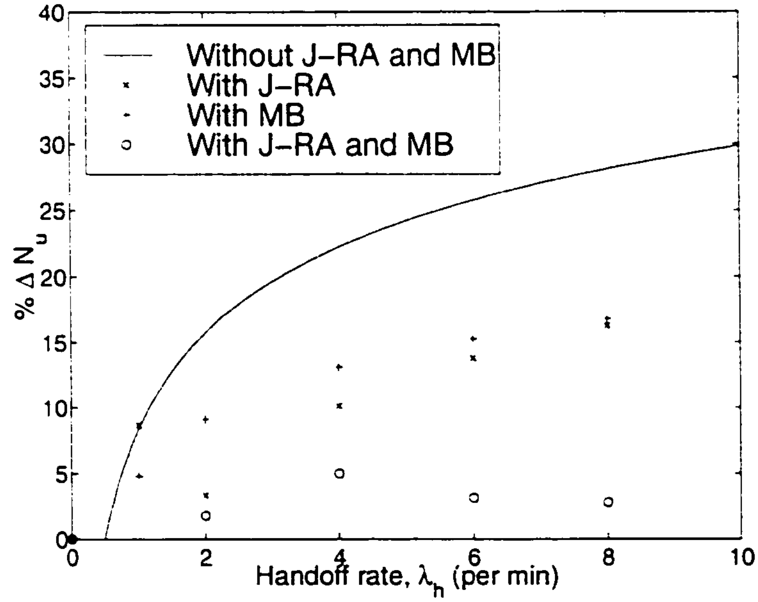

Figure 6.6: % utilization degradation as a function of mobility for NRT service ($c$ = 1.5 Mbps, $\mu^{-1}$ = 2 mins, $P_n^* = P_f^* = 10^{-3}$, $\text{PLR}^* = 10^{-6}$)

Table 6.2: Performance of the combined scheme for the NRT service in terms of utilization with $P_n^* = P_f^* = 10^{-3}$. from $\lambda_h = 0$ to $\lambda_h = 8$

| Scheme | $N_{u_{NRT}}$ | $\Delta N_{u_{NRT}}$ | $\% \Delta N_{u_{NRT}}$ |
|---|---|---|---|
| Without J-RA and MB | 12.68 → 9.12 | 3.56 | 28.01% |
| With J-RA only | 14.48 → 12.12 | 2.36 | 16.30% |
| With MB only | 19.49 → 16.22 | 3.27 | 16.78% |
| Combined | 22.64 → 22.01 | 0.63 | 2.78% |

are used as the benchmark, the improvement by the joint-levels RA scheme, MB channel allocation scheme, and the combined scheme, are clearly shown. Using the combined scheme, the utilization is increased by 9.96 Erlangs (78.52%), at $\lambda_h = 0$. At $\lambda_h = 8$, the utilization is increased by 12.89 Erlangs (141.36%). Significant gain in utilization is achieved when using the combined scheme. From Figures 6.5 and 6.6, the utilization degradation and the percentage utilization degradation are also significantly reduced.

## 6.2 Call Admission Control

The integration of the joint-levels RA scheme and the MB channel allocation scheme as a combined method is shown to be a viable resource allocation technique. This section uses the combined method to structure an effective call admission control mechanism.

CAC makes admission decision on the basis that the QoS requirements of the request as well as all the on-going calls are met. The QoS requirements consist of the packet-level factor, packet loss rate, PLR$^*$, and the call-level factors, new call blocking probability, $P_n^*$ and forced termination probability, $P_f^*$. Two traffic classes are considered, real-time (RT) and non-real-time (NRT).

The capacity of the transmission link is $c$ Mbps. $c$ is offered to the users through $K_c$ transmission channels. In TDMA, $K_c$ is the number of time slots, while in FDMA, $K_c$ is the number of frequency slots. $K_c$ is chosen such that each channel provides a transmission rate of $R$ bps while satisfying the physical layer bit error rate specification. In CDMA, $K_c$ is the number of spreading codes. $K_c$ is such that (3.3) satisfies.

Provided that the physical layer bit error rate specification is always met by lower layer functional blocks such as antenna design and source coding, the nominal capacity is determined by the packet-level QoS constraint, PLR$^*$. For a given link rate of $c$, it is assumed that $c_{\mathrm{RT}}$ is allocated to RT traffic and $c_{\mathrm{NRT}}$ is allocated NRT traffic. With a knowledge of $c_{\mathrm{RT}}$, $c_{\mathrm{NRT}}$ and PLR$^*$, packet-level resource allocation method such as the effective bandwidth may be used to determine the maximum number of users $K'_{\mathrm{RT}}$ and $K_{\mathrm{NRT}}$, which $c$ can support.

With the QoS constraints of $P_n^*$ and $P_f^*$, the utilization $N_{u_{\mathrm{RT}}}$ can be determined. Using the MB channel allocation scheme, the utilization of the NRT service $N_{u_{\mathrm{NRT}}}$ is enhanced. The enhancement is due to the ability to utilize idle RT channels so that the NRT allocation is increased to $K_{\mathrm{NRT}}(1+\zeta)$ effectively. The joint-levels RA scheme can then be used within the RT and NRT allocations to increase the nominal capacity from $K_{\mathrm{RT}}$ to $K'_{\mathrm{RT}}$, and $K_{\mathrm{NRT}}$ to $K'_{\mathrm{NRT}}$. This improvement comes from the incorporation of the call-level traffic behaviour in the packet-level traffic analysis. The CAC mechanism is summarized as follows.

1. Use effective bandwidth to determine the nominal capacity (maximum number of users) of each traffic class, $K_{\mathrm{RT}}$ and $K_{\mathrm{NRT}}$.

2. Use the MB channel allocation scheme to increase the effective nominal capacity for the NRT traffic from $K_{\mathrm{NRT}}$ to $K_{\mathrm{NRT}}(1+\zeta)$.

3. Use the joint-levels RA scheme to increase the nominal capacity from $K_{\mathrm{RT}}$ to $K'_{\mathrm{RT}}$, and $K_{\mathrm{NRT}}(1+\zeta)$ to $K'_{\mathrm{NRT}}$.

With the combined scheme, the nominal capacity for both the RT and NRT traffic are increased. At the same time, the utilization of the NRT allocation is enhanced. An increase in nominal capacity and utilization means an increase in user population. It is conjectured that the combined scheme offers a means to the network provider to increase revenue.

# Chapter 7

# Conclusions and Future Works

## 7.1 Conclusions

This thesis is concerned with call admission control in a wireless environment. CAC is performed from the perspective of resource allocation. It consists of three major building blocks.

First, the effect of user mobility on QoS provisioning is studied. It is shown that user mobility entails standby resources. The call-level QoS constraints, $P_n^*$ and $P_f^*$, are encapsulated in a single parameter, $P^*$, which is the generic blocking requirement at the radio cell. It is found that with a given capacity, the presence of user mobility degrades the utilization. In order to maintain the utilization with the introduction of user mobility, excess capacity needs to be provided. These results are used to assess the performance of the proposed resource allocation and channel allocation schemes in the presence of mobility. A linkage between the link layer specification and network layer performance, in a CDMA environment, is then established. The tradeoff between $E_b/I_0$ specification and system utilization is provided. This work provides a fundamental understanding of the salient feature, namely user roaming, of wireless mobile networking.

Second, the joint-levels resource allocation (RA) scheme is proposed. It is based on simultaneous satisfaction of the packet-level and call-level QoS factors. This scheme is enlightened from the observation that at the call level, calls are non-persistent. The joint-levels RA scheme incorporates the call-level behaviour in the packet-level traffic analysis so as to increase the nominal capacity. The improvement in the nominal capacity in turn enhance the utilization significantly.

Third, the movable boundary (MB) channel allocation scheme is proposed. It is a form of virtual partitioning in the sense the NRT traffic can borrow idle RT channels, on the basis that it may be preempted if RT traffic needs them back at a later time. The MB channel allocation scheme increases the NRT channels utilization without affecting the RT service performance. By allowing NRT traffic to borrow idle RT channels, the number of NRT channels are increased effectively to realize the utilization enhancement.

The call admission control mechanism is based on the integration of the joint-levels RA scheme and the MB channel allocation scheme. The former is used to increase the nominal capacity in order to support more users. The latter is used to increase the carried traffic intensities to yield a higher utilization. The CAC mechanism based on the combined scheme increases the utilization drastically. It provides a means to the network provider to increase the user population, and hence, revenue.

The contributions of this thesis are summarized as follows.

1. A fundamental understanding of the effect of user mobility on QoS provisioning is provided.

2. A linkage between the link layer specification and network layer QoS requirements is established. The tradeoff between the $E_b/I_0$ specification and utilization performance is provided.

3. A resource allocation scheme based on simultaneous satisfaction of packet-level and call-level QoS constraints is proposed. By taking the call-level behaviour into the packet-level analysis, more multiplexing is achieved and it results in an increase in utilization.

4. A channel allocation scheme based on movable boundary, supporting two traffic classes is proposed. By allowing the NRT traffic to utilize idle RT channels, the capacity is utilized more efficiently.

5. A combined scheme based on the integration of the joint-levels RA scheme and MB channel allocation scheme is proposed. It inherents the benefits of both schemes and utilization is significantly increased.

6. An effective CAC mechanism based on the combined scheme is structured.

## 7.2 Future Works

The mean handoff time $\lambda_h^{-1}$ is assumed exponential throughout the thesis. As pointed out in [83, 84], the mean handoff time may be better described by a gamma distribution. It is interesting to investigate how the handoff time distribution affects the results obtained in Chapter 3, and the dependency of the effects of mobility on the handoff time distributions.

The packet-level analysis in Chapter 4 is based on the effective bandwidth approach. In Chapter 6, two traffic classes are considered. It is interesting to generalize the joint-levels RA scheme to handle $M$ traffic classes with different traffic characteristics and QoS requirements, using results from, e.g., [8]. The findings can then be used to upgrade the combined scheme to be used in structuring the CAC mechanism.

In order to facilitate the analysis to demonstrate the viability of the joint-levels RA scheme and the MB channel allocation scheme, new calls and handoff calls are treated equally likely. In practice, handoff calls should have higher priority than new calls because interruption of an ongoing call is more annoying than receiving a busy signal by a new call. It is of practical interest to extend the findings of the effects of user mobility and enhance the joint-levels RA scheme and the MB channel allocation scheme to handle handoff calls with a higher priority. This is also crucial for CAC.

This thesis is focused on the wireless subnet of the end-to-end communications system, as shown in Figure 1.2. It is interesting to extend the works to the hybrid wireline/wireless networks. As pointed out in Chapter 1, the end-to-end system consists of a collection of wireless subnets, interconnected through a backbone wireline network. A mobile terminal communicates with the serving base station via a radio link. The base station connects to the backbone network by wireline links. A base station is in effect the wireline attachment point for each connection. In a hybrid wireline/wireless network, when a mobile user migrates to a neighbouring radio cell, the necessary handoff may trigger a change in the wireline attachment point. This is an interesting phenomenon, unique to the interconnected environment. Issues such as the effects of the change in wireline attachment point on resource management and QoS provisioning are critical. With a good understanding of these issues, effective resource management schemes and call admission control schemes can be devised. Communications with high resource utilization and QoS guarantee, supporting multimedia traffic. is thus made possible. This is of paramount importance for the success of the third generation wireless systems.

# Appendix A

# Forced Termination Probability

The forced termination probability, $P_f$, given by (3.11) in Section 3.2, can be expressed as

$$\theta P^2 - \theta(1 - P_f)P + P_f(1 - \theta) = 0. \tag{A.1}$$

Using the quadratic formula, $P$ can be expressed in terms of $P_f$ and $\theta$ as

$$P = \frac{\theta(1 - P_f) \pm \sqrt{\theta^2(1 - P_f)^2 - 4\theta P_f(1 - \theta)}}{2\theta}. \tag{A.2}$$

There are two issues involved: 1) whether the discriminant is non-negative, and 2) which of the two solutions should be picked. To ensure a positive discriminant in (A.2), it is required that $\theta^2(1 - P_f)^2 > 4\theta P_f(1 - \theta)$. Consider $P_f = 10^{-3}$, it is necessary that $\theta > 0.016$. For $P_f = 10^{-6}$, $\theta > 1.000002 \times 10^{-6}$. Therefore, $\theta$ is to be very small to result in a negative discriminant term in (A.2). In this case, mobility is very low and $P_f \approx 0$ from (3.11). From (3.9), $\theta$ is small when $\lambda_h$ is very small or $\mu$ is very large. In the former case, we may simply treat $\lambda_h = 0$ and $P_f = 0$. In the latter case, call duration is very small, perhaps in the range of milliseconds. This will not happen in practice. Therefore, in practice, we may assume the discriminant in (A.2) to be always positive.

The two solutions to (A.1) are:

$$P_1 = \frac{\theta(1 - P_f) + \sqrt{\theta^2(1 - P_f)^2 - 4\theta P_f(1 - \theta)}}{2\theta}. \tag{A.3}$$

and

$$P_2 = \frac{\theta(1 - P_f) - \sqrt{\theta^2(1 - P_f)^2 - 4\theta P_f(1 - \theta)}}{2\theta}. \tag{A.4}$$

Note that the square root term is always less than $\theta(1 - P_f)$; therefore, $P_1$ and $P_2$ are always positive. In addition, $P_1 > P_2$. In the case of $P_1$, $P_f$ is satisfied since the blocking probability of the queueing system is *too large*. Most calls are being blocked upon initial request. As a result, $P_f$ is kept below the desired level.

In the case of $P_2$, $P_f$ is satisfied because the blocking probability of the queueing system is kept very small by having enough *standby resources*. Clearly, this is the desired situation whereas $P_1$ is not. Therefore, we use the second solution where the negative of the discriminant is used, i.e., $P$ is given by the right-hand side of (A.4).

# Appendix B

# Typical Values of Parameters

Call duration typically ranges from a few seconds to, perhaps, an hour. Table B.1 shows some typical values of call duration and the corresponding call completion rate, $\mu$. $\mu$ ranges between 0.01 and 10 (per min). Short calls of a few seconds can arise when the callee is busy or unavailable, while calls of half an hour in duration can arise when the two parties are discussing, perhaps interesting research problems.

Three different types of communication cells are considered: macrocells, microcells and picocells. Table B.2 shows some typical values of handoff rate, $\lambda_h$. $\lambda_h$ ranges between 0.1 and 10 (per min). Notice that a user speed of 100 km/h is typical when the user is traveling on a highway. Speed of 50–60 km/h arises when user is traveling along urban streets. On the other hand, user speed of 2–4 km/h occurs when a user is walking along city streets or within a building.

Table B.1: Typical call duration and $\mu$.

| Call duration (min) | 0.167 | 0.5 | 1 | 2 | 3 | 4 | 5 | 10 | 30 | 60 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mu$ (min$^{-1}$) | 6 | 2 | 1 | 0.5 | 0.333 | 0.25 | 0.2 | 0.1 | 0.0333 | 0.0167 |

Table B.2: Typical user speed and handoff rate $\lambda_h$.

| Cell Type | Diameter | Speed | $\lambda_h$ (min$^{-1}$) |
|---|---|---|---|
| Macrocell | 2 km | 100 km/h | 0.833 |
| | | 60 km/h | 0.5 |
| Microcell | 500 m | 50 km/h | 1.67 |
| | | 4 km/h | 0.13 |
| | 200 m | 50 km/h | 4.16 |
| | | 4 km/h | 0.33 |
| Picocell | 30 m | 4km/h | 2.22 |
| | | 2km/h | 1.11 |
| | 10 m | 4km/h | 6.67 |
| | | 2km/h | 3.33 |

# Appendix C

# Fluid Flow Analysis

This appendix gives the procedures to assess the PLR using the fluid flow approach as described in [15] and [87].

The buffer occupancy distribution is denoted by

$$\boldsymbol{F}(x) \triangleq [F_0(x), F_1(x), \cdots, F_K(x)], \tag{C.1}$$

where $F_i(x)$ is the stationary probability that the buffer occupancy is less than or equal to $x$, when $i$ sources are in talkspurt. $K$ is the number of multiplexed voice sources. The set of differential equations that describes the buffer occupancy is given by

$$\begin{aligned}(i-C)\alpha_s\frac{dF_i(x)}{dx} =&[K-(i-1)]\eta_s F_{i-1}(x) - \\ &[(K-i)\eta_s + i\alpha_s]F_i(x) + (i+1)\alpha_s F_{i+1}(x), \qquad 0 \le i \le K.\end{aligned} \tag{C.2}$$

with $F_{-1}(x) = F_{K+1}(x) = 0$. Note that there are $K+1$ equations. Eq.(C.2) may be written in a compact matrix form:

$$\begin{aligned} D\frac{\boldsymbol{F}(x)}{dx} &= M\boldsymbol{F}(x) \\ \frac{\boldsymbol{F}(x)}{dx} &= D^{-1}M\boldsymbol{F}(x) \end{aligned} \tag{C.3}$$

where $D$ is an $(K+1) \times (K+1)$ diagonal matrix:

$$D = \text{diag}[-C\alpha_s, (1-C)\alpha_s, \cdots, (K-C)\alpha_s]. \tag{C.4}$$

and $M$ is derived from (C.2):

$$M = \begin{bmatrix} -K\eta_s & \alpha_s & 0 & \cdots \\ K\eta_s & -[\alpha_s + (K-1)\eta_s] & 2\alpha_s & \cdots \\ 0 & (K-1)\eta_s & -[2\alpha_s + (K-2)\eta_s] & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{C.5}$$

The solution to (C.3) is a sum of exponentials of the form

$$\boldsymbol{F}(x) = \sum_{i=0}^{K} a_i F_i e^{z_i x} \tag{C.6}$$

where $z_i$ is the $i^{th}$ eigenvalue of $D^{-1}M$, with the corresponding eigenvector, $\Phi_i$. Each row in (C.6) can then be written as

$$F_j(x) = \sum_{i=0}^{K} a_i \Phi_{ij} e^{z_i x}. \qquad 0 \le j \le K. \tag{C.7}$$

To obtain $\boldsymbol{F}$(x), we need to solve for the coefficients $a_i$'s in (C.7). By noting that $F_i(0) = 0$ when $i > C$ and the fact that $z_0 = 0$ and $\Phi_0 = \boldsymbol{\pi}$ is a valid eigenvalue-eigenvector pair, (C.7) reduces to

$$F_j(0) = 0 = \pi_j + \sum_{i=\lceil C \rceil}^{K} a_i \Phi_{ij}. \qquad \lceil C \rceil \le j \le K. \tag{C.8}$$

$\lceil C \rceil$ is the smallest integer larger than or equal to $C$ (the ceiling function of $C$). $\boldsymbol{\pi}$ is the steady state probability vector of dimension $K + 1$ where the $i$th element is given by

$$\pi_i = \binom{K}{i} \left( \frac{\eta_s}{\eta_s + \alpha_s} \right)^i \left( \frac{\alpha_s}{\eta_s + \alpha_s} \right)^{K-i}. \qquad 0 \le i \le K. \tag{C.9}$$

Eq. (C.8) gives $K - \lfloor C \rfloor$ equations. $\boldsymbol{F}(x)$ is obtained by solving for the $K - \lfloor C \rfloor$ values of $a_i$. $\lfloor C \rfloor$ is the largest integer smaller than or equal to $C$ (the floor function of $C$). PLR$(x) = 1 - \sum_{j=0}^{K} F_j(x)$, where $x$ is the normalized buffer size.

# Bibliography

[1] A. Acampora, “Wireless ATM: A perspective on issues and prospects,” *IEEE Personal Communications*, pp. 8–17, August 1996.

[2] A. S. Acampora and M. Naghshineh, “Control and quality of service provisioning in high-speed microcellular networks,” *IEEE Personal Communications Magazine*, vol. 1, no. 2, 1994.

[3] E. Ayanoglu, K. Y. Eng, and M. J. Karol, “Wireless ATM: Limits, challenges, and proposals,” *IEEE Personal Communications*, pp. 18–34, August 1996.

[4] M. Naghshineh and A. S. Acampora, “QoS provisioning in micro-cellular networks supporting multiple classes of traffic,” *Wireless Networks 2*, pp. 195–203, 1996.

[5] D. Raychaudhuri, “Wireless ATM: An enabling technology for multimedia personal communication,” *Wireless Network 2*, pp. 163–171, 1996.

[6] M. Schwartz, "Network management and control issues in multimedia wireless networks," *IEEE Personal Communications*, pp. 8–16, June 1995.

[7] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "On the effectiveness of effective bandwidths for admission control in ATM networks," in *Proc. of ITC-14*, pp. 411–420, 1994.

[8] H. W. Lee and J. W. Mark, "Capacity allocation in statistical multiplexing of ATM sources," *IEEE/ACM Trans. on Networking*, vol. 2, pp. 139–151, April 1995.

[9] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. on Networking*, vol. 1, pp. 329–343, June 1993.

[10] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Systems*, vol. 9, pp. 17–27, 1991.

[11] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its applications to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 968–981, September 1991.

[12] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, vol. 9, pp. 5–16, 1991.

[13] L. Kosten, "Stochastic theory of a multi-entry buffer (I)," *Delft Progress Report*, vol. 1, pp. 10–18, 1974.

[14] I. Hsu and J. Walrand, "Admission control for multi-class ATM traffic with overflow constraints," tech. rep., Department of EECS, University of California, Berkeley, 1994.

[15] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *The Bell System Technical Journal*, vol. 61, pp. 1871–1894, October 1982.

[16] D. D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an Integrated Services Packet Network: Architecture and mechanism," in *Proc. of ACM SIGCOMM '92*, August 1992.

[17] S. Jamin, P. B. Danzig, S. Shenker, and L. Zhang, "A measurement-based admission control algorithm for Integrated Services Packet Networks," in *Proc. of ACM SIGCOMM '95*, pp. 2–13, August 1995.

[18] H. Saito and K. Shiomoto, "Dynamic call admission control in ATM networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 982–989, September 1991.

[19] Z. Dziong, O. Montanuy, and L. G. Mason, "Adaptive bandwidth management

in ATM networks," *International Journal of Communication Systems*, vol. 7, pp. 295-306, 1994.

[20] R. J. Gibbens, F. P. Kelly, and P. B. Key, "A decision-theoretic approach to call admission control in ATM networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1101-1114, August 1995.

[21] H. Saito, "Simplified dynamic connection admission control in ATM networks," in *Proc. of ICCCN '94*, pp. 217-223, 1994.

[22] M. Cheung and J. W. Mark, "Measurement-based connection admission control in ATM networks," in *Proc. of the 18th Biennial Symposium on Communications*, (Queen's University, Kingston, Ontario, Canada), pp. 70-73, June 1996.

[23] M. Cheung, "Measurement-based connection admission control in ATM networks," Master's thesis, University of Waterloo, 1996.

[24] M. Cheung and J. W. Mark, "A hybrid analytical/measurement-based decision-making approach to connection admission control," in *Proc. of the CISS (Conference on Information Sciences and Systems) '98*, (Princeton, NJ, U.S.A), pp. 194-199, March 1998.

[25] D. Yates, J. Kurose, D. Towsley, and M. G. Hluchyj, "On per-session end-

to-end delay distributions and the call admission problem for real-time applications with QoS requirements," in *Proc. of ACM SIGCOMM '93*, pp. 2–12, 1993.

[26] A. M. Viterbi and A. J. Viterbi, "Erlang capacity of a power controlled CDMA system," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 892–900, August 1993.

[27] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," *Wireless Networks*, vol. 3, pp. 29–42, March 1997.

[28] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. on Vehicular Technology*, vol. VT-35, pp. 77–92, August 1986.

[29] E. C. Posner and R. Guérin, "Traffic policies in cellular radio that minimize blocking of handoff calls," in *Proc. of ITC-11*, 1985.

[30] C.-C. Chao and W. Chen, "Connection admission control for mobile multiple-class personal communications networks," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1618–1626, October 1997.

[31] M. Asawa, "Optimal admissions in cellular networks with handoffs," in *Proc. of IEEE ICC '96*, 1996.

[32] Q.-A. Zeng, K. Mulumoto, and A. Fukuda, "Performance analysis of two-level priority reservation handoff scheme in mobile cellular radio systems," in *Proc. of IEEE VTC '97*, pp. 974–978, 1997.

[33] O. T. W. Yu and V. C. M. Leung, "Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1208-1225, September 1997.

[34] S. S. Rappaport and C. Purzynski, "Prioritized resource assignment for mobile cellular communication systems with mixed platforms," *IEEE Trans on Vehicular Technology*, vol. 45, pp. 443–458, August 1996.

[35] D. Calin and D. Zeghlache, "Performance and handoff analysis of an integrated voice-data cellular system," in *Proc. of IEEE PIMRC '97*, pp. 386–39., September 1997.

[36] M. D. Kulavaratharasah and A. H. Aghvami, "Teletraffic performance evaluation of microcellular personal communication networks (PCN's) with prioritized handoff procedures," *IEEE Trans. on Vehicular Technology*, vol. 48, pp. 137–152, January 1999.

[37] P. Ramanathan, K. Sivalingam, P. Agrawal, and S. Kishore, "Dynamic re-

source allocation schemes during handoff for mobile multimedia wireless networks," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1270–1283, July 1999.

[38] A. S. Acampora and M. Naghshineh, "An architecture and methodology for mobile-executed handoff in cellular ATM networks," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 1365–1375, October 1994.

[39] T. C. Wong, J. W. Mark, and K. C. Chua, "Connection admission control in cellular wireless ATM access networks," in *Proc. of IEEE ICC '98*, pp. 1095–1098, June 1998.

[40] D. Levine, I. F. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Trans. on Networking*, vol. 5, pp. 1–12, February 1997.

[41] J. Li, N. B. Shroff, and E. K. P. Chong, "Channel carrying: A novel handoff scheme for mobile cellular networks," *IEEE/ACM Trans. on Networking*, vol. 7, pp. 38–50, February 1999.

[42] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE J. Select. Areas Commun.*, vol. 14, pp. 711–717, May 1996.

[43] A. Sutivong and J. M. Peha, "Performance comparisons of call admission

control algorithms in cellular systems," in *Proc. of IEEE GLOBECOM '97*, pp. 1645–1649, November 1997.

[44] A. Sutivong and J. M. Peha, "Novel heuristic call admission control algorithms for cellular systems: Proposal and comparison," in *Proc. of IEEE ICUPC '97*, October 1997. http://www.ece.cmu.edu/afs/ece/usr/peha/peha.html.

[45] W. Yue and Y. Matsumoto, "Performance analysis of integrated voice/data transmission in slotted CDMA packet radio communication networks," in *Proc. of Globecom '98*, pp. 3288–3294, November 1998.

[46] Y. Ishikawa and N. Umeda, "Capacity design and performance of call admission control in cellular CDMA systems," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1627–1635, October 1997.

[47] T.-K. Liu and J. A. Silvester, "Joint admission/congestion control for wireless CDMA systems supporting integrated services," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 845–857, August 1998.

[48] Z. Liu and M. E. Zarki, "SIR-based call admission control for DS-CDMA cellular systems," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 638–644, May 1994.

[49] M. Andersin, Z. Rosberg, and J. Zander, "Soft and safe admission control in

cellular networks," *IEEE/ACM Trans. on Networking*, vol. 5, pp. 255–265, April 1997.

[50] Z. Dziong, M. Jia, and P. Mermelstein, "Adaptive traffic admission for integrated services in CDMA wireless-access networks," *IEEE J. Select. Areas Commun.*, vol. 14, pp. 1737–1747, December 1996.

[51] S. Sun and W. A. Krzymien, "Call admission policies and capacity analysis of a multi-service CDMA personal communiation system with continuous and discontinuous transmission," in *Proc. of IEEE VTC '98*, pp. 218–223, May 1998.

[52] R.-F. Chang and S.-W. Wang, "QoS-based call admission control for integrated voice and data in CDMA systems," in *Proc. of IEEE PIMRC '96*, pp. 623–627, October 1996.

[53] J. S. Evans and D. Everitt, "Effective bandwidth-based admission control for multiservice CDMA cellular networks," *IEEE Trans. on Vehicular Technology*, vol. 48, pp. 36–46, January 1999.

[54] W.-B. Yang and E. Geraniotis, "Admission policies for integrated voice and data traffic in CDMA packet radio networks," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 654–664, May 1994.

[55] J. Q. J. Chak and W. Zhuang. "Capacity analysis for connection admission control in indoor multimedia CDMA wireless communications," *Wireless Personal Communications*. no. 12. pp. 269-292. 2000.

[56] A. Sampth and J. M. Holtzman. "Access control of data in integrated voice/data CDMA systems: Benefits and tradeoffs," *IEEE J. Select. Areas Commun.*, vol. 15. pp. 1511 1526. October 1997.

[57] D. A. ad Anthony Ephremides. "Cellular multicode CDMA capacity for integrated (voice and data) services," *IEEE J. Select. Areas Commun.*, vol. 17. pp. 928-938. May 1999.

[58] W. Huang and V. K. Bhargava. "Performance evaluation of a DS/CDMA cellular system with voice and data services," in *Proc. of IEEE PIMRC '96*. pp. 588-592. October 1996.

[59] C.-N. Wu. Y.-R. Tsai. and J.-F. Chang. "A quality-based birth-and-death queueing model for evaluating the performance of an integrated voice/data CDMA cellular system," *IEEE Trans. on Vehicular Technology*. vol. 48. pp. 83-89. January 1999.

[60] A. Baiocchi. F. Sestini. and F. D. Priscoli. "Effects of user mobility on a

CDMA cellular network." *European Trans. on Telecommun.*, vol. 7, pp. 305-314, July/August 1996.

[61] F. D. Priscoli and F. Sestini, "Effects of imperfect power control and user mobility on a CDMA cellular network," *IEEE J. Select. Areas Commun.*, vol. 14, pp. 1809-1817, December 1996.

[62] G. J. Foschini, B. Gopinath, and Z. Miljanic, "Channel cost of mobility," *IEEE Trans. on Vehicular Technology*, vol. 42, pp. 414-424, November 1993.

[63] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley, "On the capacity of a cellular CDMA system," *IEEE Trans. on Vehicular Technology*, vol. 40, pp. 303-312, May 1991.

[64] A. J. Viterbi, "The evolution of digital wireless technology from space exploration to personal communication services," *IEEE Trans. on Vehicular Technology*, vol. 43, pp. 638-644, August 1994.

[65] J. S. Evans and D. Everitt, "On the teletraffic capacity of CDMA cellular networks," *IEEE Trans. on Vehicular Technology*, vol. 48, pp. 153-165, January 1999.

[66] F. D. Priscoli and F. Sestini, "Fixed and adaptive blocking thresholds in

CDMA cellular networks," in *Proc. of IEEE ICC '95*, pp. 1090–1093, June 1995.

[67] G. Brussaard, "Erlang capacity of ATM-based CDMA satellite system," *Electronics Letters*, vol. 35, pp. 613–614, April 15 1999.

[68] C. Lee and K. Kim, "Capacity enhancement using an ARQ scheme in a voice/data DS-CDMA system," *Electronics Letters*, vol. 34, pp. 259–261, February 19 1998.

[69] M. Cheung and J. W. Mark, "Effect of mobility on QoS provisioning in wireless communication networks," in *Proc. of the WCNC (Wireless Communications and Networking Conference) '99*, (New Orleans, LA, U.S.A), pp. 306–310, September 1999.

[70] R. Bolla, F. Davoli, and M. Marchese, "Bandwidth allocation and admission control in ATM networks with service separation," *IEEE Communications Magazine*, vol. 35, pp. 130–137, May 1997.

[71] D. Mitra, M. I. Reiman, and J. Wang, "Robust dynamic admission control for unified cell and call QoS in statistical multiplexers," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 692–707, June 1998.

[72] I. Rubin and T. Cheng, "Admission control for multi-layer management of

high-speed packet-switched networks under observation noise," in *Proc. of IEEE INFOCOM '91*, pp. 570–578, 1991.

[73] M. Beshai, R. Kositpaiboon, and J. Yan, "Interaction of call blocking and cell loss in an ATM network," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 1051–1058, August 1994.

[74] M. J. Fischer and T. C. Harris, "A model for evaluating the performance of an integrated circuit- and packet-switched multiplex structure," *IEEE Trans. on Communications*, vol. COM-24, pp. 195–202, February 1976.

[75] G. F. Williams and A. Leon-Garcia, "Performance analysis of integrated voice and data hybrid-switched links," *IEEE Trans. on Communications*, vol. COM-32, pp. 695–706, June 1984.

[76] H. Okada, "Delay behavior of data traffic in an integrated voice/data muliplex structure: Multi-capacity limits (MCL) property," *IEEE Trans. on Communications*, vol. COM-34, pp. 300–303, March 1986.

[77] L.-P. Chin and J.-F. Chang, "Integrated voice/data transmission in a high speed common channel using demand assigned movable-boundary TDMA multiplexer," *Journal of the Chinese Institute of Engineers*, vol. 18, no. 4, pp. 471–480, 1995.

[78] J. W. Mark and G.-L. Wu, “Resource management at an atm network element,” in *Proc. of First Workshop on ATM Traffic Management WATM '95*, pp. 331–338, 1995.

[79] D. Mitra and I. Ziedins, “Virtual partitioning by dynamic priorities: Fair and efficient resource-sharing by several services,” in *Broadband Communications: Proc. 1996 International Zurich Seminar on Digital Communications (IZS '96)*, pp. 173–185, February 1996.

[80] S. C. Borst and D. Mitra, “Virtual partitioning for robust resource sharing: Computational techniques for heterogeneous traffic,” *IEEE J. Select. Areas Commun.*, vol. 16, pp. 668–678, June 1998.

[81] L. Chen, S. Yoshida, and H. Murata, “A dynamic channel assignment algorithm for voice and data integrated TDMA mobile radio,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E80-A, pp. 1204–1210, July 1997.

[82] S. Zhu and J. W. Mark, “Power distribution law and its impact on the capacity of multimedia multirate wideband CDMA systems,” Tech. Rep. CWC'07, Centre for Wireless Communications, University of Waterloo, 1999.

[83] M. M. Zonoozi and P. Dassanayake, “User mobility modeling and characteriza-

tion of mobility patterns," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1239–1252, September 1997.

[84] Y. Fang, I. Chlamtac, and Y.-B. Lin, "Channel occupancy times and handoff rate for mobile computing and PCS networks," *IEEE Trans. on Computers*, vol. 47, pp. 679–692, June 1998.

[85] D. McMillan, "Traffic modelling and analysis for cellular mobile networks," in *Proc. of ITC-13*, 1991.

[86] L. Kleinrock, *Queueing Systems, Volume I: Theory*. John Wiley & Sons, 1975.

[87] M. Schwartz, *Broadband Integrated Networks*. Prentice Hall PTR, 1996.

[88] M. Noll, "Does data traffic exceed voice traffic?," *Communications of the ACM*, vol. 42, pp. 121–124, June 1999.