

# **An Intelligent-Agent Approach for Managing Congestion in W-CDMA Networks**

Soamsiri Chantaraskul

Submitted for the degree of Doctor of Philosophy

Department of Electronic Engineering  
Queen Mary, University of London

August 2005

*To my parents*

# Abstract

Resource Management is a crucial aspect in the next generation cellular networks since the use of W-CDMA technology gives an inherent flexibility in managing the system capacity. The concept of a “Service Level Agreement” (SLA) also plays a very important role as it is the means to guarantee the quality of service provided to the customers in response to the level of service to which they have subscribed. Hence there is a need to introduce effective SLA-based policies as part of the radio resource management.

This work proposes the application of intelligent agents in SLA-based control in resource management, especially when congestion occurs. The work demonstrates the ability of intelligent agents in improving and maintaining the quality of service to meet the required SLA as the congestion occurs.

A particularly novel aspect of this work is the use of learning (here Case Based Reasoning) to predict the control strategies to be imposed. As the system environment changes, the most suitable policy will be implemented. When congestion occurs, the system either proposes the solution by recalling from experience (if the event is similar to what has been previously solved) or recalculates the solution from its knowledge (if the event is new). With this approach, the system performance will be monitored at all times and a suitable policy can be immediately applied as the system environment changes, resulting in maintaining the system quality of service.

# Contents

<b>Abstract</b>	<b>3</b>
<b>Contents</b>	<b>4</b>
<b>List of Figures</b>	<b>7</b>
<b>Acknowledgements</b>	<b>11</b>
<b>Abbreviations</b>	<b>12</b>
<b>Chapter 1 Introduction</b>	<b>15</b>
1.1 Research Motivation	15
1.2 Research Scope	16
1.3 Contributions	16
1.4 Organisation of the Thesis	17
<b>Chapter 2 Cellular Networks</b>	<b>18</b>
2.1 Evolution	18
2.2 Cellular Concept	22
2.3 GSM System Overview	24
2.3.1 GSM Architecture	25
2.3.2 The Radio Interface	26
2.4 Third Generation Cellular System (UMTS)	28
2.4.1 UMTS Structure	29
2.4.2 Services and QoS Classes in UMTS	31
2.4.3 Radio Access Technique in UMTS (W-CDMA)	33
2.5 Radio Resource Management in UMTS	35
2.5.1 Power Control	35
2.5.2 Handover Control	37
2.5.3 Admission Control	39
2.5.4 Load Control (Congestion Control)	41
2.5.5 Packet Scheduling	42
2.6 Previous Works of RRM in UMTS	42
2.7 Summary	43

<b>Chapter 3 Agent-Based SLA Management</b>	<b>44</b>
3.1 Service Level Agreement	44
3.2 Multiagent System in Resource Management	47
3.2.1 Agents and Multi-Agent Systems	47
3.2.2 Agents and Resource Management	51
3.2.3 Functional Architecture	53
3.2.4 Internal Agent Architecture	55
3.2.5 NPRA	56
3.3 Case-Based Reasoning Approach	57
3.3.1 Why use CBR?	58
3.3.2 CBR Process Model	59
3.3.3 Cases and Retrieving Algorithm	61
3.4 CBR in SLA-Based Resource Control	63
3.5 Summary	64
<b>Chapter 4 Simulation Model and Validation</b>	<b>65</b>
4.1 Traffic Model	68
4.1.1 Voice Traffic	69
4.1.2 Video Traffic	69
4.1.3 Data Traffic	71
4.2 Radio Propagation Model	72
4.3 Receiver Model	73
4.4 Power Control Model	73
4.4.1 Open-Loop Power Control	73
4.4.2 Inner-Loop Power Control	74
4.5 Assignment and Admission Scheme	76
4.6 CBR Model	78
4.7 Verification and Validation	78
4.7.1 Introduction	78
4.7.2 Introduction of the relevant model	79
4.7.3 Validation and discussion	81
4.8 Summary	82
<b>Chapter 5 Monitoring and Congestion Pattern Recognition</b>	<b>84</b>
5.1 Simulation Scenarios and SLA Assumption	84

5.1.1	Random Overload Cases	84
5.1.2	Hotspot Cases	85
5.2	Monitoring Process	86
5.3	Case Matching for Congestion Pattern Recognition	87
5.3.1	Method for Case Matching in Hotspots	89
5.3.2	Chosen Case Matching Method	92
5.4	Summary	95
<b>Chapter 6 Simulation Results and Analysis</b>		<b>96</b>
6.1	Simulation of Changing Reactive Layer Policy	96
6.2	Determining Steady-State of the System	101
6.3	Simulation of the SLA-Based Control System Using CBR Approach	102
6.3.1	Generating Cases	102
6.3.2	Result from the matching of the same case as existing ones in the library	114
6.4	Examining the System Performance under Similar Congestion Pattern to the Existing Case	115
6.4.1	First experiment for the similar case that has additional congestion area across bands	115
6.4.2	Second experiment for the similar cases that has additional congestion area across segments	116
6.5	Examining the System Performance under Unfamiliar Congestion Pattern	118
6.5.1	Setting up rules for the calculation method	118
6.5.2	Cell Shrinking Method	121
6.5.3	Buffer Mechanism	123
6.5.4	Hybrid Method	124
6.6	Summary	126
<b>Chapter 7 Conclusions</b>		<b>128</b>
7.1	Conclusions	128
7.2	Further Work	128
<b>Author's Publications</b>		<b>130</b>
<b>References</b>		<b>131</b>

# List of Figures

Figure 2.1	Mobile represents over half of the UK market (from [Oftel04])	19
Figure 2.2	The proposed new GERAN architecture [HNPR01]	21
Figure 2.3	Cellular evolution	22
Figure 2.4	Cellular frequency reuse in GSM	23
Figure 2.5	Frequency bands for GSM from GSM MOU Association: <a href="http://www.gsmworld.com/technology/spectrum/frequencies.shtml">http://www.gsmworld.com/technology/spectrum/frequencies.shtml</a>	25
Figure 2.6	GSM system architecture (based on [LC01])	25
Figure 2.7	Multiple access techniques	27
Figure 2.8	The FDMA/TDMA structure of GSM [Heine98]	27
Figure 2.9	GSM-TDMA frame structure	28
Figure 2.10	Spectrum allocation in different countries	29
Figure 2.11	Overall UMTS architecture	30
Figure 2.12	UTRAN architecture [3GPP01]	31
Figure 2.13	UMTS QoS architecture	32
Figure 2.14	Cellular frequency reuse in GSM vs. UMTS	33
Figure 2.15	Spread spectrum process	34
Figure 2.16	Principle of spread spectrum	34
Figure 2.17	Interference in CDMA	35
Figure 2.18	Power control techniques illustrated as uplink power control	36
Figure 2.19	Types of handover [Chen03]	38
Figure 3.1	Principle of new business model	44
Figure 3.2	SLA types as applied here	46
Figure 3.3	(a) The TouringMachines architecture [Fer95] (b) The INTERRRAP architecture [MPT95]	50
Figure 3.4	Benefits of using an agent control system for mobile networks (from [Bod00])	52
Figure 3.5	Illustration of functional and agent architecture adopted by SHUFFLE (from [CRTBB01])	54
Figure 3.6	General agent structure	56
Figure 3.7	NPRA internal architecture	57
Figure 3.8	Case-based reasoning process model (Based on the CBR cycle	

	in [AP94])	60
Figure 3.9	A Shared-Feature Network	62
Figure 3.10	NPRA (with CBR) internal architecture	63
Figure 4.1	Simulation Model	65
Figure 4.2	Simulation process	66
Figure 4.3	Traffic model for voice call (ON-OFF model)	69
Figure 4.4	Video source model (Discrete-state continuous time Markov process)	70
Figure 4.5	Components of Data Traffic [Tri01]	72
Figure 4.6	Power control process	74
Figure 4.7	Effect of different power control time step on simulation time usage	75
Figure 4.8	Effect of different power control time step on the call blocking rate	76
Figure 4.9	Connection admission control process	77
Figure 4.10	Comparison between the validating result from the simulation model and the results from [CR01]	82
Figure 5.1	Hotspot cases illustrated by hotspot cell layouts	85
Figure 5.2	Example of hotspot case 1 and case 6 as scenarios were generated by a simulation model	86
Figure 5.3	Monitoring process	87
Figure 5.4	Case matching process using CBR approach	87
Figure 5.5	Case library structure according to shared-feature network	88
Figure 5.6	Hotspot cell monitoring areas	89
Figure 5.7	Hotspot case identification represented by ellipsoids	90
Figure 5.8	The effect on collecting different sample size in order to obtain ellipsoid parameters	91
Figure 5.9	The layout of three different hotspot cases	91
Figure 5.10	Ellipsoids represent congestion pattern according to the hotspot cell area of congestion	92
Figure 5.11	A two-step hotspot pattern identifying method	93
Figure 5.12	Comparison of the offered traffic (kbit/s) plot between the system under monitoring time 10 s and 60 s	94
Figure 5.13	Case library structure	94
Figure 6.1	The simulation result from conventional system	97
Figure 6.2	Simulation results showing the effect of policy change	98

Figure 6.3	Comparison between conventional system and the one with SLA-based control	100
Figure 6.4	Throughput plot over simulation time	101
Figure 6.5	Simulation result from the conventional system for random overload cases	103
Figure 6.6	Simulation result showing the effect of applying new reactive layer policy for the first random overload case	104
Figure 6.7	Simulation result showing the effect of applying new reactive layer policy for the second random overload case	104
Figure 6.8	Simulation result showing the effect of applying new reactive layer policy for the third random overload case	105
Figure 6.9	Simulation result from the conventional system for the hotspot case 1	106
Figure 6.10	(a) Hotspot case 2 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 2	107
Figure 6.11	(a) Hotspot case 4 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 4	108
Figure 6.12	Transferring method as centre cell shrinks	109
Figure 6.13	(a) Hotspot case 1 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 1	110
Figure 6.14	(a) Hotspot case 3 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 3	111
Figure 6.15	(a) Hotspot case 5 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 5	112
Figure 6.16	Simulation results of applying the new reactive layer policy to the neighbouring cell for hotspot case 5	112
Figure 6.17	(a) Hotspot case 6 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 6	113
Figure 6.18	Simulation results of applying the new reactive layer policy to the neighbouring cell for hotspot case 6	113
Figure 6.19	(a) Hotspot case 1 layout (b) Simulation result showing of the hotspot cell for the case matching compared with previous result	114
Figure 6.20	(a) Hotspot cell layout (b) Hotspot cell layout of the matched case	

	(c) Result for the similar case with additional congestion area across band	115
Figure 6.21	(a) Hotspot cell layout (b) Hotspot cell layout of the matched case (c) Result for 5% of segment is an additional congestion area across segments	116
Figure 6.22	(a) Hotspot cell layout (b) Simulation result for 25% of segment is an additional congestion area across segments	117
Figure 6.23	(a) Hotspot cell layout (b) Hotspot cell layout of the matched case (c) Result for a repeated simulation by matching with the new case	118
Figure 6.24	Existing cases and their solutions	119
Figure 6.25	Summary of the rule-based algorithm for the calculation method	119
Figure 6.26	(a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – cell shrinking method	121
Figure 6.27	(a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – cell shrinking method	122
Figure 6.28	(a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – cell shrinking method	123
Figure 6.29	(a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – buffer mechanism	124
Figure 6.30	(a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – Hybrid method	125
Figure 6.31	(a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – Hybrid method	125
Figure 6.32	(a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – Hybrid method	126

# Acknowledgements

I would like to take this opportunity to express my sincere gratitude to my supervisor, Prof. Laurie Cuthbert, for his supervision and continuous support. His advice has been of greatest help at all times and his encouragement has always been invaluable throughout my study. Also I would like to thank Prof. Jonathan Pitts, Dr. John Schormans, Dr. John Bigham, Dr. Chris Phillips, Dr. Eliane Bodanese, and Dr. Yue Chen for their guidance.

I wish to extend my gratitude to all the support staff of the department, who have provided a great help, especially Lynda Rolfe. Many thanks for my colleagues and friends who have encouraged and advised me throughout my study. They have made my time here memorable. Finally, I would like to thank my parents for their love and enormous support. They always be my strongest power to work hard and get through hard times.

I would also like to acknowledge that the background concept of the agent architecture used in this work from the IST project IST-1999-11014 (SHUFFLE) that finished in 2002, the work on applying SLAs and in using learning techniques as described in this thesis is wholly that of the author and was not part of that project.

# Abbreviations

2G	Second Generation (mobile)
3G	Third Generation (mobile)
3GPP	3G Partnership Project
AI	Artificial Intelligence
AM	Amplitude modulation
AMR	Adaptive Multi-Rate
ARIB	Association of Radio Industries and Businesses
ATIS	Alliance for Telecommunications Industry Solutions
AuC	Authentication Centre
BER	Bit Error Rate
BLER	Block Error Rate
BMP	Broadcast/Multicast Control
BSC	Base Station Controller
BSS	Base Station System
BTS	Base Transceiver Station
CBC	Common Broadcast Centre
CBR	Case-Based Reasoning
CC	Call Control
CCSA	The China Communications Standards Association
CIR	Carrier-to-Interference Ratio
CM	Communication Management
CN	Core Network
CS	Circuit Switched
DAMPS	Digital Advanced Mobile Phone System
DECT	Digital Enhance Cordless Telephone
EDGE	Enhanced Data Rates for Global/GSM Evolution
EFR	Enhance Full-Rate
EGPRS	Enhanced GPRS
EIR	Equipment Identity Centre
ECSD	Enhanced Circuit-Switch Data
ETSI	European Telecommunications Standard Institute

FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FL	Fuzzy Logic
FM	Frequency modulation
GGSN	Gateway GPRS Support Node
GMSC	Gateway MSC
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
HLR	Home Location Register
HR	Half-Rate
HSCSD	High Speed Circuit-Switch Data
IMT-2000	International Mobile Telephony 2000
IS-95	cdmaOne, one of the 2 <sup>nd</sup> generation systems, mainly in America and in Korea
ITU	International Telecommunication Union
MAC	Medium Access Control
ME	Mobile Equipment
MM	Mobility Management
MPLS	Multi-Protocol Label Switching
MS	Mobile Station
MSC	Mobile service Switching Centre
MSS	Mobile Satellite System
MT	Mobile Terminal
NADC	North American Digital Communication
NN	Neural Network
NP	Network Provider
NPRA	Network Provider Resource Agent
NSS	Network and Switching Subsystem
OSI	Open System Interconnect
PC	Power Control
PCS	Personal Communication Systems
PDCP	Packet Data Convergence Protocol
PHS	Personal Handy phone System
PIN	Personal Identity Module

PS	Packet Switched
QoS	Quality of service
RAB	Radio Access Bearer
RLC	Radio Link Control
RNC	Radio Network Controller
RNP	Radio Network Planning
RNS	Radio Network Subsystem
RRC	Radio Resource Control
SGSN	Service GPRS Support Node
SIM	Subscriber Identity Module
SIR	Signal-to-Interference Ratio
SLA	Service Level Agreement
SLM	Service Level Management
SLS	Service Level Specification
SMS	Short Message Services
SP	Service Provider
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TE	Terminal Equipment
TIA	Telecommunications Industry Association
TP	Internet Protocol
TPC	Transmit Power Control
TTA	Telecommunications Technology Association
TTC	Telecommunication Technology Committee
UMTS	Universal Mobile Telecommunication System
USDC	United States Digital Communication
USIM	UMTS Service Identity Module
UT	User Terminal
UTRAN	UMTS Radio Access Network
VLR	Visitor Location Register
WARC	World Administrative Radio Conference
WRC	World Radio communication Conference
W-CDMA	Wideband Code Division Multiple Access

# Chapter 1 Introduction

## 1.1 Research Motivation

In TDMA (Time Division Multiple Access) second generation (2G) cellular systems, the system capacity, or the number of users, is governed by the number of timeslots available. Therefore, the decision whether to admit a call request is concerned only with the timeslot (or resource) availability: in most 2G situations, each call has the same bandwidth demand<sup>1</sup> so it is merely a matter of checking whether a spare timeslot is available. In well-planned 2G systems, the co-channel interference arising from other transmitters using the same frequency is a second or third order effect.

With W-CDMA (Wideband Code Division Multiple Access) being used for the third generation cellular networks (3G networks), the system capacity becomes more flexible since all users share the same spectrum allocation and use codes to identify themselves from others. Hence the whole bandwidth can be reused in every cell and the system capacity is limited by the total interference that occurs from other users (in the case of the network being uplink - capacity limited) or other base stations (in the case of the network being downlink - capacity limited) and the background noise. Hence, providing the flexible, higher bandwidth services, and maintaining the best system capacity leads to more complexity in Radio Resource Management (RRM).

Playing a crucial part in 3G networks, RRM controls the system capacity and affects the management of service quality. Congestion is a major event that leads to a deterioration of system Quality of Service (QoS). When congestion occurs, action(s) needs to be taken as part of load control algorithm.

The concept of a Service Level Agreement (SLA) is becoming one of the main interests in 3G networks as an SLA allows network providers to offer different levels of service guarantees to different customers (who are paying different rates). In other words, customers' QoS should be guaranteed according to the SLA they have made with the provider.

---

<sup>1</sup> GPRS is an exception since it is packet based; however it is implemented by assigning timeslots to the total GPRS demand and there is no real resource management for individual GPRS users.

The motivation of this research comes from the requirement to introduce efficient congestion management by offering SLA-based policies to maintain the QoS as guaranteed in the SLA, especially when congestion occurs. The exploitation of intelligent agents in SLA-based control is introduced in this work. Employing multilayer agent technology offers different decision timescales for resource management. In a normal traffic situation, the decision can be made immediately, but as the circumstances change, the Case-Base Reasoning (CBR) approach proposed here (as part of the agent system) can identify the congestion pattern and find the best policy (RRM configuration).

## **1.2 Research Scope**

This work is the first to investigate a learning approach, coupled with an intelligent agent system, to manage the resources to keep within SLA boundaries when congestion occurs.

To investigate this, a detailed model of the system had to be created that allowed the system to be monitored when congestion occurs. Initially, a homogeneous congestion pattern was used, but this has then been expanded to include a range of congestion patterns.

Although a variety of learning techniques could have been used, CBR was chosen for the reasons given in the body of this thesis. This has been employed to recognise congestion situations as they occur and to get the RRM to execute appropriate action.

## **1.3 Contributions**

The author's papers are listed on page 130.

The main contributions in this work are:

- A detailed modelling of the system behaviour that implements intelligent agents and the CBR learning approach.
- A CBR implementation in RRM that recognises congestion and takes action to manage the load on the network.

- A novel application of the CBR approach to allow fast convergence of the agent policies and the calculation method used in the cases where an unfamiliar situation occurs.

It should be noted that while the background concept of the agent architecture used in this work came from the IST project IST-1999-11014 (SHUFFLE) that finished in 2002, the work on applying SLAs and in using learning techniques as described here is wholly that of the author and was not part of that project.

## **1.4 Organisation of the Thesis**

In Chapter 2, background information about the cellular networks used in the research is presented. This chapter also gives an overview of 3G cellular networks and also focuses on the resource management method being used in the system modelling.

Chapter 3 is dedicated to the background review of service level agreements including the definition, the problem and SLA management challenges. The introduction of multi-agent systems in resource management and CBR for learning is explained in this chapter.

Chapter 4 presents the simulation model being used in the research. It consists of traffic model, radio propagation model, receiver model, power control model, assignment and admission scheme, and CBR model. The important processes of validation and verification, essential when using simulation-based research, are also described in this chapter.

Chapter 5 illustrates the simulation scenarios that are used as the basis for the experiments here, including the monitoring process, case matching algorithm and the calculation method.

Chapter 6 gives the numerical results that show the performance management of the SLA-based control system and provides discussions on the results.

Chapter 7 presents the conclusions and potential future work.

# Chapter 2 Cellular Networks

## 2.1 Evolution

The idea of using cells for wireless communication originated in 1947 from the US AT&T Bell Laboratories and was published in 1979, the first system being developed in Tokyo, Japan. [Parry02] That network was operated by Nippon Telephone and Telegraph company (NTT) using 600 FM duplex channels (25 kHz for each one-way link). [LWN02] These first generation (1G) cellular systems (of which the Nordic Mobile Telephone System (NMT) and the American Mobile Phone System (AMPS) are examples) are also known as analogue cellular systems because frequency modulation was used with no digital coding.

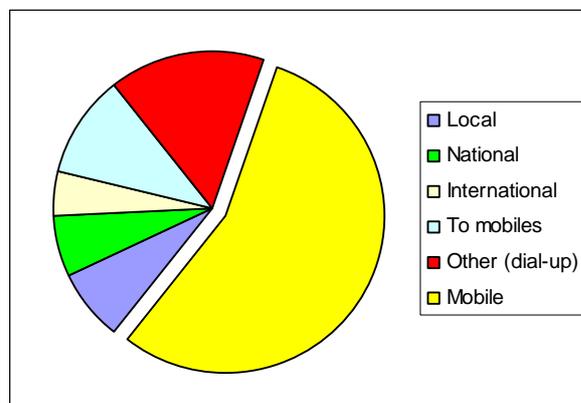
In 1G mobile, the focus was on speech and basic mobility, but the networks could also support data communication albeit with a very low data rate (up to 2.4 kbit/s). Because of the low data rate, it was only suitable for a few applications (such as paging, reading text email and downloading small files), although it has to remember that, at the time, the data rate offered by wired modems was of the same order of magnitude.

In addition to the low bit rate for data applications, the most serious limitation of 1G networks was the lack of substantial roaming capability, apart from in the Nordic countries. This was exacerbated by the different frequencies and communication protocols being used.

As mobile communications became more popular, the need for global roaming arose and this was one of the main drivers in the second generation mobile systems. Other drivers were overcoming the capacity limitation in 1G networks with lower cost and better speech quality, as well as the need for privacy, security and measures to stop cloning. With large-scale integrated circuit technology, digital communications became more practical and more economic than analogue technology. The second generation, 2G, mobile systems began to be deployed in the early 1990s with GSM being introduced in Europe and later becoming a *de facto* world standard (apart from a few countries). GSM and mobile networks in general have been a phenomenal success. In 2004, there were well over a billion mobile customers worldwide with

around 70% of them using GSM [GSM04] and in the UK in 2004, mobile represented over half of the annual US\$29 billion call-revenue market (Figure 2.1 and [OfTel04]).

In this generation, more advanced mobility and a greater variety of services were implemented to solve the problems of the first generation, with data service being supported up to 9.6 kbit/s. This system also solved the incompatibility problem as European countries dedicated the same frequency bands for cellular telephone service and this was followed by most of the world. GSM900 was introduced first in Europe and followed by GSM1800 to increase the system capacity. GSM has then been adopted in the US, albeit on a different frequency band, GSM1900 [LWN02].



**Figure 2.1 Mobile represents over half of the UK market (from [OfTel04])**

Dramatic changes in services and networks happened during the 1990s. The demand from users increased radically, including the demand for using wireless Internet access through mobile terminals. As a result of this overall increase in demand, congestion in radio traffic became more of a problem, leading to network operators having to install more and more base stations. However, there was still the problem that the very nature of GSM meant that really high bit rate services are not possible. Hence, the third generation (3G) cellular system (also called the Universal Mobile Telecommunication System (UMTS)), has been introduced with the aim of (a) being spectrally more efficient and (b) allowing more Internet-like services at higher bit rates [LWN02] [KALNN01].

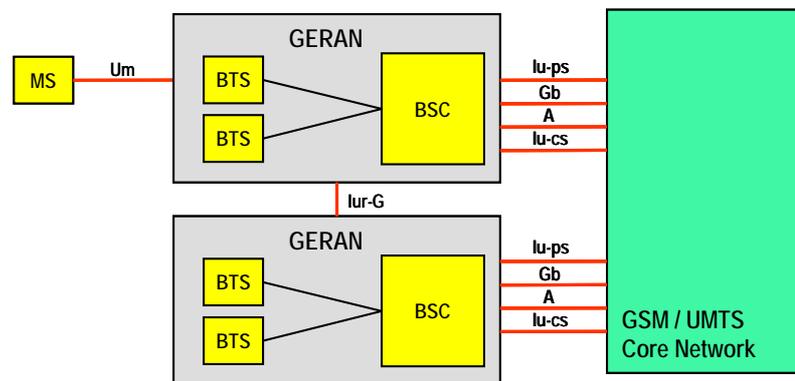
The important evolutionary path from 2G to 3G should be considered here as GSM has been highly successful. The so called 2.5G technologies were released to provide more satisfactory Internet service by offering higher data rate and uninterrupted web access while making voice calls. Examples of 2.5G technologies are General Packet Radio Service (GPRS), High Speed Circuit-Switch Data (HSCSD), and Enhanced Data rate for GSM Evolution (EDGE).

HSCSD and GPRS were introduced first to increase the user data rate by allowing a mobile station to access more than one timeslot per TDMA frame. HSCSD is a circuit-switched technology, which offers a data channel limited to a single 64 kbit/s circuit while GPRS delivers packet-based service over GSM networks. The GPRS radio channel is flexible: ranging from one to eight radio interface timeslots in a TDMA frame and delivering data bit rates range from 9 kbit/s up to more than 150 kbit/s per user.

EDGE was introduced in 1999 to offer more efficient modulation, coding, and retransmission schemes, so allowing enhanced data rates and better overall system capacity. EDGE is designed as an add-on to enhance the existing service such as GPRS and HSCSD as the technique can be used to transmit both packet-switched and circuit-switched voice and data services. In general, for the higher bit rate service, a maximum bit rate of 521 kbit/s is achievable for users moving with speeds up to 100 km/h and up to 182 kbit/s for speeds between 100-500 km/h and 4.7 Mbit/s for an indoor environment. For packet-switched data, EDGE can be introduced as a packet-switched enhancement for GPRS, known as Enhanced GPRS or EGPRS. For the circuit-switched data enhancement, it is known as Enhanced Circuit-Switch Data (ECSD) [SLG01].

New radio access network architecture, GSM/EDGE Radio Access Network (GERAN), has been developed in the 3GPP Release 5 in 2002 based on GSM/EDGE technologies to support UMTS networks. Figure 2.2 shows the GERAN proposed architecture. It has been designed to be fully integrated with the UTRAN (Universal Terrestrial Radio Access Network) of the UMTS system and can be connected to the UMTS core network through the Iu interface. The two legacy interfaces are Gb interface (interface between BSS and SGSN for EGPRS) and A interface (interface between BSS and MSC for ECSD).

3GPP (The Third Generation Partnership Project) was established in December 1998 as a collaboration agreement. The six Organizational Partners are ARIB (Association of Radio Industries and Businesses), CCSA (The China Communications Standards Association), ETSI (European Telecommunications Standard Institute), ATIS (Alliance for Telecommunications Industry Solutions), TTA (Telecommunications Technology Association), and TTC (Telecommunication Technology Committee). It is developing technical specifications for IMT-2000, the International Telecommunication Union's (ITU) framework for third-generation standards.



**Figure 2.2 The proposed new GERAN architecture [HNPR01]**

UMTS networks offer better service performance by using an advanced access technology called wideband CDMA (Code Division Multiple Access). The system is designed for multimedia communication and integrated services. The two IMT-2000 radio interface proposals are W-CDMA (ARIB/ETSI) and cdma2000 (TIA). Since this thesis focuses on W-CDMA 3G networks, more details are given later, in §2.4. To summarise this section, Figure 2.3 shows the cellular evolution.

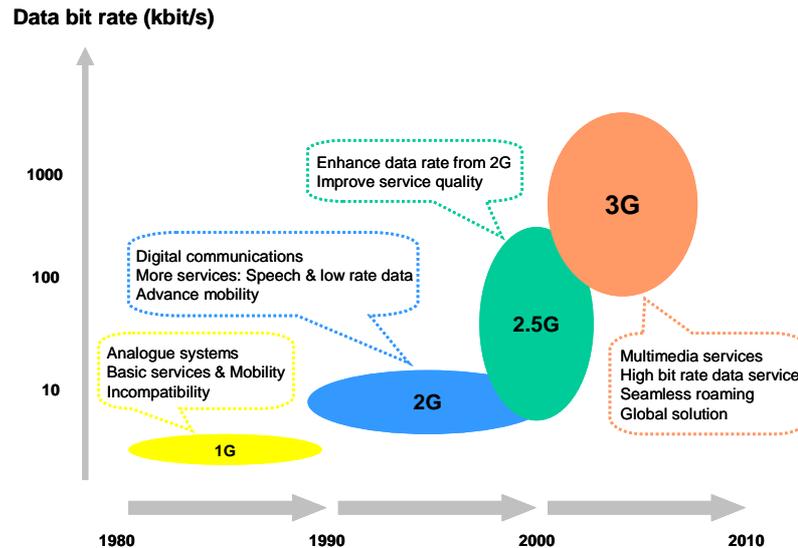


Figure 2.3 Cellular evolution

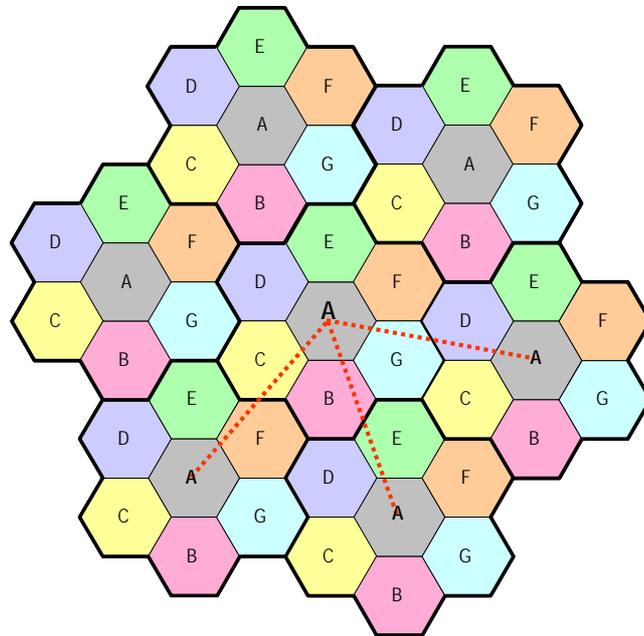
## 2.2 Cellular Concept

As mentioned in §2.1, the cellular concept was published in 1979. The motivation came from the growing demand for wireless communications in a wide area, together with the limitation of frequency spectrum. The main aim was to provide an efficient use of spectrum and at the same time expanding the coverage area by allowing spectrum to be reused. It was also designed to handle mobility, as users are able to roam from cell to cell and the connection will be handed over to maintain the communication.

The basic concept came from the fact that as the distance between transmitter and receiver increases, the signal strength decays. Hence, same frequency can be reused again without causing interference to the existing communication as long as the distance between the two transmitters is sufficient. From this idea, the whole area can be divided into small hexagonal cells (as it is the nearest shape to the ideal circular, radiation pattern that does not have any overlap or gaps between cells). A number of cells,  $N$ , are clustered and the whole bandwidth is evenly distributed among  $N$  cells. The range of  $N$  can be calculated using equation below.

$$N = i^2 + ij + j^2 \quad (2-1)$$

Where,  $i$  and  $j$  are positive integers. This gives the typical value of  $N$  to be 1, 3, 4, 7, 9, 12, 19 etc. A cluster of seven cells is commonly used in the UK in GSM systems [Far96]. Figure 2.4 illustrates the frequency reuse in GSM.



**Figure 2.4 Cellular frequency reuse in GSM**

From Figure 2.4, the cells that have the same colour and name (A-F) utilise the same frequency; they are called co-channel cells. The distance between the two neighbouring co-channel cells,  $D$ , or the so-called *reuse distance* ([SLG01]) can be calculated as follows.

$$\frac{D}{R} = \sqrt{3N} \quad (2-2)$$

Where,  $R$  is the cell radius and  $N$  is a number of cells in a cluster.

In actual cell planning, a vital concern is the trade-off between co-channel interference and the capacity of the frequency reuse. As the number of cells in a cluster increases, less interference will be obtained as the reuse distance becomes longer, but the capacity will be decreased since the spectrum available in each cell will be reduced.

A number of technologies have been developed in order to increase the capacity; these include techniques such as cell splitting and sectoring [Rap96]. A number of

smaller cells are created within an ordinary cell for the cell splitting technique, which means that more base transceiver stations will be located in a cell and hence the existing spectrum can be reused more often, resulting in increasing capacity.

For the sectoring technique, several directional antennas are utilised to replace a single omni-directional antenna; each antenna radiates within a particular section of the cell. Usually, a cell is partitioned into three 120° sectors or six 60° sectors. By doing so, the co-channel interference can be decreased as the interference only comes from the co-channel cells that have antennas radiating toward. Hence, the system capacity is increased without reducing the transmit power. Although techniques such as these provide higher capacity, the cell design can be more complicated and also the hand-off rate is increased.

### **2.3 GSM System Overview**

GSM is a digital wireless network standard that was initially standardized by the European Telecommunications Standard Institute (ETSI). The system needed to achieve certain basic requirements including quality of service, integration of voice and data service, security, good radio frequency utilization and, finally, low cost.

GSM mainly supports voice service but it also supports low bit rate data service (9.6 kbit/s) and Short Message Services (SMS). It has become a very successful system as it achieved 60% market share in 2001 and climbed up to 70% in 2004 [Parry02]. The original frequency band for GSM was 890 MHz to 915 MHz for the uplink and 935 MHz to 960 MHz for the downlink, with the carrier separation of 200 kHz. Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) are used together as the multiple access technique.

Frequency bands now in use are:

GSM Frequencies	
Frequency	Range
GSM400	450.4 - 475.6 MHz paired with 460.4 - 467.6 MHz or 478.8 - 486 MHz paired with 488.8 - 496 MHz
GSM850	824 - 849 MHz paired with 869 - 894 MHz
GSM900	880 - 915 MHz paired with 925 - 960 MHz
GSM1800	1710— 1785 MHz paired with 1805 - 1880 MHz
GSM1900	1850 - 1910 MHz paired with 1930 - 1990 MHz

Figure 2.5 Frequency bands for GSM from GSM MOU Association [GSM05]

### 2.3.1 GSM Architecture

Figure 2.6 shows the GSM system architecture, which consists of three main sections.

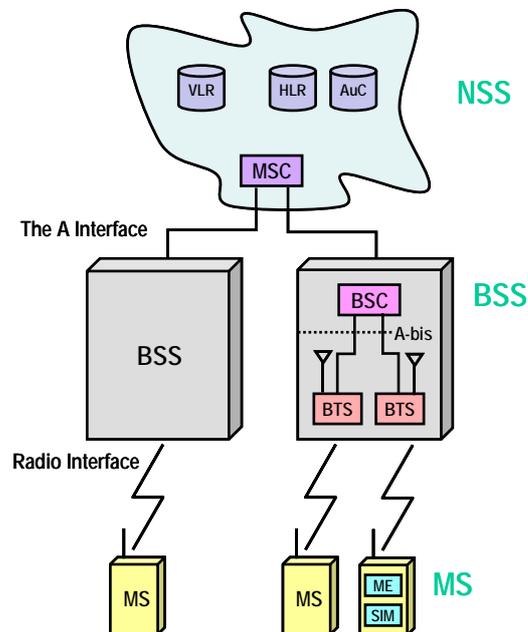


Figure 2.6 GSM system architecture (based on [LC01])

- **Mobile Station (MS)** consists of 2 parts: the **Mobile Equipment (ME)**, which is the portable terminal, and the **Subscriber Identity Module (SIM)**

that contains subscription and authentication information about that particular user.

- **Base Station System (BSS)** consists of the **Base Transceiver Station (BTS)** which contacts MSs through the radio interface and the **Base Station Controller (BSC)** which controls a group of BTSs.
- **Network and Switching Subsystem (NSS)**: The main function of the NSS is the **Mobile Switching Centre (MSC)**. Each MSC serves a number of BSCs and implements the basic switching function and handles calls between the mobile network and other networks (other mobile networks as well as fixed networks). The **Home Location Register (HLR)** and **Visitor Location Register (VLR)** are the databases that maintain the current location of each MS as well as information about particular customers. An **Authentication Centre (AuC)** is used to authenticate subscribers and it provides the HLR, the authentication parameters and ciphering keys.

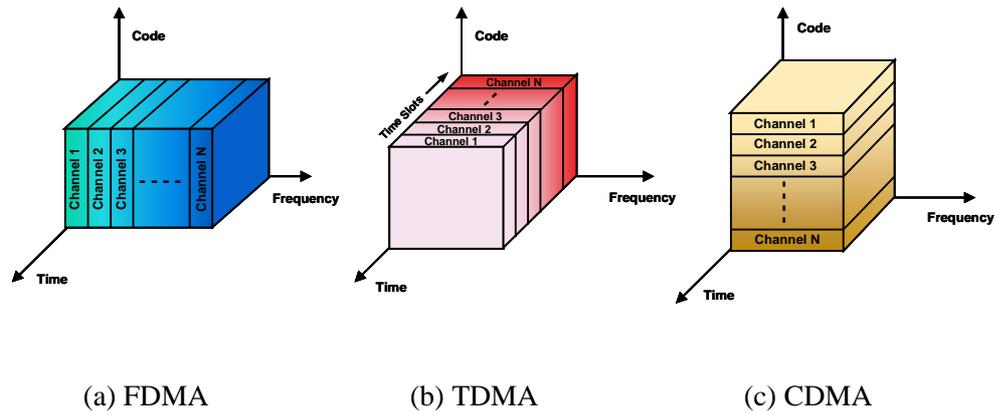
It should be noted that the BSS connects the MS to the NSS.

### ***2.3.2 The Radio Interface***

The radio interface (air interface) is generally the crucial part in a mobile system because the resources are limited by the spectrum available. Naturally it also needs to be standardised and one of the major achievements of GSM is the global acceptance of the standard so that roaming is feasible.

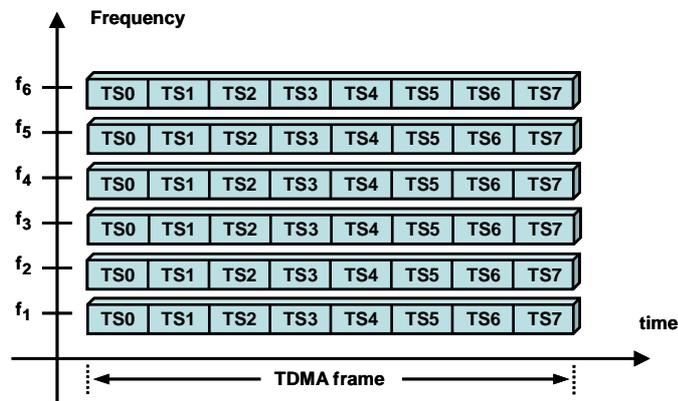
The radio interface also determines the spectrum efficiency.

Figure 2.7 illustrates the comparison between different access techniques: TDMA, FDMA (which, as explained previously, are used together in GSM) and CDMA (that is used in 3G ).



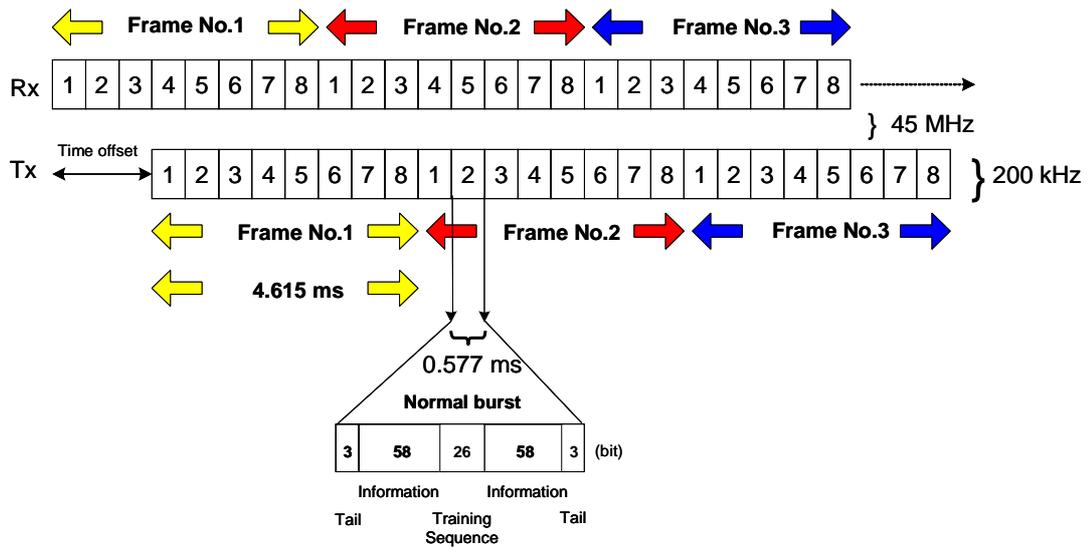
**Figure 2.7 Multiple access techniques**

In FDMA system, an individual user occupies a particular frequency channel; in TDMA, the radio spectrum is divided into time slots and only one user is allowed to transmit or receive in each slot; and in CDMA all users share the same frequency and may transmit simultaneously, but each has a unique code that is used to identify the communication at the receiving end. More detail on CDMA is given in §2.4.3.



**Figure 2.8 The FDMA/TDMA structure of GSM [Heine98]**

Figure 2.8 shows the FDMA/TDMA structure used in GSM systems. By using FDMA, 25MHz in both the uplink and downlink frequency bands are divided into a number of carriers with the carrier separation being 200 kHz. In each carrier, TDMA is applied by dividing the whole 200 kHz in each carrier into time slots. Eight time slots together become one time frame. Figure 2.9 shows the GSM-TDMA frame structure.



**Figure 2.9 GSM-TDMA frame structure**

As shown in Figure 2.9, the number of active users is limited by the number of carriers and also by the co-channel interference (since the transmitters that use the same frequency can directly interfere with each other), although good radio planning should ensure that interference is a low-order effect. In a congestion situation, new call arrivals will be rejected if there is no available bandwidth.

## 2.4 Third Generation Cellular System (UMTS)

UMTS is the common name used in Europe and by 3GPP for the 3G mobile system. According to a study released by In-Stat/MDR ([Mobile03]), 3G services increased rapidly in 2004 and will increase even more rapidly in period between now and 2009 based on the number of chip sets being manufactured.

UMTS uses W-CDMA (Wideband-CDMA) as a multiple access technology in the frequency band around 2 GHz. The actual frequencies bands are 1885-2025 MHz and 2110-2200 MHz and were allocated in March 1992 at the World Administrative Radio Conference (WARC) of the ITU. Figure 2.10 illustrates the spectrum allocation in different countries.

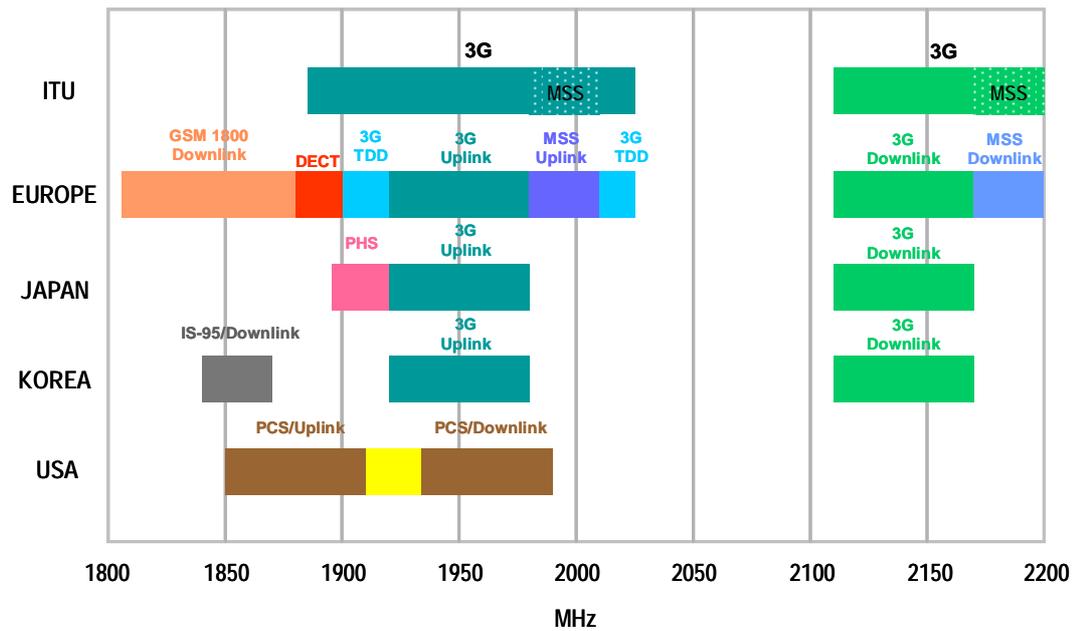


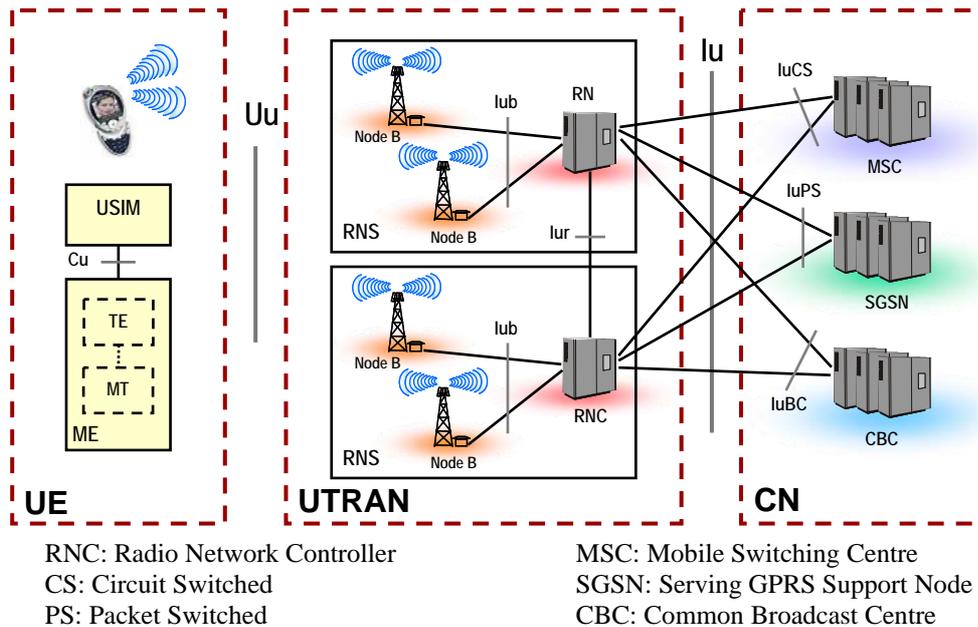
Figure 2.10 Spectrum allocation in different countries

Later on in the World Radio communication Conference 2000, the following bands were allocated for use by IMT-2000 [HT02].

- 1710-1885 and 2500-2690 MHz
- 806-960 MHz for mobile service on a primary basis

### 2.4.1 UMTS Structure

In the 3GPP specification TR 21.905, the 3G network can be divided into two *strata* according to the protocol and their area of responsibility. The first one is the *access stratum*, which contains the protocol handling activities between the User Equipment (UE) and access network. The other is the *non-access stratum*, which contains protocols handling activities between the UE and Core Network (CN) (circuit-switched or packet-switched). Figure 2.11 shows the general architecture of UMTS [HT02] [KALNN01].

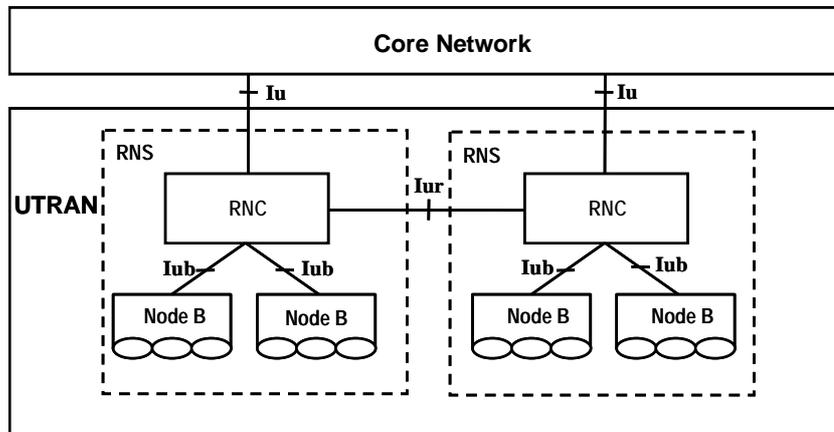


**Figure 2.11 Overall UMTS architecture**

The system architecture consists of three main functions:

- **UE (User Equipment):** the 3G network terminal, which contains two separate parts: **Mobile Equipment (ME)** and **Universal Subscriber Identity Module (USIM)**. The interface between the USIM and the ME is called Cu. The ME domain can be subdivided into several entities typically consists of **Mobile Terminal (MT)**, which performs radio transmission and related functions, and the **Terminal Equipment (TE)**, which contains the end-to-end applications.
- **UTRAN (Universal Terrestrial Radio Access Network):** the subsystem controlling the wideband radio access, W-CDMA. Radio Access Bearers (RAB) for the communication between UE and CN are created and maintained in the UTRAN. Referring to [3GPP01], the UTRAN consists of a set of **Radio Network Subsystems (RNS)** connected to the core network through the Iu interface. Figure 2.12 illustrates the UTRAN architecture [3GPP01]. There are three types of Iu interface, Iu-CS used when connecting to the circuit-switched domain; Iu-PS when connecting to the packet-switched domain, and Iu-BC for the common broadcast centre. The UTRAN

connects to a UE through the Uu interface, which is W-CDMA radio interface.



**Figure 2.12 UTRAN architecture [3GPP01]**

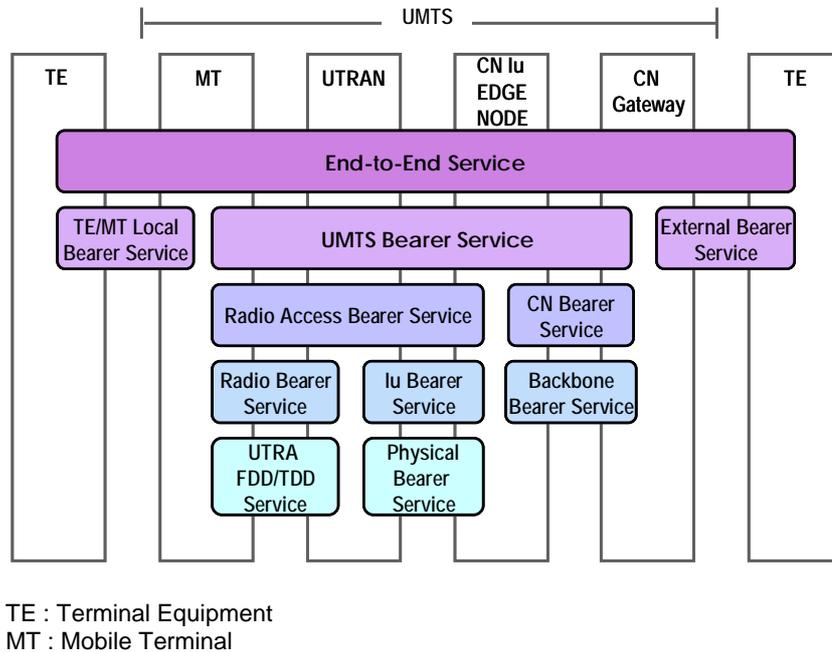
The **Radio Network Subsystem** (RNS) consists of a **Radio Network Controller** (RNC) and one or more **Node B**, essentially the UMTS base station. Each RNS is responsible for the resource of its set of cells. A Node B is responsible for the transmission in a cell or a number of cells and it is connected to the RNC through the Iub interface. Iu, Iub and Iur are logical interfaces [3GPP01].

- The **Core Network** (CN) covers all the network elements needed for switching and subscriber control. The CN maps the end-to-end QoS requirements to the UMTS bearer service and also maps onto the available external bearer service when inter-connecting to the other networks [3GPP01].

### ***2.4.2 Services and QoS Classes in UMTS***

UMTS is required to support the transmission with a bit rate up to 2Mbit/s for such services as video conference and video streaming. This is a requirement for use indoors, urban, hotspot area case, or at speeds up to walking speed. For coverage in the wider urban area where the user may be in moving vehicle or outdoors, a bit rate of 384 kbit/s is required. Finally, 144 kbit/s is the minimum bit rate requirement for rural or suburban areas where movement speeds will be higher.

Figure 2.13 depicts the layered architecture of a UMTS bearer service. A bearer service in each layer offers its services using those provided by the layer below [3GPP99b].



**Figure 2.13 UMTS QoS architecture**

There are four traffic classes (or QoS classes) in UMTS, differentiated mainly by level of delay sensitivity.

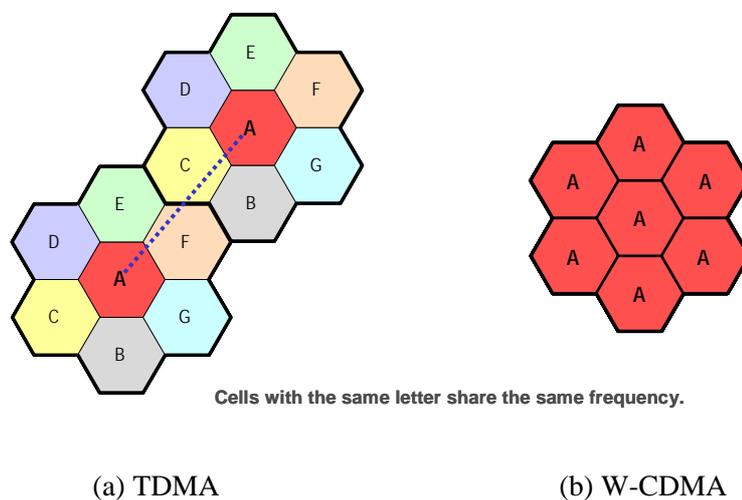
- **Conversational class** is that for the most delay sensitive traffic. The most common example of this class is speech (voice); video telephony and video games are other examples. These services must be transmitted as real-time connections over the radio link: there is no buffering and a guaranteed bit rate is required.
- **Streaming class** services are also transmitted as a real-time connection. The delay is slightly variable in this class and buffering is allowed. An example application is streaming multimedia, which uses a reconstruction technique that makes it appear as a steady and continuous stream. In this class, the bit rate is also guaranteed.

- **Interactive class** is used for data communications, such as web browsing and interactive network games, where the delay is moderately variable. The bit rate is not guaranteed for services in this class.
- **Background class** tolerates the most delay and background downloading of email is an example of a service in this class. Buffering is necessary and the bit rate is not guaranteed.

### 2.4.3 Radio Access Technique in UMTS (W-CDMA)

In 3G UMTS networks, the radio access method is W-CDMA. The main advantage of W-CDMA compared with TDMA used in 2G is that it gives higher system capacity and more flexible use of the limited radio spectrum.

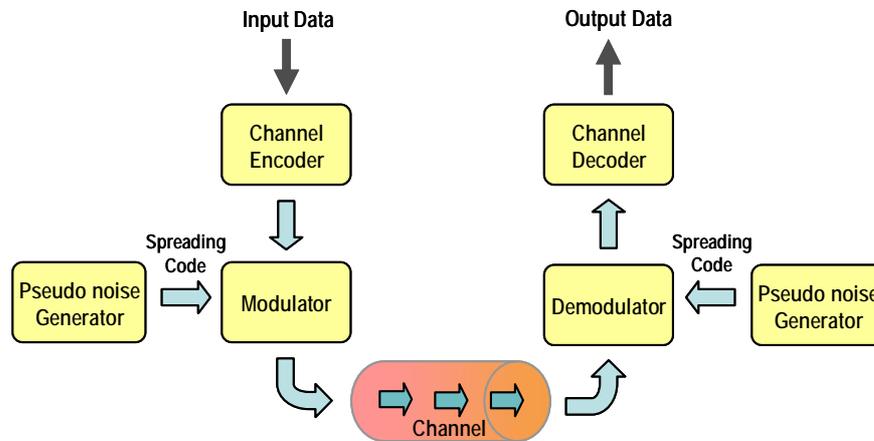
From a simplistic point of view, with the previous approaches, the bandwidth available in a cell depended not only on the total bandwidth of the spectrum allocated, but also on the frequency re-use pattern. For example, if a re-use pattern of 7 cells was used, the bandwidth available in each cell is one seventh of the total as illustrated in Figure 2.14 (a) while the whole bandwidth can be reused in every cell for UMTS as W-CDMA technique is employed (Figure 2.14 (b)).



**Figure 2.14 Cellular frequency reuse in GSM vs. UMTS**

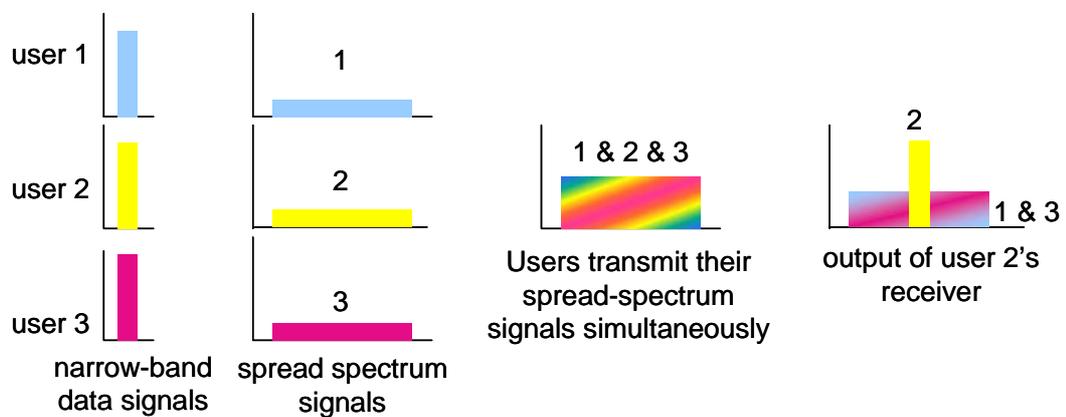
In W-CDMA, the use of spread spectrum provides sharing of the resource, illustrated in Figure 2.15. Each user employs spread spectrum modulation and the message signal is spread over a wide band by multiplying it with a pseudo-random spreading

signal. To demodulate, the signal at receiver is cross-correlated with an exact replica of the spreading function in order to distinguish itself from other users and noise.



**Figure 2.15 Spread spectrum process**

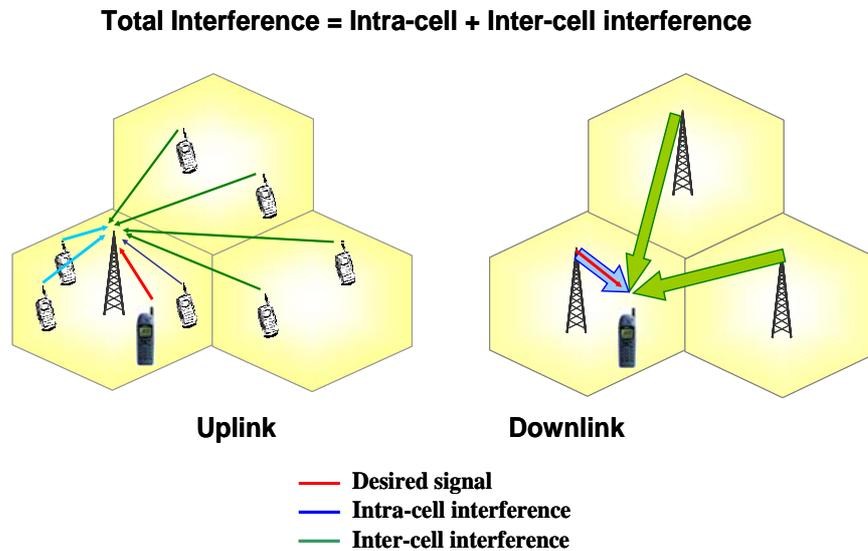
With a unique code being assigned to each user, all the users can coexist even though they use the same frequency band. Figure 2.16 illustrates how the signal from different users is modulated and transmitted at the same time in the same spectrum and how it can be recovered.



**Figure 2.16 Principle of spread spectrum**

As a result, all cells use the same frequency band and each has the full capacity: i.e. effectively there is re-use pattern of 1. On the other hand, determining the system capacity is a complex concept. The system is considered a soft capacity system because the number of users is only limited by the total interference that occurs from

others (Figure 2.17) and noise, which also depends on the geographical distribution of users so the determination of capacity is a complex concept.



**Figure 2.17 Interference in CDMA**

## 2.5 Radio Resource Management in UMTS

Radio Resource Management (RRM) is one of the most important entities in the UTRAN. It is responsible for maximising the system capacity and delivering the required QoS, given the finite radio resources at the air interface. The RRM algorithms can be divided into five functions:

- power control;
- handover control;
- admission control;
- load control (congestion control); and
- packet scheduling.

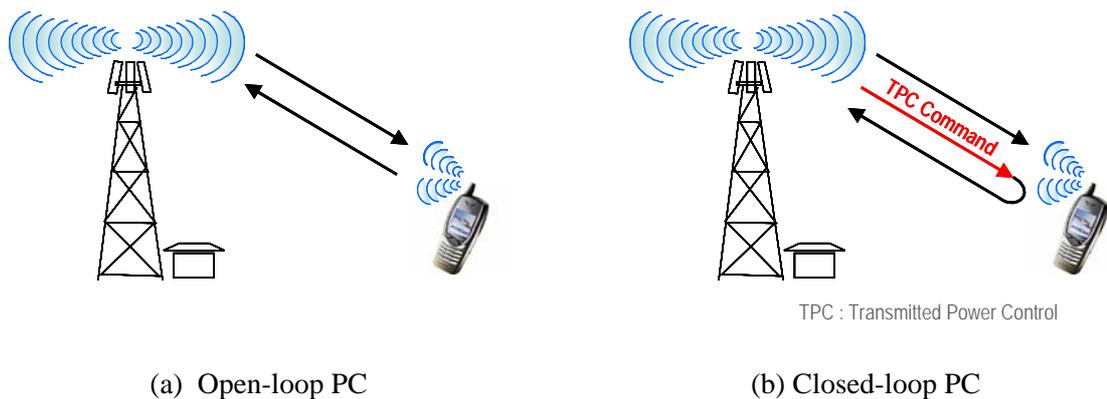
### 2.5.1 Power Control

Since many users are operating on the same frequency band, power control is essential to minimise the interference by adjusting the level of transmitted power to the optimum level, which means just enough to maintain the link quality: the aim is not only to maximise the battery life of the mobile terminal and to provide the required QoS, but also to make the most efficient use of the radio resource and crucially for CDMA, to reduce the interference levels and so maximise the capacity.

There are two broad categories of power control techniques used in UMTS (Figure 2.18).

**Open-Loop Power Control:** The transmit power of the UE and base station is decided based on the measurement of the power received from the opposite end.

**Closed-Loop Power Control:** The transmit power of the UE and base station is based on feedback from the opposite end.



**Figure 2.18 Power control techniques illustrated as uplink power control**

These broad categories are used as follows:

**Open-Loop Power Control:** responsible for setting the initial uplink and downlink transmitted power when a new connection is allowed onto the network.

- In the uplink, the UE uses parameters broadcasted in the cell and the received signal-code power to initialise its transmitted power.
- In the downlink, the base station relies on the measurement report from the UE in deciding the transmit power.

**Inner-Loop Power Control** (or Fast Closed-Loop Power Control): as this is a closed-loop power control, the UE or base station uses the received Signal-to-Interference Ratio (SIR) level at the opposite end to decide the transmit power.

- In the uplink, the base station uses the target SIR set by the outer-loop power control to compare with the received SIR from the UE once every time slot (0.666ms). If the received SIR is greater than the SIR target, the base station

transmits a TPC (Transmit Power Control) command 'down' or '0' to the UE. If the received SIR is lower than SIR target, the TPC command 'up' or '1' is sent to the UE. According to [BM00], the power control step size is dependant on the UE speed. An update rate of 1500Hz (once every time slot) is used. A step size of 1 dB can effectively track a typical Rayleigh fading up to a Doppler frequency of about 55Hz, corresponding to a UE speed of 30 km/h. With UE speeds up to 80km/h, it is better to use a step size of 2 dB as the power control can no longer follow the fades with the lower step size. For a UE speed lower than about 3 km/h, a step size smaller than 1dB can be beneficial because of the slow fading rate of the channel.

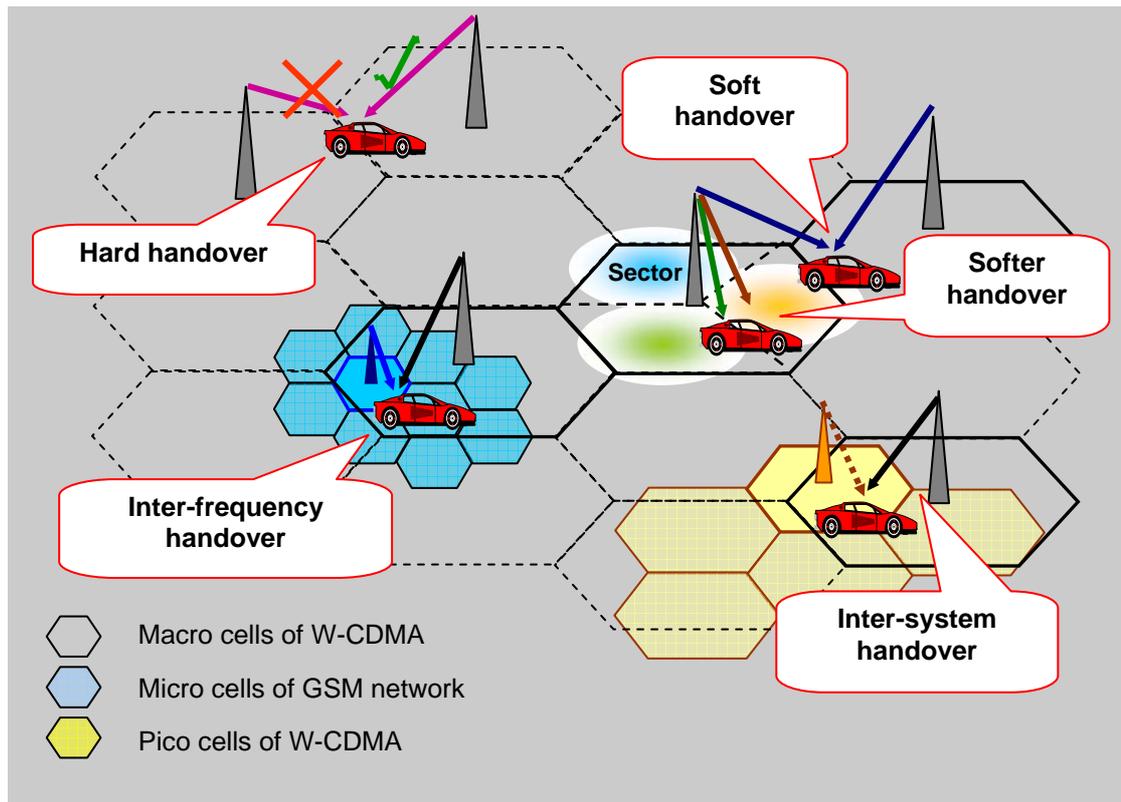
- In the downlink, the UE estimates the downlink SIR from the pilot symbol of the downlink channel and the estimated SIR is then compared to the target SIR. The TPC command 'down' or 'up' is sent to the base station according to the comparison result, as with the uplink. Step sizes of 1 or 0.5dB can be used.

***Outer-Loop Power Control:*** this adjusts the target SIR for the fast closed-loop power control based on the estimated quality in order to maintain the required QoS. Typically, the estimated quality is based on a certain target Bit Error Rate (BER) or BLock Error Rate (BLER).

- The uplink outer-loop power control is responsible for setting the target SIR in the base station for the uplink inner-loop power control. The step size is typically set between 0.1 to 1.0 dB.
- The downlink outer-loop power control is responsible for adjusting the target SIR for the downlink inner-loop power control. This process is done in the UE.

### ***2.5.2 Handover Control***

Handover control enables the network to maintain a user's connection as the terminal continues to move across cell boundaries. There are five types of handover technique as shown in Figure 2.19 [Chen03].



**Figure 2.19 Types of handover [Chen03]**

**Hard handover:** In this technique, a definite decision whether to handover the call to other base station or not is made so the user is dropped from one cell before being connected to another and the user cannot have simultaneous traffic flows from two or more cells. This procedure is easy and relatively uncomplicated.

**Soft handover:** When a mobile terminal moves into the soft handover zone it is allocated traffic channels from two (or more) base stations simultaneously. There is an algorithm based on the strength of the pilot signal from the base stations involved that determines when the UE has moved out of the soft handover zone and is served again by only one base station. Particularly in the downlink there are advantages and disadvantages for soft handover as explained in [Chen03]: basically these are macro diversity gain from there being two transmissions, although there is also extra interference.

**Softer handover** is similar to soft handover, but is between the sectors within a cell; in this situation there is only one active power control loop.

**Inter-frequency handover** occurs between cells on different W-CDMA carriers, which occupy different frequency bands, for instance between pico and macro-cells.

**Inter-system handover** takes place between cells belonging to two different types of networks. This will be increasingly important in the deployment of 3G as the 3G coverage will initially be limited and UEs will have to be able to handover between 3G and GSM/EDGE systems as they move out of the coverage area of 3G.

### ***2.5.3 Admission Control***

Connection Admission Control (CAC) determines which base station will have power control over a connection. This means that the base station must have sufficient bandwidth to support the new connection and also must ensure that none of the existing ones would be dropped, otherwise the new connection request will be rejected. Therefore, a good CAC should be able to guarantee low call dropping rate, even under high traffic load, and also have a reasonable call blocking rate. The admission control entity is located in the RNC, where it can monitor load information across cells. As new connection requests arrive, the load increment from both uplink and downlink has to be calculated separately and the connection is only admitted when both are satisfied.

A great deal of work has been done in this area and the following techniques are examples of the proposed methods.

**Number based CAC** [IU97]: This is a rather simple technique as the admission decision is made according to the number of active calls. The CAC threshold is set for each base station as the maximum number of users that can be accepted. Therefore, a new connection request will be blocked if the number of existing users has reached the threshold.

**Power based CAC** [HY96] [LWN02]: This method can be considered in two different forms, transmit-power based CAC and receive-power based CAC.

With the transmit-power approach, the new connection request is blocked if it causes the current connections to transmit with the maximum power and the system only admits the new connection if all existing ones can maintain acceptable carrier-to-interference ratio (CIR).

The receive-power based CAC (also called **interference based CAC**) will block the new connection when the total received power measured at the base station exceeds the threshold.

**Throughput based admission control** [LWN02]: This technique makes accept or reject decisions based solely on the total load calculated after admitting the new connection according to the threshold defined by the radio network planning (RNP). The criteria can be expressed as follows:

In the uplink,

$$\eta_{oldUL} + \Delta L \leq \eta_{thresholdUL} \quad (2-3)$$

In the downlink,

$$\eta_{oldDL} + \Delta L \leq \eta_{thresholdDL} \quad (2-4)$$

Where,  $\eta_{oldUL}$  and  $\eta_{oldDL}$  represent the network load in the appropriate direction before the new request, estimated with equations, (2-5) and (2-6), and  $\Delta L$  is the load increase generated by the new request calculated with equation (2-7).

$$\eta_{UL} = \sum_k \frac{1}{1 + \frac{W}{\rho_k \cdot R_k \cdot v_k}} \cdot (1 + i) \quad (2-5)$$

$$\eta_{DL} = \left[ (1 - \bar{\alpha}) + i_{DL} \right] \cdot \sum_{k=1}^N \left( \frac{\rho_k \cdot R_k \cdot v_k}{W} \right) \quad (2-6)$$

$$\Delta L = \frac{1}{1 + \frac{W}{\rho \cdot R \cdot v}} \quad (2-7)$$

In these equations,  $W$  is the chip rate, and  $\rho_k$ ,  $R_k$  and  $v_k$  are the  $E_b/N_0$  (bit energy divided by noise spectral density) requirement, the bit rate and the service activity of connection  $k$  ( $N$  is the total number of connections, respectively;  $v$  is the service activity of the new bearer;  $i$  is the average other-to-own-cell interference ratio; and  $\bar{\alpha}$  is the average orthogonality [LWN02]).

**SIR based CAC:** In this technique, the admission decision is made by referring to the measurement of the SIR. In the uplink, for each connection request, the base station will check the SIR that would be achieved for all existing connections plus the one from the requested connection. If all of them have acceptable SIR (meet the set target SIR), the new connection can be admitted to the system. In [AKKC02], the performance of this CAC method was investigated and in [LZ94] and [KSL00] the performance of different algorithms for SIR based CAC in the downlink side was considered.

In [HY96], a comparison was made between two interference based CAC methods: a scheme based on transmit-power CAC (TPCAC) that protects the established calls and one based on receive-power (RPCAC) that blocks new calls when the total received power at a base station exceeds a threshold. The result shows that the RPCAC scheme is found to offer significant reductions in the weighted combination of call blocking and call dropping. In [CR01], the number-based CAC and interference-based CAC were compared. The conclusions from their work are that under proper acceptance threshold setting, interference-based CAC would result in lower dropping probability than the number-based algorithm, which however seems to be more flexible with non-uniform traffic distributions. SIR-based CAC (signal-to-interference based CAC) has been proposed in [LZ94], the benefit of this scheme being in improving the system performance under hot spot traffic load.

#### ***2.5.4 Load Control (Congestion Control)***

Load control takes place when the system starts to become overloaded in order to get back to the stable state. The target load (or the threshold) is set by the radio network planning and the aim of a good load control is to quickly return the system to normal state once overload happens. The following methods are some examples of load control action according to [HT02]. The first two methods are the fast actions that happens within a base station.

- Deny downlink power-up commands for the downlink fast load control
- Reduce SIR target of uplink inner-loop power control for the uplink fast load control

Slower actions are carried out in the RNC as follow:

- Handover to another W-CDMA carrier or another network (GSM)

- Bit rate reduction for real time users
- Throughput reduction for packet data traffic.

### ***2.5.5 Packet Scheduling***

Packet scheduling is responsible for the initialisation of packet transmission and also determines the bit rate to be used. It is only relevant for non-real-time traffic and it determines which data, when, and at what rate, should be transmitted for how long [LWN02].

## **2.6 Previous Works of RRM in UMTS**

In the previous section, a general view of RRM was presented. As it is a main factor directly affecting the system QoS, a large amount of research work have been carried out in the area. This section will look into the published work in the area.

From literatures observed in the area of power control, more attention has been paid to the closed-loop power control. As mentioned in [HT02], the open-loop power control is only used to provide a coarse initial power for the mobile station in attempting to eliminate the slowly varying near/far and shadowing effects. The inner-loop power control is the solution for overcoming the fast fading environment particularly on the reverse link. [NTT00] In conventional inner-loop power control, the power control step size is negotiated at connection set-up stage and then fixed for the entire connection time. [ZM04], [NTT00], and [SGSS99] are examples of research which proposed new adaptive step-size algorithms for controlling the transmitted power of the mobile stations in order to improve system performance.

Since more benefits could be achieved by using soft handover rather than traditional hard handover ([Chen02]), a lot of research work has been conducted in this area. [BHC00], [ABG023], and [FSW03] are some examples of published work presenting the performance evaluation of soft handover.

Previous work in admission control area, which proposed several CAC techniques, was mentioned in §2.5.3. Because CAC has been a major area of attention for the last two decades, a great deal of work has therefore been done. [Ahmed05] presents the classification of CAC schemes. CAC schemes can be separated into centralized schemes, where only one entity controls admission for the whole network

([CKCN00][YGM00]), and distributed schemes, where each cell or each BS performs admission control [HY96][ZSM00].

CAC techniques can also be classified as a proactive (parameter-based) or reactive (measurement-based). [DJM96] and [ES00] are examples of work on the proactive CAC schemes. [HY96] and [VL02] are for the reactive ones.

[JK99] presents the taxonomy of admission control algorithms by referring to the information required by the CAC process into two approaches: cell-occupancy approach or a spatial mobility approach. The work based on a cell-occupancy approach can be found in [RNT96] and [CC97]. [LAN97] is an example of work proposing the scheme based on a spatial mobility approach.

§2.5.4 has introduced load control and its methods. Previous work [MB01] presents a theoretical framework for load control and discusses its applications. In [BIVK00], load control strategies based on resource reservation have been developed and evaluated. The strategies enable quality guarantees to data services, as result showing an improvement in non real-time service quality.

[SPRAC03] presents a range of representative case studies with several innovative algorithms for RRM and supports the idea by the simulation results in a realistic UMTS Terrestrial Radio Access Network scenario. The studies include a decentralised uplink transmission rate selection algorithm in the short term, a congestion control mechanism to cope with overload situations, and downlink scheduling for layered streaming video packets.

## **2.7 Summary**

In this chapter, an overview of cellular networks has been given including the evolution, basic cellular concept, and generations of systems prior to the UMTS. This chapter then focused on W-CDMA networks and its radio resource management together with the literature review on work being done in the area as it is the main focus of this research.. In the next chapter, the use of intelligent agents to provide SLA management is introduced

# Chapter 3 Agent-Based SLA Management

## 3.1 Service Level Agreement

It is widely expected that as future mobile systems develop, users will not be restricted to only one network or service provider (SP) : they will be able to choose which service provider they want to connect to, according to their perception of the QoS (in the general sense) they are receiving or in terms of the service offerings available. The SP will then be able to choose which network to use to carry the service requested.

In a limited sense this exists already with Mobile Virtual Network Operators (MVNOs). Virgin Mobile is an MVNO that uses T-Mobile's network; however, in the future this choice of network is likely to be dynamic, as is already the choice of international wired capacity through bandwidth brokers for example Band-X [www.band-x.com](http://www.band-x.com) in London and Arbinet [www.arbinet.com](http://www.arbinet.com) in New York.

An outline of the business model is shown in Figure3.1.

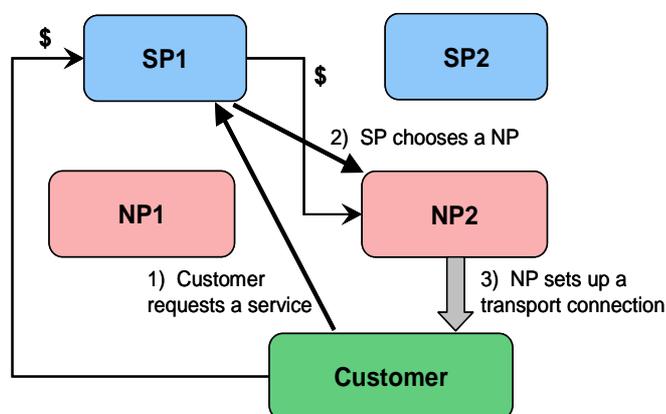


Figure 3.1 Principle of new business model

In setting up such an approach the service provider at least will require some sort of service level agreement (SLA) with the network operators and will use these SLAs as part of the decision-making process when allocating traffic. Corporate customers will also want SLAs between themselves and the service provider; the service provider will also monitor the service it delivers to all its customers.

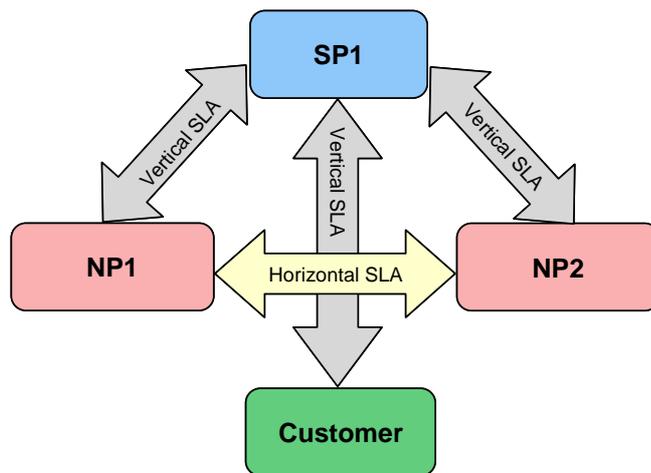
SLAs allow service providers to differentiate themselves from their competitors and allow them to offer different levels of service guarantees.

Specific terms have been defined for use with SLAs:

- **User:** Those who make use of the telecommunication services provided by the service provider; they can be organisation, companies, or individuals.
- **Customer:** Those who pay for the telecommunication services provided by the service provider; they can be organisation, companies, or individuals. In general the user and customer can be separate entities (for example an employee in a corporate customer is the user, but the corporate itself is the customer). In this thesis this distinction is ignored and the term “customer” is used.
- **Service Provider:** The companies who provide the communication services.
- **Network Provider:** The companies that own/operate the underlying networks.
- **Service Level Agreement (SLA)** is a contract between Service Providers and Network Providers, or between Service Providers and Customers, that specifies, usually in measurable terms, what services the service provider will furnish (the supporting services, service parameters, acceptable/unacceptable service levels, liabilities, and action to be take in specific circumstance) and what penalties the service provider will pay if it cannot meet the committed goals [MMPG02] [LR99].
- A **Horizontal SLA** is “an SLA between two providers being at the same OSI layer (for example two IP domains or two optical transport network domains)” (from [MMPG02]). Here, it would be between Service Providers, or between Network Operators (Figure 3.2).
- A **Vertical SLA** is “an SLA between two providers at two different OSI layers (for instance between the core MPLS network and an optical network)” (from [MMPG02]). Here it would be between the Service Provider and Network Provider, or Service Provider and Customer (Figure 3.2).
- **Service Level Specification (SLS)** is the technical specification deriving from the SLA: it can be a precise specification directly related to the SLA, but it can also be

an interpretation of the SLA, an adaptation depending on the provider or on the service. SLA/QoS management requires a definition of services, SLS parameters and a classification of these services depending on the SLS parameters [MMPG02].

- **Service Level Management** (SLM) refers to the process of negotiation, SLA articulation, checks and balances, and reviews between the supplier and customer regarding the services and service levels that support the consumer's business path [LR99].



**Figure 3.2 SLA types as applied here**

SLAs are usually generated with the following process (from [Alcatel03]):

- Understand the business path, business objectives and the user requirements.
- Compare actual performance against the long-term objective.
- Translate strategy into service and the service into performance metrics that are measurable and meaningful.

The generation of an SLA can be a very complicated task but once it has been created, good management is also important in maintaining an efficient service.

To provide a reliable service that meets an SLA, it is necessary to be able to monitor reliably how the SLA is being met – and in this network context that means resource admission control, allocation management, and resource allocation handling. In the

case of an unsatisfactory scenario, the SLA management has to be able to react rapidly to reinstate the promised QoS.

When writing an SLA that will be used within a network environment, it is very important that the specification focuses on the commitment to provide a particular level of service, rather than on the actions and liabilities that will be brought into play in the event of a problem happening. This means that the SLS should be clear and unambiguous and written in terms of attributes that are able to be identified and measured.

However, it must be realised that in a telecommunications application it may not be easy to define ways of representing a fair level for both customer and provider. For example, consider a simple level of service that might be *“95% of all call attempts in a month for a particular customer should succeed”*. If the customer is making many calls a day then any measurement will represent a fair average. However, suppose that the customer makes only 10 calls during most of that month and then encounters congestion lasting for 1 minute, during which time he makes 10 unsuccessful redial attempts. This results in the customer receiving only a 50% success rate, but in fact it could be that the network was only congested for that one minute, i.e. it was available for 99.9977% of the time, just that the customer did not want to use it then! While an extreme example, this does illustrate the difficulty with writing SLAs and interpreting them to SLSs.

## **3.2 Multiagent System in Resource Management**

### ***3.2.1 Agents and Multi-Agent Systems***

As stated in [Weiss99], an agent is a computational entity, such as a software program or a robot, that can be viewed as perceiving and acting upon its environment and that is autonomous in the sense that its behaviour depends on its own experience to some extent.

There are four main agent properties according to [LAd'I04] including *autonomy*, *reactiveness*, *proactiveness* and *social ability*:

- **Autonomous:** the agent must be able to function without direct command from a programmer or user, but in accordance with a set of tendencies, which could be in form of a goal or satisfaction that the agent attempts to achieve.
- **Reactiveness:** agents are capable of monitoring their environment and responding in such a way that they move efficiently towards the goal from their current situation.
- **Proactiveness:** agents have over-reaching goals that direct their behaviour over relatively long periods of time towards their goals and with complex tasks this means they can instigate actions to move towards those goals.
- **Social ability:** agents can communicate directly with other agents and act on information from those other agents to direct their own decision making.

Agents can be classified in different ways [Nwa96]. The first classification considered here is by their mobility and they are generally classified as being either *static* or *mobile*. Mobile agents can improve performance by moving from machine to machine at their chosen time and place; they can suspend their execution at an arbitrary point, transfer to another machine and resume the execution on the new machine. In telecommunications, mobile agents are generally not as popular as static agents since a mobile agent means *mobile code*, and mobile code is a security risk in terms of spreading malicious code (e.g. viruses).

Differentiating agents by their natures or habits ([LAd'I04]), there are 3 types: *reactive*, *deliberative* and *hybrid*.

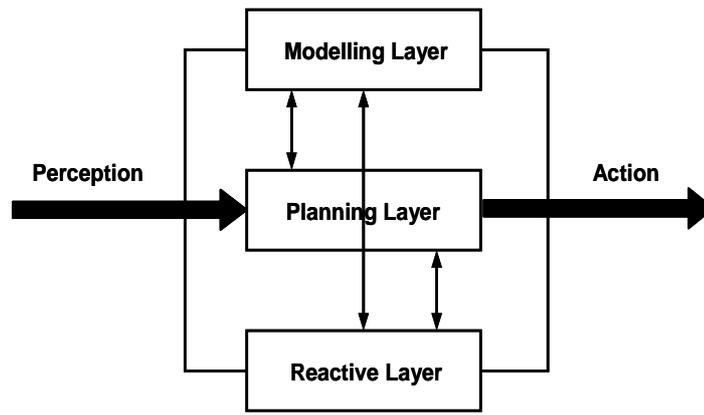
- **Reactive agents** are not able to predict what may happen in the future, so their course of action depends only on their current status and environment. They work by means of stimulus-response rules and the aim is to deal with situations that have been considered in advance. This means that reactive agents are only sufficient for limited situations where they can map to preset

rules, and this means that they are unsuitable where longer-term reasoning is required. In addition, they are very difficult to implement in a more complex environment as that would need a very large number rules. The advantages of reactive agents are that they are robust and can react very fast.

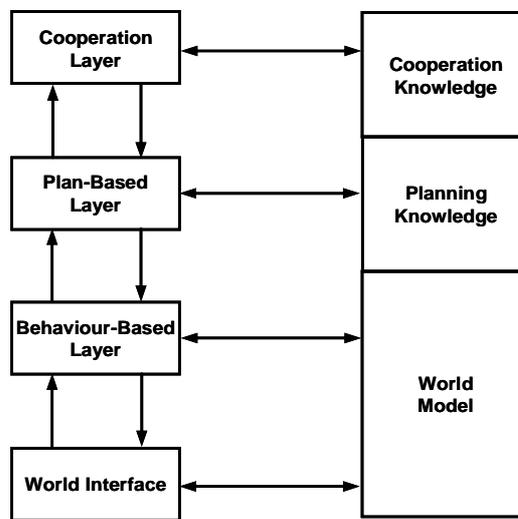
- **Deliberative agents** are capable of attaining a longer-term goal since they use the information in the “world model” ([WJ95]) to build a plan in order to achieve the goal without stimulus-response rules. Hence, the encoding of rules is not necessary because the planning system can determine goal-directed actions by itself, even for unforeseen situations. This ability to adapt is desirable in a dynamic environment, but the drawback of deliberative agents is the time consumed by the agent’s planning process since the plan generally has to be recomputed each time a decision is made.
- **Hybrid agents** are a combination between the previous two types, so they can act both reactively and deliberately. This type of agent is intended to overcome the drawbacks of the two individual types above in order to allow the agent to perform actions in a timely fashion as well as to effectively maintain system performance in complex environments.

[LAd’I04] gives two key examples of hybrid agent architectures that contain both reactive and deliberative components: TouringMachines and INTERRRAP, as illustrated in Figure 3.3 (a) and 3.3 (b).

The first hybrid agent architecture example, TouringMachines by Ferguson [Fer92], consists of three layers: *reactive*, *planning*, and *modelling*. The reactive layer gives a fast response to events not programmed in other layers; the planning layer is responsible for generating, executing, and modifying plans. Building and maintaining models of entities that are used to understand the current behaviour of the other layers and to make predictions about their future behaviour, are done in the modelling layer. The architecture shows that it is possible to operate this type of agent in a complex environment as it could produce a range of behaviour from reactive to goal-directed.



(a)



(b)

**Figure 3.3 (a) The TouringMachines architecture [Fer95]**

**(b)The INTERRRAP architecture [MPT95]**

The second hybrid agent architecture example shown in Figure 3.3 (b) is from the work by Miller and Pischel [MP94]. Their work also shows the benefit of using layered hybrid agent architectures in that they support the modelling of an agent's environment at different levels of abstraction, different time scales, and different levels of knowledge.

To consider agents intelligent, they need to be able to react in a timely fashion and interact with other agents in order to meet the designed objectives [Wool02]. In a very large application, a number of relatively small systems or agents can be involved in

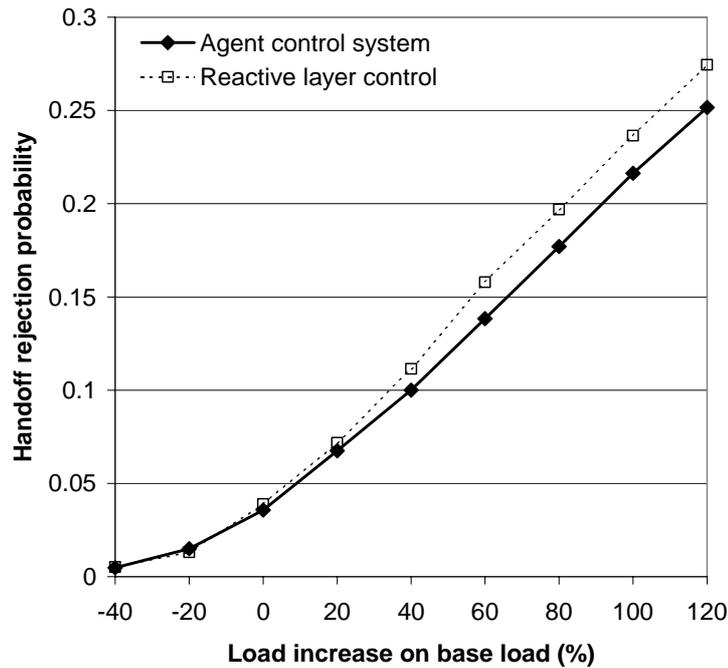
problem solving rather than using just one centralised system. Such a system is called a multi-agent system [Fer99]. In other words, multi-agent systems are computational systems in which several interacting intelligent agents pursue some sets of goals or perform some sets of tasks. Hence, they can handle more complex tasks than normally could be handle by an individual agent.

### ***3.2.2 Agents and Resource Management***

Resource management is very crucial aspect of any telecommunications system as it aims to maximise the utilisation while maintaining SLAs. It can be very complicated in wireless networks as one of their main features is that the users are moving and the traffic pattern changes rapidly. Therefore, the critical dimension in a radio network is the allocation and use of the (limited) bandwidth in the radio cells in order to avoid local congestion or degradation of the QoS and it is generally the capacity of the wireless link to the user that limits the overall system capacity, rather than any back-haul part of the network. Any work on SLAs must focus on this aspect of the network.

The first work on using intelligent agents to control mobile networks was by Bodanese [Bod00] [BC00a] [BC00b]. This resulted in a distributed resource allocation scheme for first generation mobile networks using intelligent agents that offered an efficient solution for resource allocation under moderate and heavy loads. In that work, the author restricted herself to first generation networks in order to be able to compare her results with the wealth of alternative schemes in the literature.

The main reason for using intelligent agents was to give greater autonomy to the base stations. That autonomy gives an increase in flexibility to deal with new situations in traffic load and to decrease the information load (the messaging resulting from taking, or determining control actions) on the network. An example set of results from [Bod00] is shown in Figure 3.4 where it can be seen that in this case the handoff rejection rate is significantly reduced compared to a purely reactive scheme.



**Figure 3.4 Benefits of using an agent control system for mobile networks  
(from [Bod00])**

The reason for this improvement is that the co-operation between agents allows resources to be allocated that cannot be accessed by the reactive layer alone.

A conclusion coming from the work was that the choice of planning layer was important; later on it will be shown how adapting policies to meet SLAs can implicitly improve the choice.

The most direct previous work relevant to this thesis is that of the IST Project SHUFFLE (IST-1999-11014) [Bod00]. In that project, the work of Bodanese was extended to 3G networks, the business model of §3.1 was introduced and the overall agent architecture will be described later.

However, the SHUFFLE project only offered the hypothesis that the agents could control SLAs: there was no detailed study on how to implement such control, nor were there any results on SLA management.

### 3.2.3 *Functional Architecture*

In the current mobile networks, the users make connection with a particular service provider (SP), who buys network capacity (resources) from the network provider who owns the physical network. The user will then be restricted to that particular provider until the end of the contract.

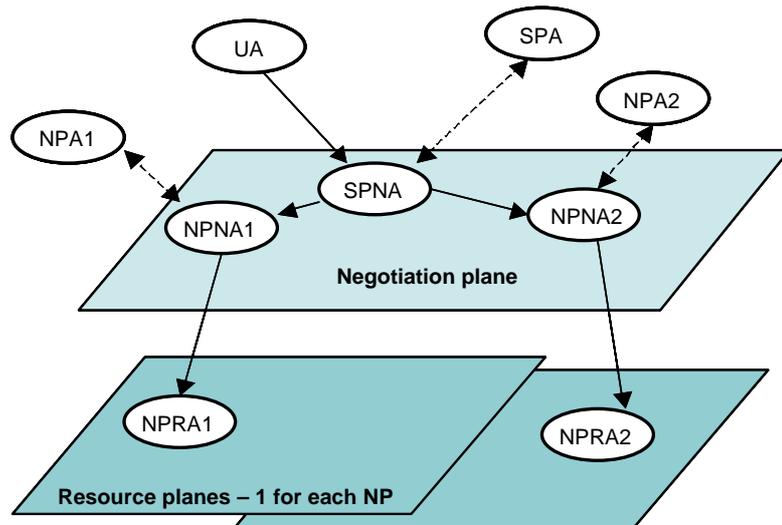
In any business environment, all parties to an agreement have their own interests: this means that users would want to be able to choose the service provider who best serves their requirements at any particular time; SPs wish to choose the NP that is offering the best price/performance package; and NPs want to sell their capacity to SPs who will give them the best return. Therefore, the trend is toward more flexible and resilient models (Figure 3.1).

Here, a multi-agent system is introduced to (i) give more flexibility in provider selection and to (ii) manage resource utilisation to allow SLAs to be met.

The functional architecture used here comes from SHUFFLE [CRTBB01]. By using agents, it is possible to allow selection of service provider according to their offering on price, QoS, or value added service type; it is also possible for service providers to choose a network provider on a similar wide range of criteria. The outline functional architecture to achieve this is shown in Figure 3.5.

The control concept is split into two functional places in the network: the *negotiation* plane and the *resource* plane.

- The **negotiation plane** is where all the interactions between the customers, the SPs and the NPs occur. The communication between the agents belonging to different entities will take place here. As users put in requests for services, SPs will handle these requests and negotiate with NPs for the transport capacity.
- The **resource plane** is where NPs manage their network resources both across and within individual radio cells to meet the SLAs they have with SPs. This plane is the domain of the NP. In work described in this thesis, the resource plane will be the focus as the Network Provider Resource Agent (NPRA), a crucial agent that manages the resource within the network, is located here.



UA: User Agent	NPA: Network Provider Agent
SPA: Service Provider Agent	NPNA: Network Provider Negotiation Agent
SPNA: Service Provider Negotiation Agent	NPRA: Network Provider Resource Agent

**Figure 3.5 Illustration of functional and agent architecture adopted by SHUFFLE (from [CRTBB01])**

Both of these planes use intelligent agents to manage their area of responsibility; these agents are as follows:

- A **User Agent (UA)** normally resides within or near the User Terminal (UT) and it will act on behalf of the user to represent his/her interest. The responsibility of the UA is to make sure that the user obtains the service as requested. In other words, it maintains SLAs with all SPs to which the user subscribes by interacting with the service provider negotiation agent.
- **Service Provider Agent and Network Provider Agent (SPA & NPA)** act on behalf of the SP/NP to control the overall policies. Their main function is to coordinate the activities between different provider agents.
- The **Service Provider Negotiation Agent (SPNA)** acts on behalf of a SP, being responsible for negotiating SLAs with the UAs of subscribing users and the network provider negotiation agents of the NPs that could be carrying the traffic from that SP.

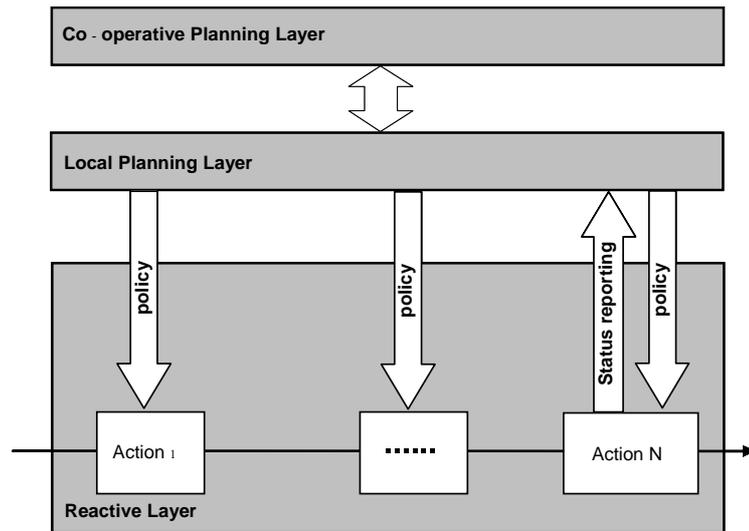
- A **Network Provider Negotiation Agent** (NPNA) acts on behalf of each NP to negotiate SLAs with SPs in order to manage the contract as actual traffic conditions vary.
- The **Network Provider Resource Agent** (NPRA) is a key component in managing the resource: it acts on behalf of the NP to implement the policies of the NPA and to manage the radio resource of the network provider. The details of this agent will be explained more in §3.2.5 as this agent is the main interest in this thesis.

### ***3.2.4 Internal Agent Architecture***

The agent structure is shown in Figure 3.6 and this is also taken from [CRTBB01]; each agent follows the concept of Bodanese [BC00b] who used three layers taking action and decisions on different timescales: *reactive*, *local planning* and *co-operation*.

As an individual connection must have any decision made in real-time, the reactive layer is designed for a very fast response. More complex, and slower acting, functions are implemented in the planning layers. Generally the local planning layer is concerned with long-term actions within its own instance, whereas the co-operative layer is concerned with long-term actions between peer agents, or with other types of agent.

The reactive layer is, therefore very simple, implementing policies being passed down by the higher layer (Figure 3.6). It also monitors its actions and feeds back this status information to the planning layers so that they can monitor the effectiveness of their actions.



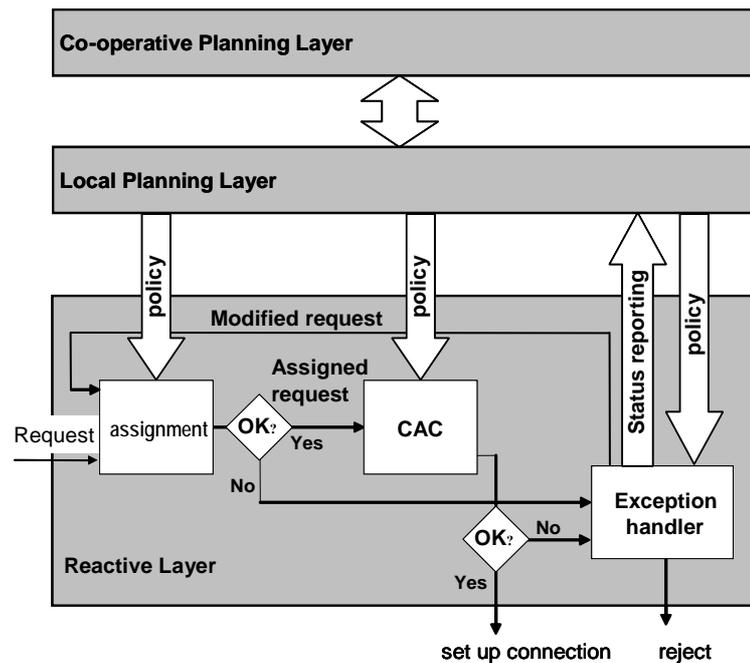
**Figure 3.6 General agent structure**

### **3.2.5 NPRA**

The architecture of NPRA is illustrated in Figure 3.7 (this is also from [CRTBB01]). The reactive layer is designed to be fast, performing the same function that would be in a conventional RNC (Radio Network Controller), assigning the connection to a Node B, and performing CAC but it does this according to policies assigned by the planning layer.

The connection request (containing information about the service provider, QoS, type of connection) is first considered for assignment to a Node B using an algorithm or set of rules passed down from the planning layer.

The assigned Node B is the one that is allowed to perform power control, and which subsequently accepts or rejects the call. The assignment and CAC scheme can be passed down as a policy, as could a new scheme. Since these can be changed dynamically, the planning layer can respond to the local resource issues by changing the assignment and CAC strategy. Hence a suitable policy can be chosen to match the current situation, reporting to the planning layer from the reactive layer in order to maintain the system performance.



**Figure 3.7 NPRA internal architecture**

The local planning layer of the NPRA is responsible for setting the policies. As explained previously, this can change the reactive layer assignment strategy, or it can change the QoS allowed on the connections within its control.

Higher-layer control (at the co-operative planning layer) depends on co-ordinated action between groups of cells and hence is likely to need handling by negotiation between several NPRAs.

### **3.3 Case-Based Reasoning Approach**

As explained earlier, intelligent agents are attractive in controlling communication networks because they can decide for themselves what they need to do to meet their objective, both by using knowledge programmed into the agent and from knowledge they have learnt.

Learning is not only about remembering successful strategies, but also learning from scenarios that did not work in order to prevent such failures to occur again.

Here the attraction in using learning is to be able to jump straight to a policy (that would be downloaded to the reactive layer) that worked in a similar situation previously. By using such a learning method the response time will be much faster

because the planning layers would not need to approach each situation *ab initio*, but would be able to recall a solution that matches, or is closer to, the situation at hand.

### 3.3.1 Why use CBR?

Possible learning approaches that could be adopted are *Neural Networks* (NN), *Fuzzy Logic*, *Case-Based Reasoning*, and *Reinforcement Learning*. Many more approaches are currently being researched worldwide but these four are the best known options.

The structure of a **Neural Network** consists of interconnected “units”. Each unit converts the pattern of incoming activities by multiplying the value with the weight and adding together all weighted inputs from other units to get the total input; this total value is sent to an input-output function which transforms that value into the outgoing activity. The network is trained by being repeatedly shown large numbers of examples for the problem under consideration. One of the most important tools for training a NN is the *back-propagation algorithm*, in which network will adjust the weight of each unit in such a way that the error between the desired output and the actual output is reduced.

A NN approach seems to have good potential in systems that need high computation rate or where many hypotheses are pursued in parallel; these can include examples such as signal processing and speech or image recognition. If NNs were to be applied in this work, the training time would be large and the system does not actually need a complex computation process to find out the right solution.

An alternative would be **Fuzzy Logic** (FL), which allows the handling of uncertainties arising from deficiencies like incompleteness or vagueness in information. FL is a problem-solving control system methodology that mimics how a person would make decisions, only very much faster. FL could be a good method for sorting and handling data or in many control system applications. However, Fuzzy Logic needs expertise in setting up the rules that are used to determine the action to be taken. The decision that can be made by FL control system is directly conceived from the set-up rules. Therefore, the use of FL will be successful only when the system that has all knowledge about the problem controlling methodology. In this work, some definite rules can be made but a new situation might not be exactly the same as that experienced before. Hence, the system that can adjust the old solution to best match the new situation and also learn from the new experience is needed.

**Case-Based Reasoning** is one Artificial intelligence approach that allows the agent to learn from past successes. CBR is a method that finds the solution to the new problem by analysing previously solved problems, called *cases*, or adapting old solutions to meet new demands. In this work, CBR seems to be the most suitable approach as different traffic patterns can be studied, and they together with the solution can be indexed in the case library to use for solving future problems.

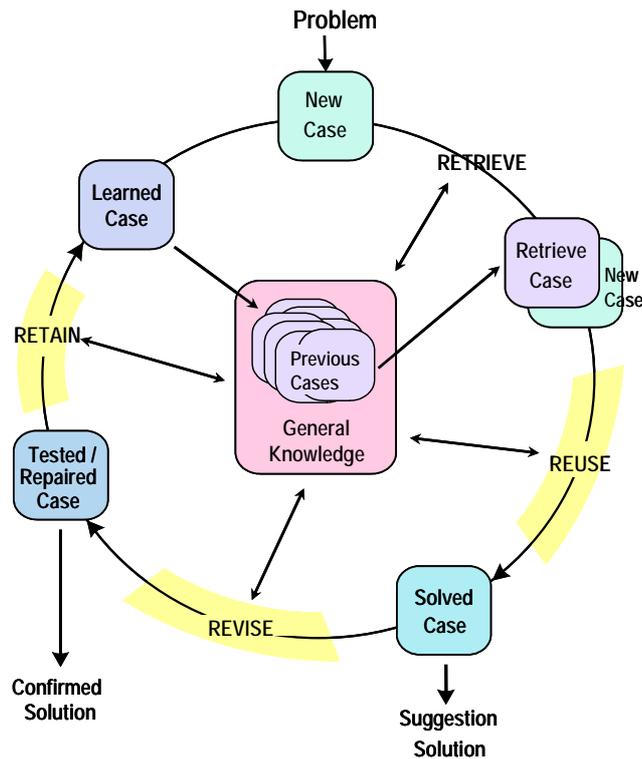
Applying CBR to agent control of mobile networks is completely new: it goes beyond what was done in SHUFFLE and is entirely the work of the author.

All the previous learning approaches are considered “supervised learning” as knowledge or examples of the problem solving method is needed at different levels for different techniques. For the non-supervised technique, **Reinforcement Learning** was reviewed. Reinforcement learning is focused on goal-directed learning by interacting with dynamic environment. The major characteristics are a trial-and-error search and delayed reward [SB99]. In this work, CBR is more suitable approach because it can solve the problem with less time consuming and more efficient as it already has knowledge of successful solutions where as reinforcement learning might have to go through a trail-and-error process. However, reinforcement learning could be an option for enhancing the performance of the revision step of CBR (§3.3.2) in the future.

### ***3.3.2 CBR Process Model***

Figure 3.8 shows the process model for CBR; it starts when there is a new problem or new case happening. There are 4 main steps: *case retrieval*, *case reuse*, *case revision*, and *case retainment*.

- **Case retrieval** is the first step in the CBR process model. In this stage, the main task is to find the best match to cases (previously solve situations) in the case library for the situation that has just arisen. The process consists of identifying features in order to achieve the event description (or characterising indexes of the event), searching for the nearest matches, and selecting the best. The best conceivable result is that which matches *all* the input features, but most likely it would be the nearest match from those that have been retrieved.



**Figure 3.8 Case-based reasoning process model**  
**(Based on the CBR cycle in [AP94])**

- **Case reuse** is the assessment of the retrieved case in terms of how different it is from the new event, and the decision on which part of the solution can then be reused: the whole solution or just the method to construct the solution. Where the retrieved case matches exactly to the new event, the solution can be reused unmodified. In reality, such a situation will not happen often so that the solution might need to be modified to fit the new situation and this leads to the next step, case revision.
- **Case revision** involves two possible processes depending on the evaluation from the case reuse stage:
  - with a successful outcome, the system will learn from the success as shown in the next step (case retainment);
  - otherwise, a monitoring process will analyse the failure and the solution will be repaired by using specific knowledge from the system.

- **Case retainment** is an important part of learning as the system needs to store what it has learnt from successful matching to use in the future. One of the most vital aspects is the indexing of the new case into the previous version of the case library because it gives a direct effect to the future retrieving process. Later in this section, more details on indexing methods will be given.

### 3.3.3 *Cases and Retrieving Algorithm*

In general, cases consist of two component parts: (i) the description of the problem that was being solved and (ii) the description of its solution. One of the most important aspects that affect the performance of CBR is the problem of making sure that the most suitable case is accessed. Crucial to this is how the case library is indexed and subsequently accessed to retrieve potential matches.

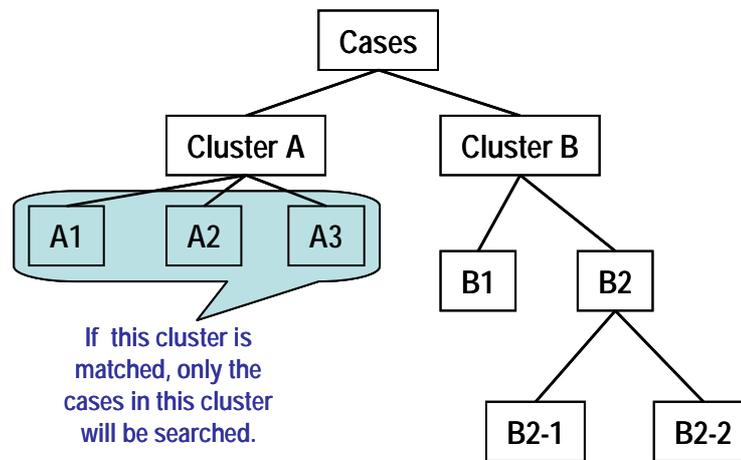
According to [Kol93], there are several possible structures and algorithms and these are described more detail below, from the simplest to the most complex. The selection of the most suitable structure and algorithm to be used depends on the number of case in the library, the complexity of indexes, the number of different tasks the case library must support, and the variety of indexes across those tasks [Kol93]. The more likely candidates are:

- **Flat memory, serial search:** cases are simply stored in the library in sequence without any organisation; on retrieval, each case will be matched in sequence and the degree of match will be returned. At the end, those cases that have highest match will be selected. The benefits of this method are that:
  - the new case can be easily added into the library;
  - by searching the entire library, it is only the matching function that affects the accuracy.

The disadvantage is obviously the time consumed in the retrieving process, especially as the case library gets larger.

- **Shared feature networks, breadth-first graph search:** this structure is suitable for large case libraries as the cases will be organised in hierarchical order. The cases will be grouped into clusters in which they share the main

characteristic. Inside the clusters, there could be sub clusters as appropriate. Figure 3.9 shows the structure of a shared-feature network.



**Figure 3.9 A Shared-Feature Network**

Breadth-first search is used for this type of structure so it will start from the highest level in the graph and continue to the deepest level for which only the cluster or sub-cluster that matches the feature will be considered. By using this method, only a relatively small subset of cases will have to be searched. The main benefit of this approach is that it is less time-consuming as the case library can be searched in more efficient way. However, adding a new case into the library is more complicated since it needs to be placed in the right location.

- **Flat memory, parallel search:** a flat memory was described earlier, but with the serial search. Here, the same structure is used, but the search is conducted in parallel, the whole library being searched at the same time. This makes the case matching and retrieving process happen in one step and gives the advantage of being less time-consuming for that part of the process as well as allowing new cases to be easily added to the case library.
- **Hierarchical memory, parallel search:** this combination offers the advantages from both the hierarchical memory and parallel search: it saves time, given that a smaller group of cases will be searched, and also the matched case is efficiently found in one step by the parallel search. *This is the method that is used in this research.*

### 3.4 CBR in SLA-Based Resource Control

There has not been a great deal of work reported on applying CBR to telecommunications networks: [CH95] gives an example of using CBR in network traffic control by using it to control traffic flow in the standard public switched telephone network of the Ile de France; in [HA1-MA1-Z01], CBR is used to correct the error estimation of the required bandwidth computed by conventional connection admission control schemes. However, there is nothing that is directly related to the work described in this thesis where CBR is used to find the solution for the particular congestion problem recognised in the network.

How this is applied is as follows. Figure 3.10 shows the internal architecture of the NPRA, but this time it is modified to include CBR in the local planning layer.

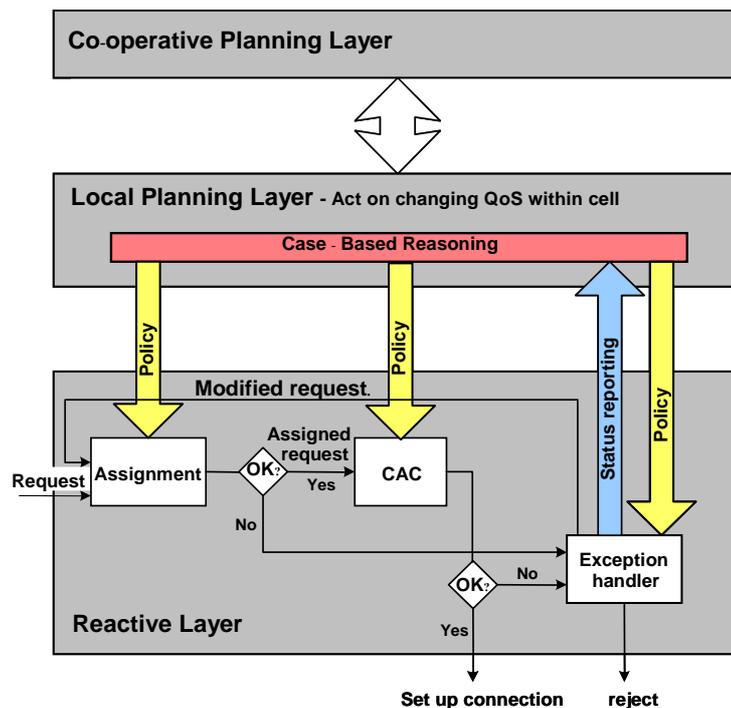


Figure 3.10 NPRA (with CBR) internal architecture

The results of monitoring what is happening to connection requests will be sent from the reactive layer to the local planning layer. When there is a congestion problem (i.e. a problem of the SLAs not being met) detected by the planning layer, it will attempt to find the best solution for that particular traffic pattern. Since this attempt uses CBR, the best solution will be found by looking at previously-solved problems as explained

earlier. The policy resulting from that solution is then passed down to the reactive layer to apply and the result will be observed in order that the agent can learn and update the case library as appropriate.

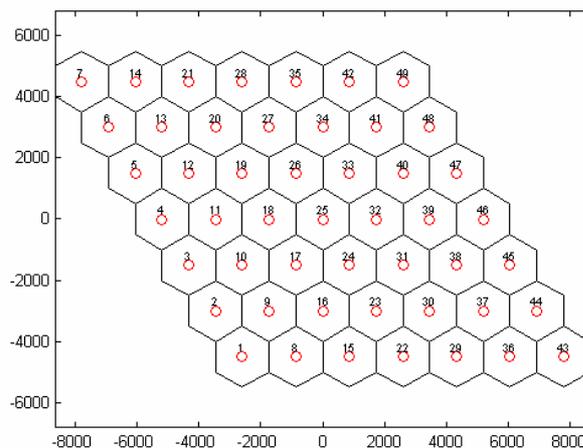
### **3.5 Summary**

This chapter presented an overview of SLAs and presented the idea of using agents in RRM, including the related work. It also gave details of the proposed multiagent system in RRM including the system functional architecture, agent internal architecture, and the NPRA, which is the focus here. After that, the consideration of agent approaches were analysed and CBR, as a chosen approach, was explained.

The most complex research issue in applying CBR is to determine a method of indexing: i.e. mapping the “view” of the network in such a way that the planning layer can identify specific cases and match to previous solutions. In the next chapter, the simulation model developed here to observe the behaviour of the proposed system will be presented followed by chapter 5, giving detail on the monitoring process and congestion pattern matching mentioned here.

# Chapter 4 Simulation Model and Validation

A simulation model to investigate the CBR approach was implemented in MatLab. The system consists of hexagonal cells (9, 25 or 49 have been used so far) and each cell has a base station with an omni-directional antenna placed at the centre of the cell; each cell has diameter of 1000m. This is a fairly standard mobile network layout for simulation. Figure 4.1 shows the layout of the simulation model being used in this work for a 49-cell system. A 49 cell layout has been created from the beginning and also the validation has been done in this layout. (§4.7) Later, the number of cell in the system was reduced to 25 and then 9 because the simulation concentrates on the cell and its immediate neighbours. Thus, a large number of cells is not necessary, and reducing the number of cells allows the simulation to run faster.



**Figure 4.1 Simulation Model**

Traffic from mobiles is generated randomly using the traffic model described later in §4.1 and the load is varied by adjusting the mean inter-arrival time between each connection request. Mobile customers are divided into three classes: bronze, silver, and gold. In the results described here, 50% are bronze, 30% are silver and 20% are gold, but this is configurable: it is assumed that gold customers will pay the highest service charge (followed by silver and bronze customers) and hence will have the most stringent SLA.

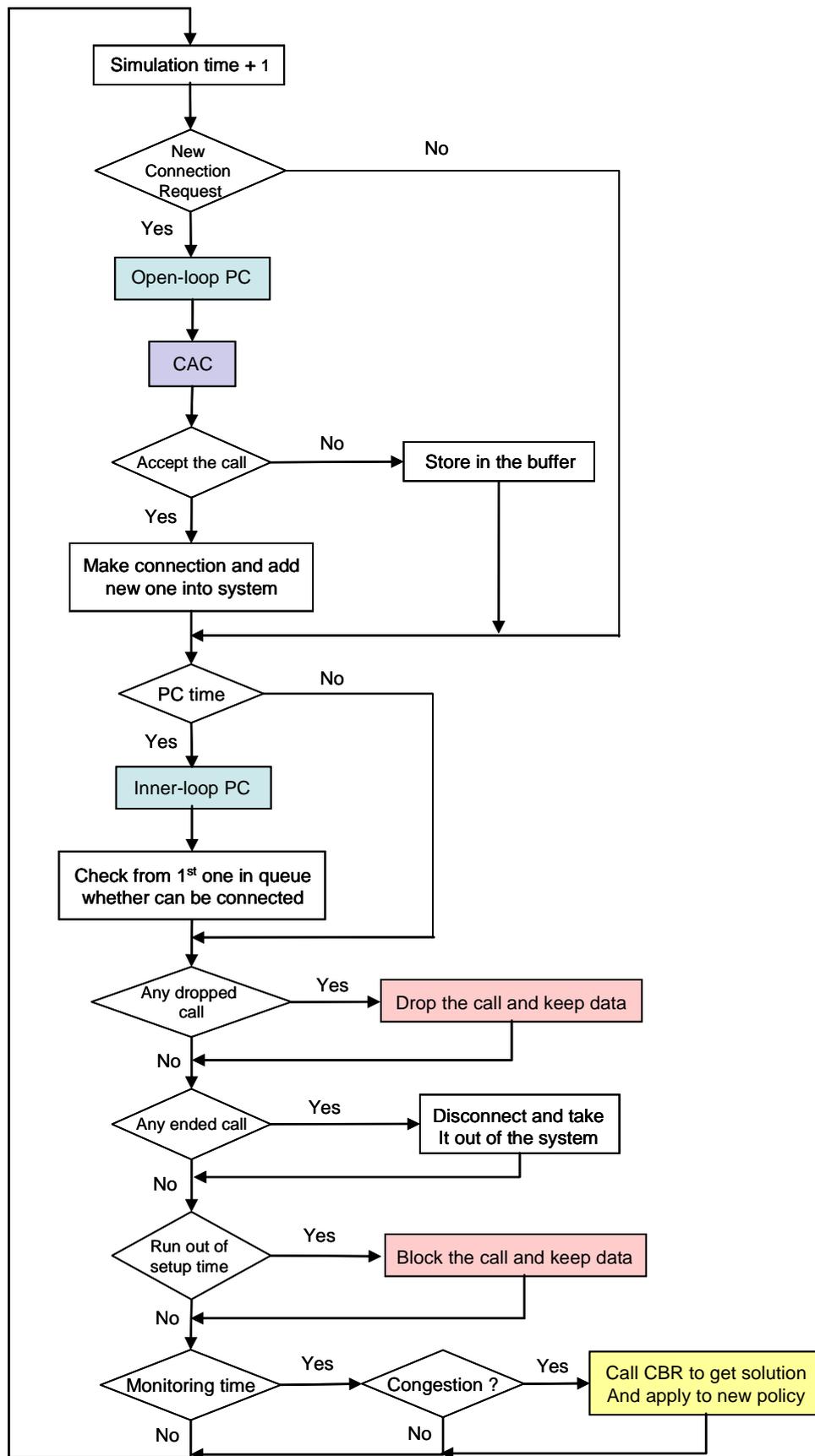


Figure 4.2 Simulation process

Figure 4.2 shows the flow diagram of the MatLab simulation process. A *fixed-increment time advance* is chosen here as the simulation method with a time step of 1ms following a feasibility study as to whether *fixed-increment* or *event-list* simulation would be better. The system developed in the simulation model is a combination between discrete and continuous systems as some state variables change continuously with respect to time and some change at discrete points in time: for example, the power control and monitoring process have to be called at regular intervals. Although some states have a variable time advance, the use of a time step with a suitable value will ensure that all the continuously variable states are reasonably well covered. Hence, the fixed-increment time advance is the more suitable mechanism.

The power control (PC) model and the CAC model will be explained in §4.4 and §4.5 respectively.

Each time step, the process starts by checking whether any new connection request has appeared; if there is a request, the open-loop PC is called in order to estimate the initial transmitted power for the new connection. After that, the new connection SIR will be recalculated and passed to the CAC process and the decision whether to accept or reject will be produced. If an *accept* decision is produced, the new connection is made and its information will be added to the system. Otherwise, this requested connection will be kept in the buffer to wait for availability.

Next, the PC timer is checked; if the timer has expired, the inner-loop PC is called to set a new transmitted power for all the existing mobiles. After each PC loop, the connection requests that are waiting in the queue will be reconsidered (starting from the first one in the queue), as there might now be more available capacity to make a connection available for them.<sup>2</sup>

After performing the power control it is necessary to check whether any connections are to be dropped<sup>3</sup>. If there is any connection to be dropped, it has to be taken out of

---

<sup>2</sup> Note that 3G capacity is interference limited; the recalculation of power levels might have reduced the interference in order to allow other connections to be admitted.

<sup>3</sup> The recalculation of power may also mean that the interference is worse and connections have to be dropped.

the system and the data kept for analysing the results. Any calls that have ended must also be taken out of the system.

Then a check must be made whether any calls are to be blocked. Each call request is put in a queue and allocated a time period that it is allowed to remain in the queue. At the end of each iteration, the time remaining for each call is reduced by the iteration step time and any that have run out of time will be blocked: i.e. they are removed from the buffer, the data being kept for analysis.

The last step is the monitoring process that occurs periodically. All the monitoring parameters are collected over period of time and are used to calculate the values of rates specified in the SLAs and determine whether those SLAs are being met. In this work, *congestion* is defined as when the SLAs are not being met and at that point the CBR model will be called to retrieve a better solution. Once the new policy has been installed, it too will be. As the situation is recovered, the ordinary policy will be released back.

The simulation time is increased by 1ms after each loop and the simulation process continues by repeating the next loop until a predetermined time. More details of each step and the models that are used in the calculation process are described in the later sections.

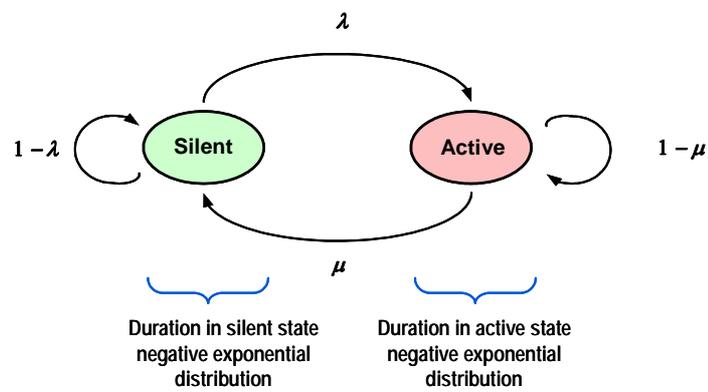
## **4.1 Traffic Model**

In the initial work by the author, the traffic model used was based on that of the SHUFFLE project. This used a multirate service with random generation of connections, subject to the constraints that 50% of calls needed 16kbit/s, 25% needed 32kbit/s, 12.5% needed 64kbit/s, and so on up to 0.390625% at 2Mbit/s; the connections were allocated to the cells randomly. The call holding time for each connection were generated by using exponential distribution with a mean of 180s. The traffic load could be controlled by changing the mean inter-arrival time between each connection, which is again exponential distributed. This initial version of traffic model was used for the simulation at the beginning of the work and the early simulation results illustrated in §6.1 (Figure 6.1) were produced by using this traffic model.

Later in the work, a new traffic model was introduced to more closely match work in the literature. This new traffic model has three types of traffic that are explained in the following sections.

### 4.1.1 Voice Traffic

Voice traffic is considered to be real-time traffic. The common model for a single voice source uses an ON-OFF process [PS01]. It consists of 2 states, active (ON) and silent (OFF) stage, with a transition rate  $\mu$  from ON to OFF and  $\lambda$  from OFF to ON stage. Figure 4.3 illustrates the ON-OFF model.

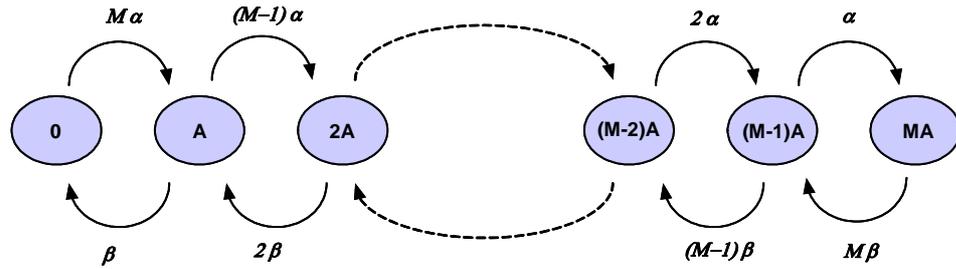


**Figure 4.3** Traffic model for voice call (ON-OFF model)

In this work, it is assumed that the voice traffic has a bit rate of 8 kbit/s during the ON period and the ON periods represent 45% of the time. Referring to the approach in [KM99] to simplify the simulation, an activity factor of 0.45 has been used. To be more specific, an activity factor is the ratio of the ON period over the total time. The value used here, 0.45, therefore means that for 45% of the time, the signal will be present at the channel. The mean holding time is assumed to be 180s.

### 4.1.2 Video Traffic

Video traffic is also considered as real-time traffic. The common model for video source is illustrated by the Discrete-state continuous time Markov process shown in Figure 4.4 [MASKR88].



**Figure 4.4 Video source model**  
**(Discrete-state continuous time Markov process)**

The bit rate of video traffic is quantized into finite discrete levels with step of  $A$  bits/pixel ( $0, A, 2A, \dots, MA$ ). Transitions between levels occur with negative exponential transition rates that may depend on the current level [SC02].

The parameters  $\alpha$  and  $\beta$  are the basic state transition probabilities and they are obtained by:

$$\beta = \left[ \frac{3.9}{1 + \frac{5.04458N}{M}} \right] \quad (4-1)$$

$$\alpha = 3.9 - \beta \quad (4-2)$$

Where,  $N$  = Number of aggregated video sources (typical assumption 1) and  $M$  = Number of quantization levels (typical assumption 8).

As might be expected from the description of the model, it is very complicated to implement in the simulator, and it slows the simulator down greatly. This is a real problem as the detail being modelled anyway makes the simulator slow.

Implementing this video traffic model in the simulation causes simulation times to be very long. Many authors therefore simplify this by using the same approach as voice traffic, activity factor [KM99] [AKKC02]. Here, an activity factor of 1 has been assumed for the real-time video source as use in [KM99]. The transmission rate for video traffic is assumed to be 64, 144, or 384kbit/s and mean holding time is 300 s.

### 4.1.3 Data Traffic

In general, data traffic is considered non real-time traffic. In the simulation model used in this work, data traffic consists of two types

- i) Real-time data traffic (circuit-switched) with a transmission rate of 64, 144, 384 kbit/s. For this type of data service, an activity factor of 1 and a mean holding time is 300s are assumed. The method is the same as that for voice traffic.
- ii) Non real-time data traffic (packet switched) with a transmission rate of 8, 64, 144, or 384kbit/s. The traffic model is based on the traffic model used in [Tri01] and is illustrated in Figure 4.5. It has a three-layer structure of a *session*, *packet call* and *packet*. A session consists of several packet calls with some inactivity periods between them. Each packet call consists of several packets. The following figures are the assumptions made in the simulation model.
  - Number of packet calls per session = Geometrically distributed random number with mean equal to 5
  - Inter-arrival time between packet calls = Geometrically distributed random number with mean of 120s.
  - Number of packets per packet call = Pareto distribution random number with mean of 25
  - Packet size = 480 bytes (from [Tri01] for www, email, ftp, and FAX)
  - Inter-arrival time between packets = Geometrically distributed random number with mean of 0.067

This data traffic model has been included in the simulation model for some part of the work. In the simulation results chapter (Chapter 6), the results shown in §6.1 (Figure 6.2 and Figure 6.3) were simulated by using this version of traffic model that included the data traffic. As a result, the simulation took too long to be really usable

so that later work excluded the data model. This simplified traffic model has then been used for all the simulations covered in §6.3 onwards.

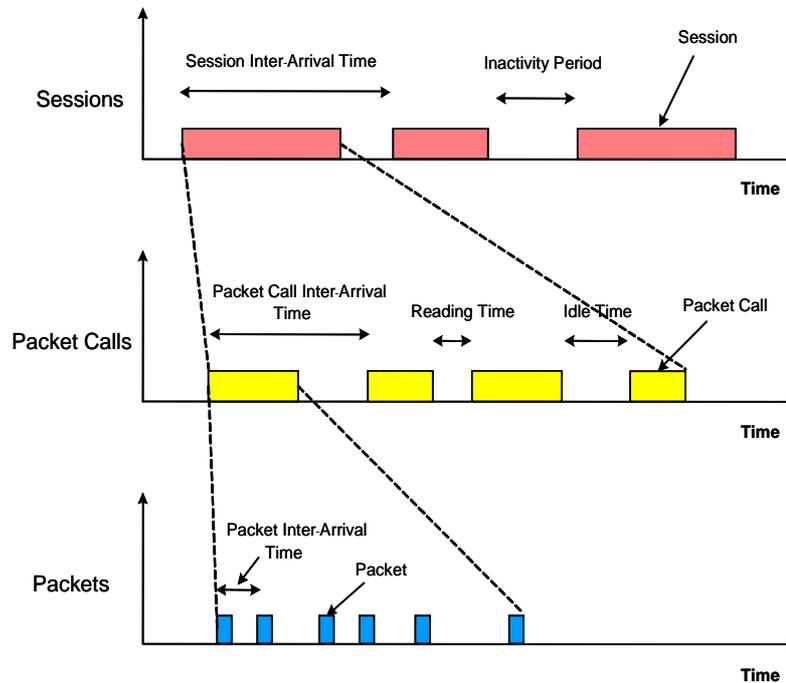


Figure 4.5 Components of Data Traffic [Tri01]

## 4.2 Radio Propagation Model

In cellular systems, radio propagation is influenced by the path loss depending on the distance, shadowing, and multipath fading. The relationship between the transmitted power and received power can be expressed as [LZ94].

$$P(r) = 10^{\xi/10} \cdot r^{-\alpha} \cdot P_0 \quad (4-3)$$

Where,  $P(r)$  is the received power;  $P_0$  is the transmitted power,  $r$  is the distance from the base station to mobile,  $\xi$  in decibels has a normal distribution with zero mean and standard deviation of  $\sigma$  (typical value of 8 dB), and  $\alpha^2$  represents the gain. (Typical values of  $\alpha$  in a cellular environment are 2.7-4.0.)

## 4.3 Receiver Model

Assuming the network to be uplink capacity limited, the SIR of each transmission is calculated at the base station and it can be expressed as follows, (based on [CR01])

$$SIR_i = \left( \frac{W}{R_i} \right) \cdot \frac{Pr_i}{(I_{intra} + I_{inter}) \cdot AF + N_{thermal}} \quad (4-4)$$

Where,  $W$  = Chip rate (3.84Mcps for W-CDMA);  $R$  = Bit rate;  $(W/R)$  = the processing gain;  $Pr$  = the received signal strength;  $I_{intra}$  = the sum of the received signal powers of other transmissions within the same cell;  $I_{inter}$  = the sum of the received signal powers from the other cells;  $N_{thermal}$  = the thermal noise power given below;  $AF$  = activity factor.

$$N_{thermal} = F * k_B * T \quad (4-5)$$

Where,  $F$  = Noise figure assumed 10dBW;  $T$  = Temperature in Kelvin assumed 298 Kelvin;  $k_B$  = Boltzmann constant  $1.380658 * 10^{-23}$  J/K

## 4.4 Power Control Model

Power control was introduced in §2.5.1. As the simulation focuses on a network limited by uplink capacity, only uplink power control is considered. Two types of power control are applied: open-loop power control and inner-loop power control. The target SIR is assumed to be 6dB [KM99] with a power control step of 1dB [BM00] [TB02]. The connection will be dropped if the SIR is below the threshold, which is assumed here as 4dB [TB02].

### 4.4.1 Open-Loop Power Control

The open-loop power control is applied when new connection requests arrive in the system. The total interference at the base station is calculated as it is the parameter that the UE needs to use in the estimation of its initial transmitted power. According to the parameter and the target SIR, the UE makes the estimation and uses that value as an initial transmit power.

#### 4.4.2 Inner-Loop Power Control

For an uplink capacity limited system, inner loop power control is used to optimise the transmitted power of all the UEs. Figure 4.6 illustrates the process implemented here. Firstly, a base station calculates the received SIR from each UE. If the SIR is less than the target SIR, the TPC command “up” is sent to the UE and the UE increases its transmitted power by one step. If the SIR is more than the target SIR+1, the TPC command “down” is sent to the UE and the UE decreases its transmitted power by one step. Otherwise, the UE maintains the same transmitted power. After the power control cycle has been performed, the new SIR for each mobile is calculated. Any mobile that has an SIR value less than the threshold will not be dropped immediately; instead the system will try to reallocate that mobile to another base station nearby that still has available bandwidth and can provide the link quality. If it is possible, the mobile will be handed over to the next base station; otherwise the mobile will be dropped.

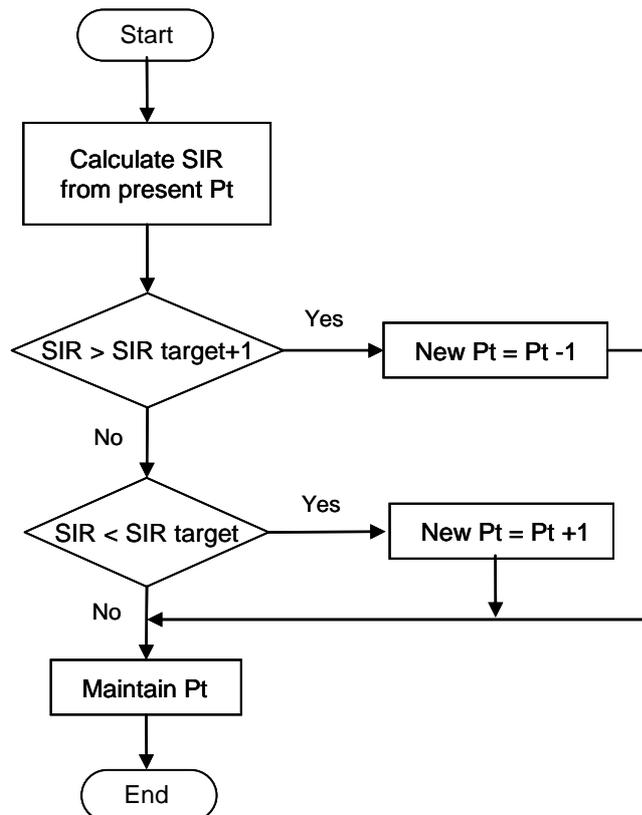
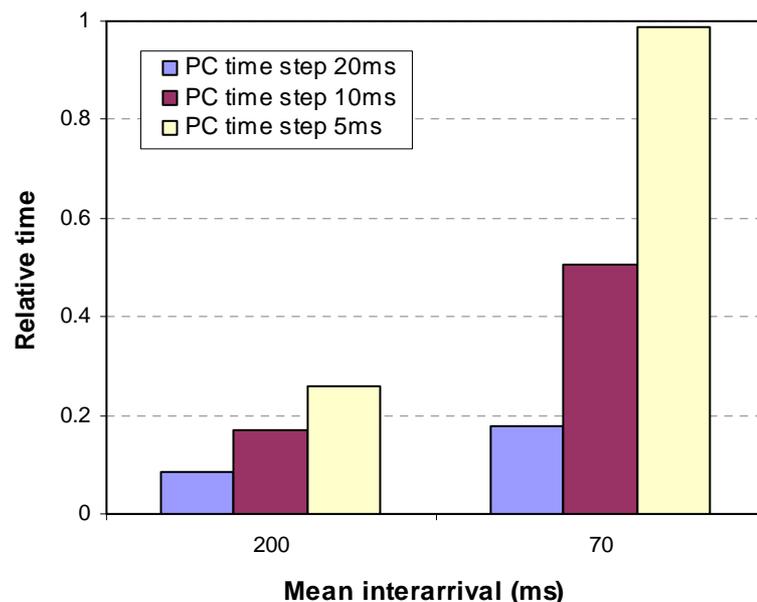


Figure 4.6 Power control process

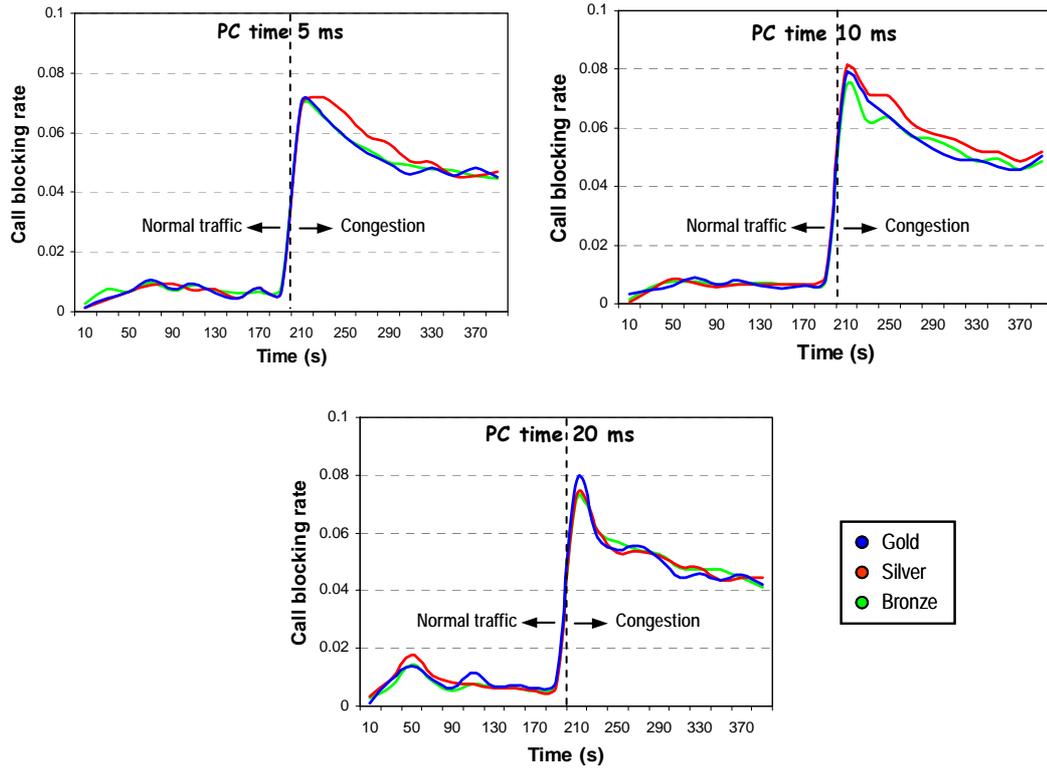
As mentioned in §2.5.1, this inner-loop power control is executed at a rate of 1500 cycles per second which means that transmitted power will be updated every 0.667ms.

In order to optimise the time consuming for the simulation model, a study was made of choosing different power control times. 1ms is representative of a real network but produces simulation times that are too long to be useful. In the early work, a compromise value of power control cycle time was chosen to be 5ms as the result is reasonably representative, but leads to much shorter simulation times; it is also the value used in a number of research works [KM99].

Later on, experiments were done to investigate the effect of varying the power control time from 5ms up to 20ms. From the comparison, a time step of 10ms has been chosen as there is not much difference in the results from 5ms or 10ms, but a more marked different with 20ms. From the simulation point of view, it is preferable to use the highest of time as this reduces the time for the simulation to run. Figure 4.7 shows the simulation time as the traffic load is varied for different power control time steps under the same simulation environment. Figure 4.8 shows the comparison between system operated under different power control time steps (5, 10, and 20ms) for the call blocking rate over time as the congestion occurs randomly over the system.



**Figure 4.7** Effect of different power control time step on simulation time usage



**Figure 4.8 Effect of different power control time step on the call blocking rate**

In some related works, perfect power control algorithm has been assumed for inner-loop power control in order to reduce the simulation time since the time advance for this type of power control can be higher [CR01]. In this case, transmitted power will be recalculated using the equation below:

$$\text{New } P_t = P_t - [\text{SIR} - \text{SIR target}] \quad (4-6)$$

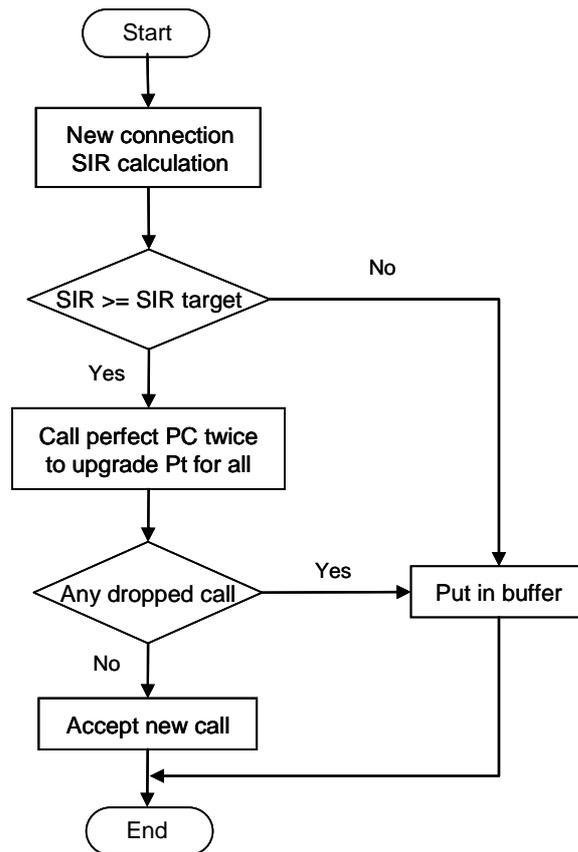
Therefore, the transmitted power is regulated according to the difference between the current SIR and the target one.

The simplification is not used in the power control model here but it is used in the CAC model.

## 4.5 Assignment and Admission Scheme

In this research, a combination between the ideal scheme and SIR-based CAC is chosen. As for the ideal scheme, the system has to make sure none of the existing connections will be dropped when accepting a new connection that request. In §2.5.3,

which introduced admission control schemes, the literature on SIR-based CAC was presented. As it is being widely used in many research works, some showing it outperforms other methods, the SIR-based CAC is chosen here. Uplink capacity-limitation has been assumed in this work, so that it is the signal-to-interference of the received signal from the mobile to the base station that is calculated. The CAC process applied here is illustrated as a flowchart in Figure 4.9.



**Figure 4.9 Connection admission control process**

As the system receives a new connection request from the UE, the new UE's transmitted power is estimated in order to get the target SIR. If the estimated transmitted power is in the acceptable range, it means the new UE can make a connection. Otherwise, it will be blocked or held in the buffer.

By adding the new connection into the system and performing the first perfect power control (described in §4.4.2) loop, the new transmitted power that should give each connection the target SIR can be determined. The second perfect power control loop is performed to achieve the actual SIR for each connection as a result of

accepting a new connection request. If any existing mobile would be dropped (by having an SIR less than the threshold) the new connection is rejected otherwise it is accepted and the connection is made.

## **4.6 CBR Model**

In §3.3.3, several schemes of organizational structure and retrieval algorithms for CBR were described. As explained there, the hierarchical memory with parallel search is used in this work as it provides an efficient retrieval that is less time consuming plus lower complexity.

All cases in the library are organised according to the hierarchical memory scheme. Here, the cases that share the same feature will be clustered together and the most important feature will be used on top to differentiate all the cases; this is followed by less important ones (see Figure 3.9).

With the parallel search, the whole library will be searched for each characterizing index vector in one step. Together with the hierarchical memory, the search will start from the most important feature, followed by lesser ones in decreasing order of importance; the cluster of matched cases will get smaller and smaller until the best match can be achieved.

If the new case is to be retained in the library, the library index has to be re-sorted according to the priority of the characterizing index of the new case. More detail will be illustrated with the cases being tested in this work in Chapter 5.

## **4.7 Verification and Validation**

### ***4.7.1 Introduction***

One of the most important aspects in developing the simulation model is its credibility: without this, the result of the simulation model will be in doubt. Therefore, *validation* and *verification* of the simulation model are essential [LK91].

- *Verification* determines whether the simulation model performs as intended. Thus, verification checks the translation of the conceptual simulation model into a correctly working program.

- The *validation* process determines whether the conceptual simulation model is an accurate representation of the system under study.

During the development of the simulation, different methods for verifying the model were carried out. Debugging subprograms was performed as they were written since it is easier and more accurate to debug the subprograms rather than trying to debug the whole model once it becomes large. Results from a particular state have been printed out and compared with calculation by hand to make sure that the simulation is operating as intended.

There are several techniques for validating the simulation model. In this case, comparing with results from similar simulation models has been used.

#### ***4.7.2 Introduction of the relevant model***

The simulation model was validated by comparing the result with the equivalent results from [CR01]. In [CR01], a comparison was made between the performance of two CAC methods for voice, one based on the number of active calls and the other on run-time measures of the power emitted by the base station or the total received interference. The second approach in [CR01] is that chosen for this work because [CR01] shows that it performs better and also because it is a more reasonable choice for the multi-service system.

In [CR01], the simulator represents a system composed of 49 hexagonal cells that lay on a torus surface to avoid border effects. The base stations are located at the centre of each cell and radiate with omni-directional antennas. The relationship between the received power ( $P_r$ ) and transmitted power ( $P_t$ ) is given by

$$P_r = P_t \cdot \alpha^2 \cdot 10^{\frac{\xi}{10}} \cdot \frac{1}{L} \quad (4-7)$$

where,  $L$  is the path loss,  $10^{\frac{\xi}{10}}$  accounts for the loss due to slow shadowing with  $\xi$  being a normally distributed variable with zero mean and  $\sigma^2$  variance, and  $\alpha^2$  represents the gain, with an exponential distribution of unit mean, due to fast fading.

The path loss is expressed as

$$10 \log L = 128.1 + 37.6 \log r(\text{dB}) \quad (4-8)$$

where,  $r$  (in metres) represents the distance between the mobile and base station, assuming the shadowing standard deviation is equal to 5 dB.

Two different traffic patterns have been considered, homogeneous and hotspot. In the validation, the homogeneous traffic pattern was considered as it is less complicated.

At the receiving side, the SIR after despreading is expressed for each transmission, as

$$SIR = SF \cdot \frac{P_r}{I_{intra} + I_{inter} + P_N} \quad (4-9)$$

where,  $P_r$  is the received signal strength,  $P_N$  is the thermal noise power assumed equal to  $-103$  dBm in the uplink,  $I_{inter}$  is the sum of the signal powers received from the other cells,  $I_{intra}$  is the sum of the signal powers due to other transmissions within the same cell, and  $SF$  is the spreading factor assumed equal to 128.

In the power control model, the transmitted power is adjusted at each iteration to maintain the SIR at the target value, target SIR ( $SIR_{tar}$ ). The new power control level is evaluated as

$$P_{new} = P_{old} \cdot \frac{SIR_{tar}}{SIR} \quad (4-10)$$

Each channel cannot exceed a transmitted power of 21 dBm in the uplink. The  $SIR_{tar}$  values adopted in the simulation are 6.1 dB for the uplink and the  $SIR_{min}$  is equal to 3 dB. Green lines in Figure 4.10 show the results for the interference-based CAC for the uplink with the homogeneous traffic distribution from [CR01]. The offered traffic against accepted traffic for different values of the threshold level is plotted.

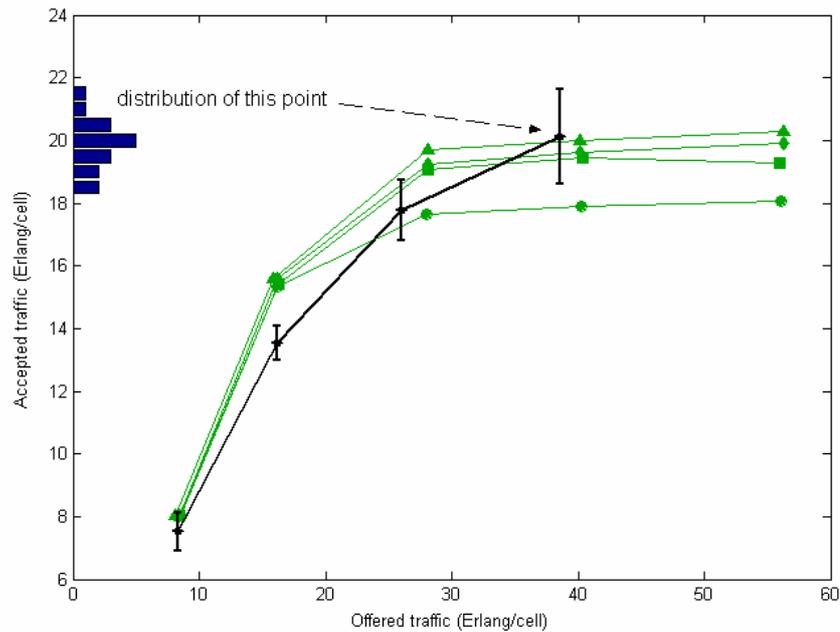
### **4.7.3 Validation and discussion**

In this section the performance of CAC for the multi-service and multi-class system was studied in order to validate against the result from [CR01]. To do this validation, the system environment and the assumptions for the main parameters must be set to the same as those in the reference work (or as nearly as possible).

This has been done as follows:-

- The simulation model consists of 49 hexagonal cells, each having one base station located at the centre of the cell radiating with omni-directional antenna.
- Voice service is assumed with a bit rate equal to 12 kbit/s.
- All the customers are of the same type and are equally served by the system.
- The traffic load has been varied by changing the mean inter-arrival time of calls.
- The traffic pattern has been set as homogenous in the mobile generator randomly generating the position of mobile.
- The main parameters have been set as follow:-
  - Maximum transmitted power = 21 dBm (uplink)
  - SIR target = 6.1 dB (uplink)
  - SIR minimum = 3 dB
  - Threshold level = -50 dBm

Running the simulation using the simulation model developed in this work with the main settings above gives the results shown by the black line in Figure 4.10, on which are also plotted the results from [CR01]: the green lines. A set of simulations was run for each traffic load value. The data points are from the average value and the error bars were plotted by analysing each set of results.



**Figure 4.10 Comparison between the validating result from the simulation model and the results from [CR01]**

From Figure 4.10, it is clear that results from the simulation model show the same trend as the corresponding results from [CR01] under a similar system environment. The values from both models are similar although there are not exactly the same because of small differences in some details: for example the method of power-control SIR calculation and some assumptions in noise value. However, the trends are sufficiently close and the confidence distributions sufficiently tight to validate the simulator and to give credibility to its results.

The process of verification and validation confirm that the simulation model is credible.

## 4.8 Summary

In this chapter, the simulation model developed in this work has been given with a detailed explanation including:

- Traffic model
- Radio propagation model
- Receiver model

- Power control model
- Assignment and admission scheme
- CBR model

The last section provided the verification and validation process done in the research to ascertain the credibility of the work.

In the next chapter, congestion scenarios that were studied will be illustrated, along with the system monitoring process and congestion pattern recognition.

# Chapter 5 Monitoring and Congestion Pattern Recognition

## 5.1 Simulation Scenarios and SLA Assumption

In these tests, the system environment is controlled to give two different types of congestion patterns in order to test the performance of the CBR approach in responding to different traffic patterns: random overload and a hotspot. These two scenarios are considered as they are common situations in reality.

### 5.1.1 Random Overload Cases

Here, the whole system traffic load is increased from a normal traffic level<sup>4</sup> when the simulation reaches stability. (§6.2 illustrates stability verification of the system.) There are three sub-cases for the random overload scenario differentiated by the accumulative value of call blocking rate for gold and silver customers because it is not generally short-term violations that are important in SLAs as an SLA might specify, for instance, that the blocking rate must not exceed a certain value during a day or over months. In other words, this specification is intended to guarantee that the customer will not be blocked again and again in a certain period of time as that could cause frustration which might lead to the customer changing network operator.

The three random overload cases consist of:-

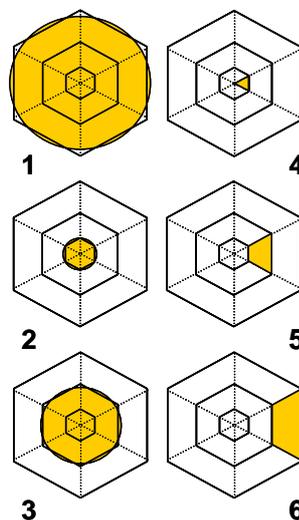
- Random overload case with accumulative blocking rate of gold customers exceeding the limit.
- Random overload case with accumulative blocking rate of gold and silver customers exceeding the limit.
- Random overload case with accumulative blocking rate of silver customers exceeding the limit.

---

<sup>4</sup> Here, *normal traffic situation* means the system is under acceptable traffic load as it uses the default policy and the QoS is maintained in accordance with the SLAs.

### 5.1.2 Hotspot Cases

For the hotspot cases, the traffic load is increased to create a congestion environment as the simulation reaches stability (§6.2 explains how the stability is determined.) in a particular cell or a particular area within that hotspot cell. In order to identify the area of congestion within the hotspot cell, the hexagonal cell is partitioned as shown in each hexagon in Figure 5.1. As the simulation runs, the traffic condition will be monitored and data is collected for each subsection of the hotspot cell. Hence, a specific area of congestion can be identified when overload occurs. There are six sub cases initially tested for the hotspot environment and Figure 5.1 shows the layout of the hotspot cell for each hotspot case.

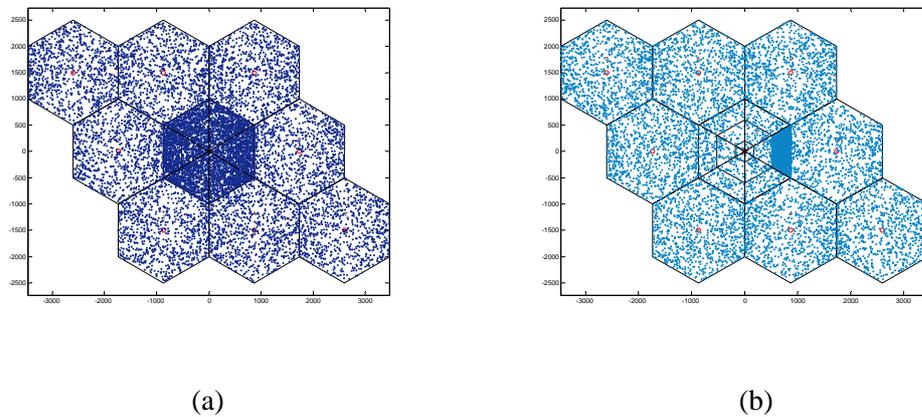


**Figure 5.1 Hotspot cases illustrated by hotspot cell layouts**

These hotspot cases cover a variety of patterns as, in reality, the hotspot area could be large or small and also it could be in any specific part of the cell, not only around the centre such as case 6 in Figure 5.1 where the congestion is located right next to the edge of the cell.

Figure 5.2 shows some examples of hotspot scenarios with mobile positions that were generated by the simulation model described in Chapter 4. The hotspot cell, (which is in this case the centre cell) has a traffic load five times higher than the other cells in the system. It should be noted that the traffic load can be controlled to change the congestion level, but in the experiments a factor five has been used throughout. The extra load has been located into a specific area according to the

traffic pattern. Figure 5.2(a) is from hotspot case 1 which has overload traffic spread over the whole hotspot cell and Figure 5.2(b) has congestion located in a smaller area at the edge of the hotspot cell.



**Figure 5.2 Example of hotspot case 1 and case 6 as scenarios were generated by a simulation model**

It has been noted previously that congestion is determined by the deviation of monitoring parameters from the SLA, and that an SLA is an agreement between provider and provider, or between provider and customer, to guarantee the service quality according to the contract. Hence, each SLA could have its own individual specifications, but here, the SLAs are call blocking rate only:-

- Maximum acceptable rate for gold : 0.03
- Maximum acceptable rate for silver: 0.05

Note that the SLA assumption for the maximum acceptable call blocking rate for bronze customer is not considered in this work.

## 5.2 Monitoring Process

The monitoring process of the system is initially performed every 10s. (60s has been used later on; this is presented in §5.3.1.) This means the monitoring parameters will be collected for 10s and sent to the local planning layer of the NPRA agent where the CBR model is located as shown in Figure 3.10. The parameters will then be compared with the SLA requirements and any deviation from the SLA can be

reported. If the SLA will be breached, the CBR model will then be used to find the best solution for the situation. Figure 5.3 illustrates the monitoring process.

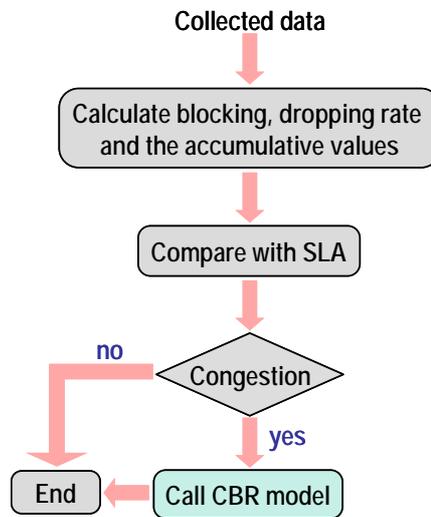


Figure 5.3 Monitoring process

### 5.3 Case Matching for Congestion Pattern Recognition

Figure 5.4 shows the CBR model (adapted from that described earlier in §3.3.2) as used in this simulation.

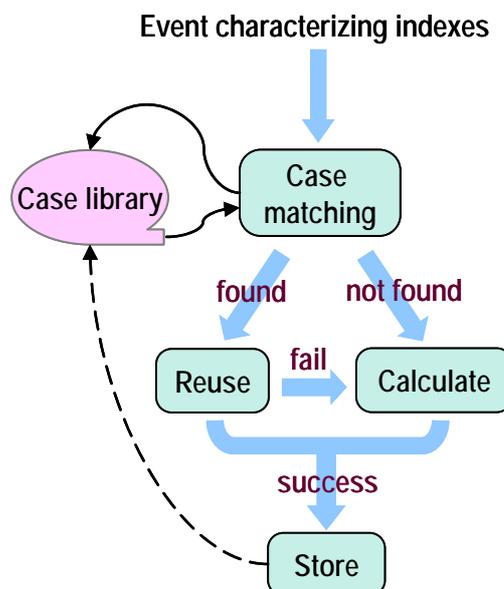
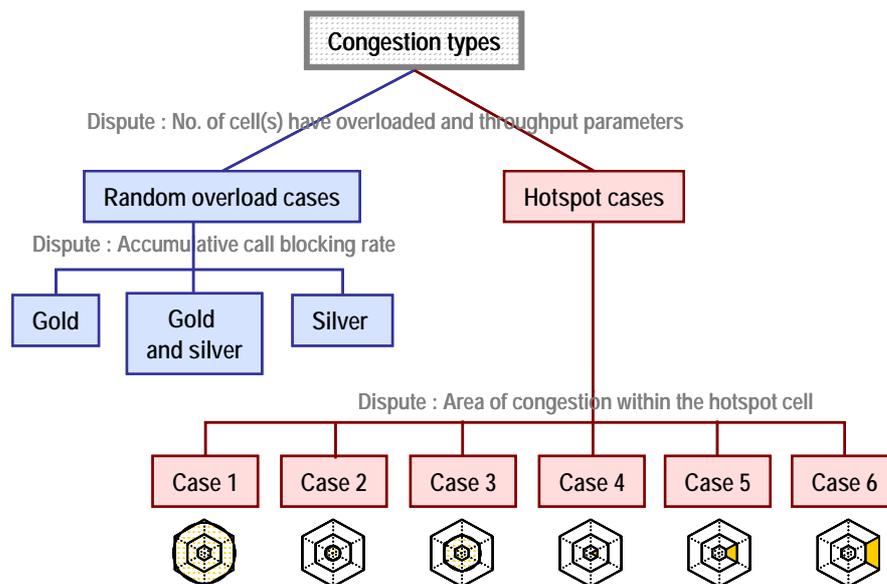


Figure 5.4 Case matching process using CBR approach

When congestion is reported, the characterising indexes of the event will be sent to the CBR model, which will try to retrieve a matching solved-case from the case library. If the exact match can be found, the solution will then be applied in the system. If an exact match cannot be found a calculation method (explained in detail later in §5.4 and in conjunction with the simulation results in Chapter 6) will be used to compute the solution. This will also happen when a matched case is found, but testing shows that it fails.

In general, CBR would retrieve the nearest matched case from the case library as mentioned in §3.3.2. and the results in chapter 6 show how well the matching works.

The shared-feature CBR approach has been chosen for this work as mentioned in §4.6. The case library structure is shown in Figure 5.5 All the random overload cases are clustered in one group and all hotspot cases are grouped together as they share the same main features<sup>5</sup>.



**Figure 5.5 Case library structure according to shared-feature network**

There are six characterising indexes to describe the main cases. As used here, they are obtained by matching the actual monitoring factors into a suitable range where the value belongs so that they are represented as small integer numbers. The factors are as follows:-

<sup>5</sup> In some parts of this thesis, the term *characterizing index* is used with the same meaning of *feature*.

- Total throughput for the whole system.
- Offered traffic for the whole system.
- Offered traffic for silver customers for the whole system.
- Offered traffic for gold customers for the whole system.
- Cell identity where the offered traffic exceeds random overload limit.
- Cell identity where the offered traffic exceeds hotspot overload limit.

From these six parameters, the system can differentiate the random overload case from the hotspot one.

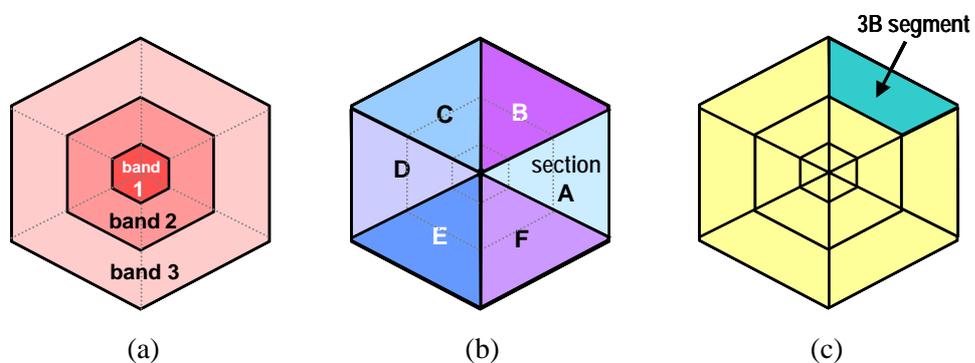
In the case of a random overload case, there are two extra parameters used to identify which sub case is occurring as follows:-

- Accumulative blocking rate for silver class.
- Accumulative blocking rate for gold class.

With a hotspot case the matching is more complicated and this is explained in the following section.

### 5.3.1 Methods for Case Matching in Hotspots

In this situation it is necessary to identify the *pattern of congestion across the cell* and to do that with a reasonable degree of approximation, the cell is divided into three concentric bands and six sectors (Figure 5.6). Using five bands was also tried, but the extra granularity did not offer any improvement, but did make for extra complexity.

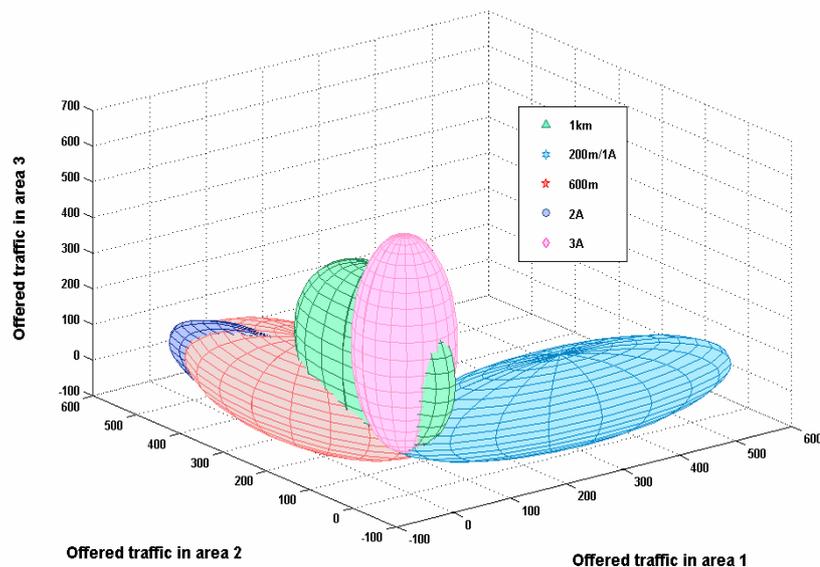


**Figure 5.6 Hotspot cell monitoring areas**

Initially the hotspot cases were represented by the case combinations shown in Figure 5.1.

In this approach, distributions of traffic are represented by ellipsoids in 3D space, with one axis for the traffic in each band (Figure 5.7). The system identifies which case a particular traffic pattern corresponds to by determining in which ellipsoid the point is located as each ellipsoid represents a particular hotspot case.

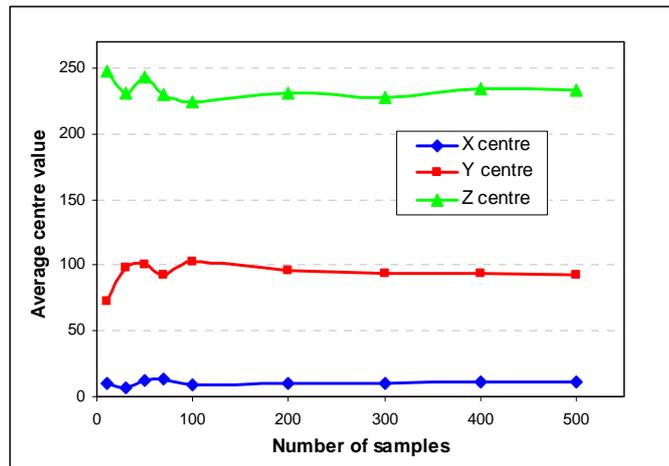
As can be seen, some ellipsoids overlap and this can lead to problems in choosing the best match. However by adding extra rules to the identification method (the distance to each ellipsoid centre and the previous monitoring case), the actual hotspot sub case can be more accurately identified.



**Figure 5.7 Hotspot case identification represented by ellipsoids**

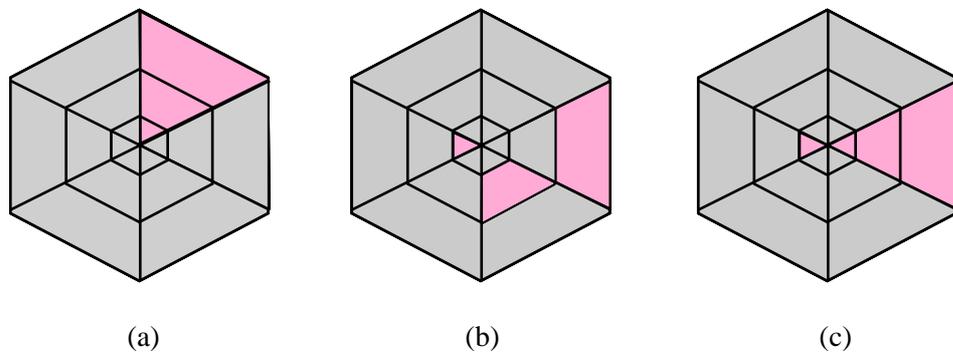
The centre of each ellipsoid is obtained from the average value of each particular case and each ellipsoid represents the maximum coverage for each case.

These ellipsoids are generated over a large number of sample periods, sufficient to ensure that a stable situation has occurred. From Figure 5.8, it can be seen that by 200 samples the parameters (in this case the centre of the ellipsoid) are stable. 200 samples have, therefore, been used.



**Figure 5.8** The effect on collecting different sample size in order to obtain ellipsoid parameters

The problem with this approach is identifying cases beyond the six used to represent the first samples. This can be quite complex and the ellipsoid approach is too simplistic. For example, Figure 5.9(a) will have the same ellipsoid *created* as Figure 5.9(b) and Figure 5.9(c) will not be easily matched at all.

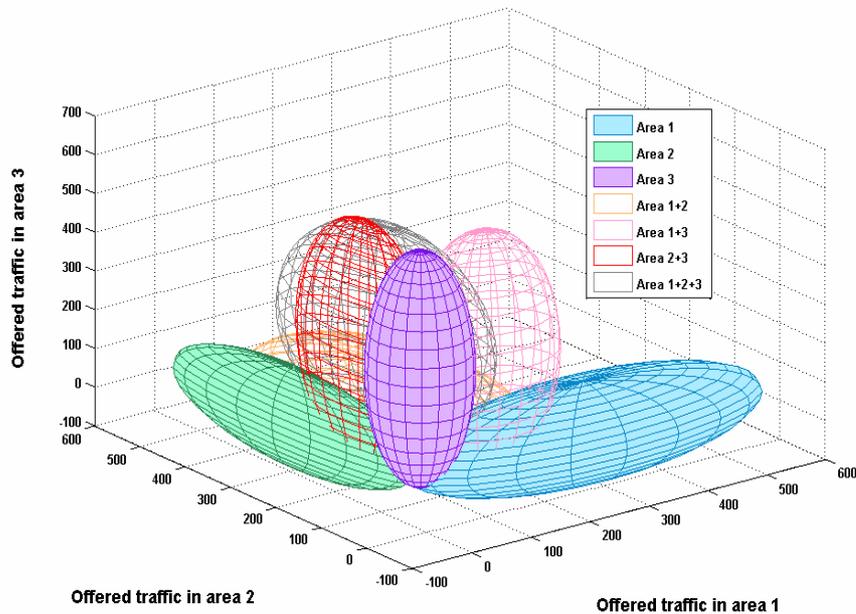


**Figure 5.9** The layout of three different hotspot cases

What is required, therefore, is a system that looks not only at the total traffic in each band, but also at the sector in which it is located.

A two-step hotspot pattern identifying method has then been introduced. The first step in the process is to identify which band has congestion; the second step will look into the congested *bands* and determine the *segments* that have congestion.

The ellipsoid approach can still be used, but instead of creating each ellipsoid for each congestion pattern, one is created for each combination of congestion patterns across the bands. There are  ${}^3C_1 + {}^3C_2 + {}^3C_3 = 7$  combinations so that seven ellipsoids have to be generated and these are shown in Figure 5.10.



**Figure 5.10 Ellipsoids represent congestion pattern according to the hotspot cell area of congestion**

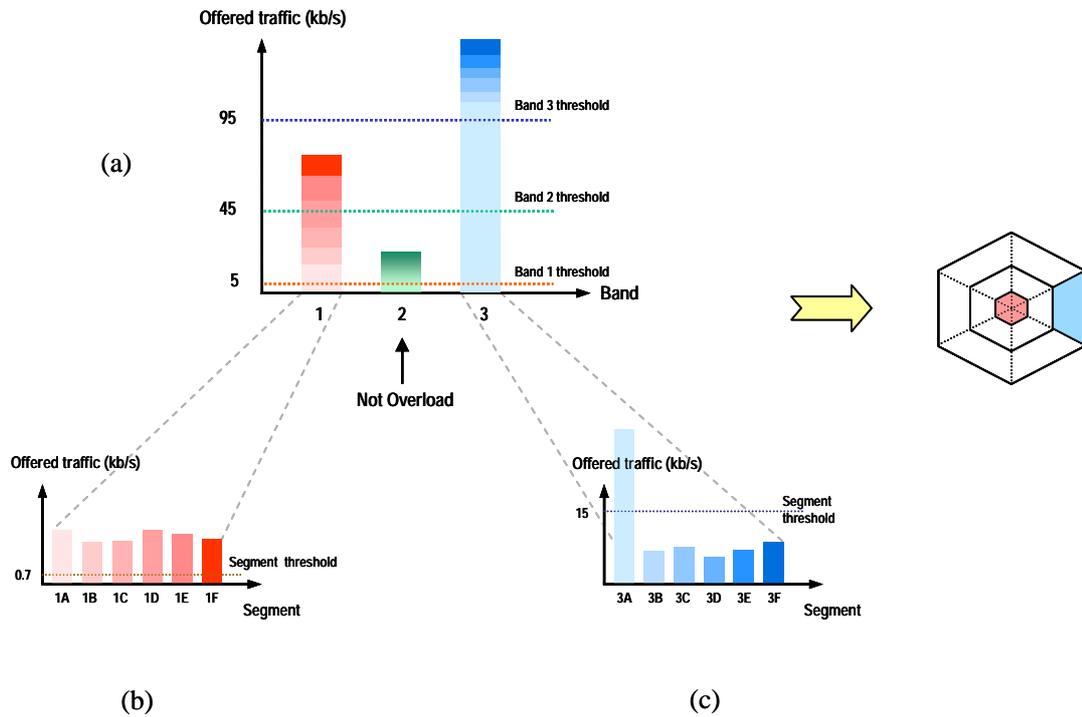
An alternative, less complicated is to use a threshold for the offered traffic in each segment. The following section gives the detailed explanation of the method.

### ***5.3.2 Chosen Case Matching Method***

From the cases, the maximum acceptable offered traffic for each band is determined and is used as the first stage process to identify the band of congestion when the value is exceeded.

The second stage is to look within the overloaded bands to find which segments are congested.

Figure 5.11 illustrates the mechanism with an example.



**Figure 5.11 A two-step hotspot pattern identifying method**

The first step is shown in Figure 5.11(a). A threshold is set for each band and it can be seen that bands 1 and 3 exceed their threshold while 2 does not. Hence the two congested bands are broken down into segments with a separate *segment threshold*, to identify which segments are congested. (Figure 5.11(b) and 5.11(c))

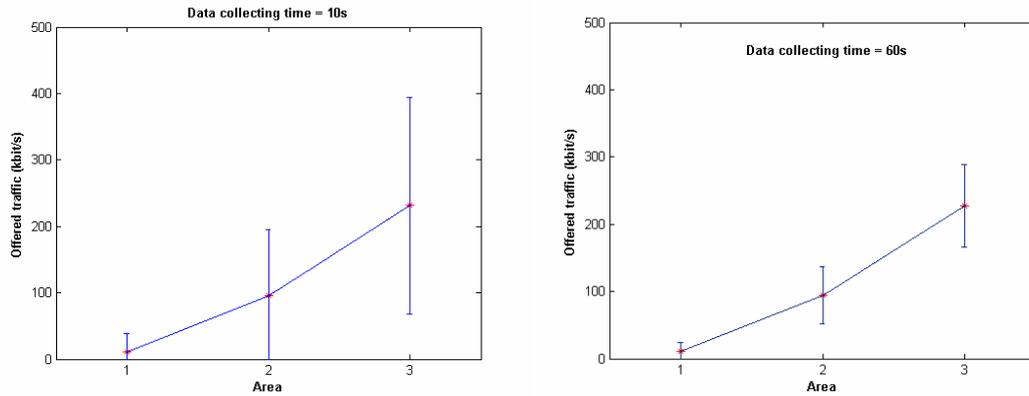
By using this two-step method, there is no need to look in detail at uncongested bands.

These two methods have been tested and the results show that the ellipsoid method is more complicated and does not give better accuracy. As can be seen in Figure 5.10, some parts of ellipsoids overlap each other, which mean more conditions need to be added otherwise unique decision making is not possible.

As a result of the previous analysis, the two-step threshold method is used for the results shown in this thesis.

Later on as the work being developed, a greater variety of hotspot area patterns was introduced and simulated. The monitoring time was then increased from 10s to 60s to

give a better averaging process. Figure 5.12 shows that there is less variability with the longer monitoring time.



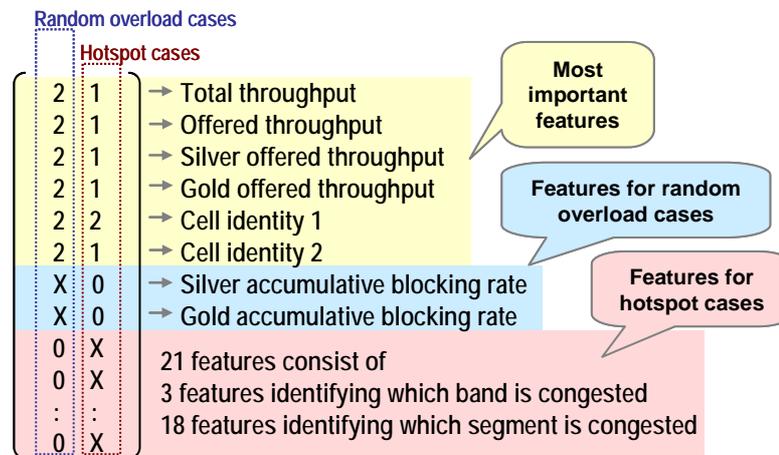
(a) Monitoring time 10s

(b) Monitoring time 60s

**Figure 5.12 Comparison of the offered traffic (kbit/s) plot between the system under monitoring time 10s and 60s**

On the other hand, because the aim of the controlled system is to recover from a congestion situation fast, the monitoring time is a compromise between better representation of the congestion condition and response time for the customers. 60s appears to be a reasonable compromise.

To summarise this section, Figure 5.13 presents the final case library structure for complete congestion pattern recognition.



**Figure 5.13 Case library structure**

## 5.4 Summary

This chapter presented the simulation environments as were tested here including the monitoring process and the case matching for the congestion pattern recognition, which is implemented when the congestion being detected.

To complete the CBR model as mentioned in §5.3, the calculation method is set up. This process will be exploited if the case matching process fails to offer the solved case or the solved case is not suitable for the new congestion. Here, the simple version of calculation method is created in order to examine the system under congestion however not one of that it experienced before. The specific rules to determine a solution according to the character of new congestion pattern is developed by referring to the knowledge from the case library. In this work, the calculation method is aimed to provide the solution for the congestion pattern that has been specified as a hotspot case nonetheless not one of the cases exists in the library.

In the next chapter, the simulation results and discussions for all congestion patterns studied here will be given. The explanation on how the calculation method being created will also be given in Chapter 6 along with the simulation results since the construction needs to refer to the solution from the existing hotspot cases.

# Chapter 6 Simulation Results and Analysis

This chapter shows the results from the simulation and these are explained in the context of a number of scenarios.

## 6.1 Simulation of Changing Reactive Layer Policy

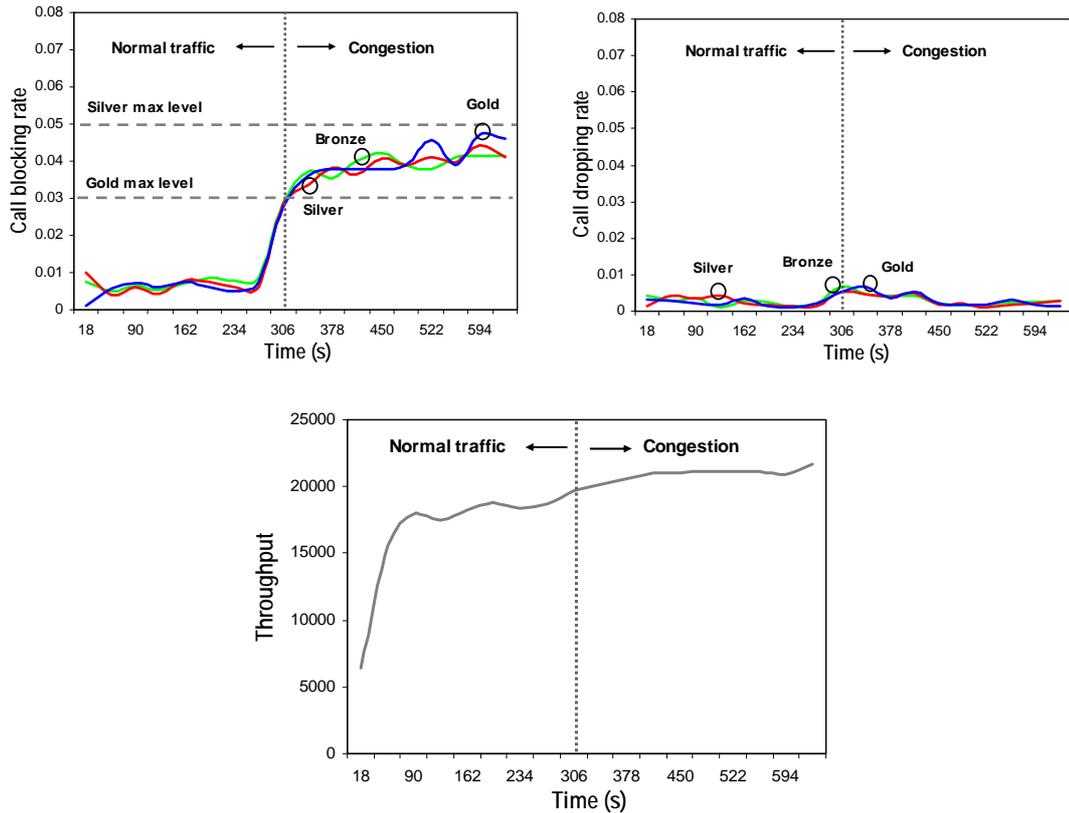
In this section, simulation results will be presented, to show the basic features of changing the reactive policy in the NPRA. Initially, the system consisted of 25 hexagonal cells and the traffic model was limited to voice service only for this part of the simulation.

The result of changing the reactive layer policy when congestion occurs is observed when a homogeneous traffic distribution is applied in the simulation environment. At the beginning, the traffic was kept to a normal level and as the system reached stability, the traffic load was increased to create a homogenous congestion pattern to the system.

The traffic load was controlled by the mean inter-arrival time between connection requests. At the beginning, the mean inter-arrival time was set at 100ms, which gives an acceptable traffic load. At some point into the simulation (after it has reached stability), the mean inter-arrival time was changed to 25ms in order to increase the traffic load. Figure 6.1 shows the simulation results as the traffic load was changed using a conventional system that does not change any reactive-layer policy.

The complete set of results includes *call blocking rate*, *call dropping rate* and *throughput* and these are plotted over the simulation time. Congestion is defined by the SLA assumptions in §5.1.

With normal traffic load, the call dropping rate has values in an acceptable range, but after increasing the load, the call dropping rate increases, then slowly declines to about the same level as before. The reason for this is because the ideal assignment and admission control model will block new arrivals as the system congests rather than dropping existing ones. Hence, the call blocking rate will increase as a greater number of mobiles attempts to connect.



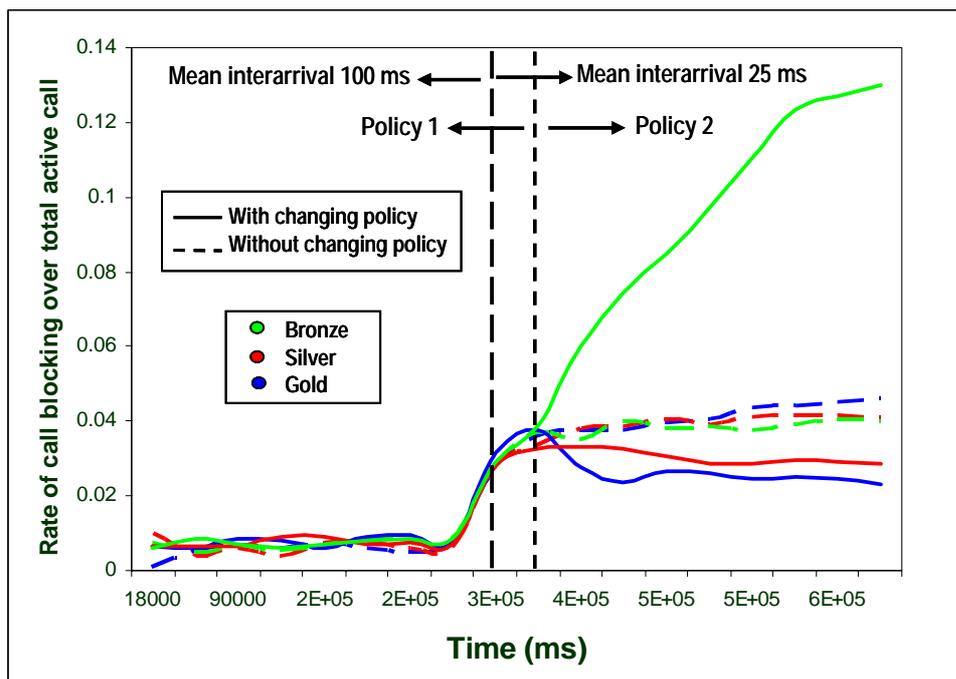
**Figure 6.1 The simulation results from conventional system**

Figure 6.1 shows that the call blocking rate for all customer classes increases when the traffic load increases as there is no change in policy: the blocking rate for gold customer exceeds the SLA constraint. From the start, the call buffering time (or call setup time) for all classes of customer and all types of service has been set to zero to give immediate accept or reject decisions for the system.

The main reason for showing these results is to verify that call dropping does not need to be considered as a factor in the later simulation results: the value is an order of magnitude less than blocking rate, and the inherent nature of the system is to favour existing connections at the expense of new ones. This is why, for simplicity, the traffic has been taken as voice only.

To investigate the effect of changing the reactive policy, a set of simulations were performed where the reactive layer was changed manually: this was to show that there was a possibility of using different cases to deal with different situations. In these test runs, video and data services were added to the traffic mix, as described in §4.1.

The same simulation scenario was rerun and this time a change in policy was made when congestion occurred. Figure 6.2 compares the results for the call blocking rate achieved from the conventional system and the one with the change in policy. The solid lines illustrate the effect of changing the reactive layer policy, a change that would normally be under the control of the agent and the dashed lines are from the conventional system. This example shows that a change in policy can be used to differentiate the service offered to customers.



**Figure 6.2 Simulation result showing the effect of policy change**

The implementation here uses a buffer mechanism so that a call request that cannot be served immediately, especially for the higher priority customer, is held for a short time in case resources become available (for example when another connection is completed, or another mobile moves out of the cell). The buffering method used here is priority buffering with a configurable buffering time.

Here, gold customers have the largest buffering time of 8s follow by silver customers with 4s, but the buffering time for bronze customers is still set to zero.

As shown by the solid lines in Figure 6.2, the policy is changed at the point of congestion to implement the above. This increases the probability of gold and silver

customers being accepted at the expense of bronze. This simulation shows just one possible strategy that could be adopted, but illustrates that by using intelligent control of these strategies, it should be possible to manage performance as required.

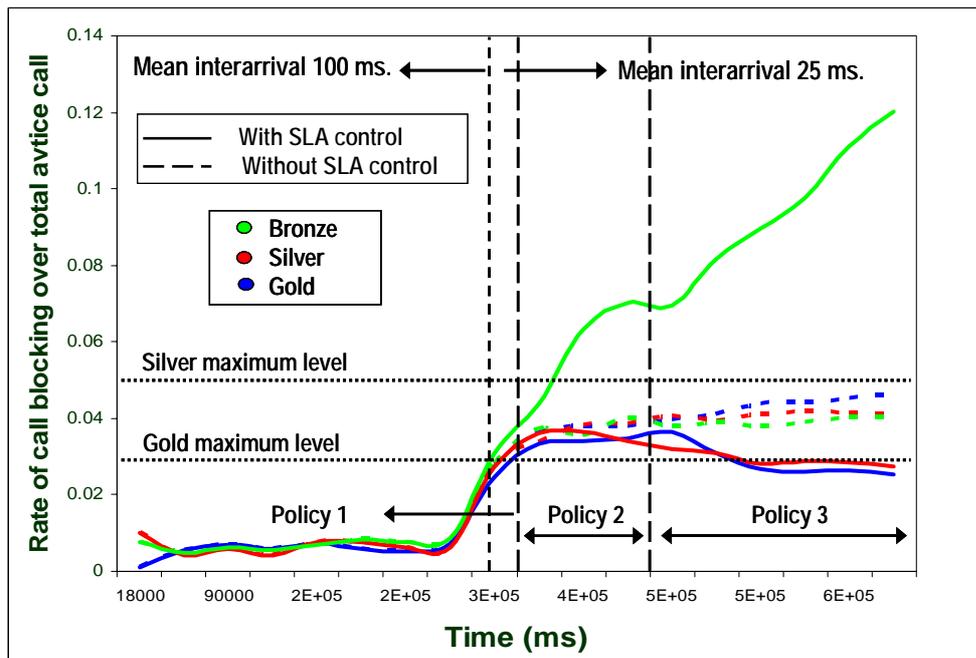
When discussing this simulation, it was mentioned that congestion was applied at some point after the system had reached stability. The question is, therefore, the definition of stability and how long the simulation should be run before any changes are applied. In this work, “reaching stability” means that with constant traffic load being applied, the system capacity (the indicator used here is *throughput*) reaches a stable level. This is considered in more detail in §6.2.

The main concern of the work is to cope with a sudden congestion event as soon as it occur so the length of the simulation time after a new event being applied is not part of this stability consideration, just the time to ensure that the system has settled down after it has started running.

A further set of basic simulations was performed to demonstrate that applying a change in policy coupled with continuous monitoring can actually control the behaviour in such a way that SLA requirements can be met.

The same simulation environment was run, but this time, SLA tracking has been implemented to monitor the result and to control the policy. Figure 6.3 shows the comparison between the result obtained from the conventional system and the result from applying SLA-based control for the call blocking rate as the traffic load increases.

Again, the result from conventional system is plotted as dashed lines while the solid lines represent the result from the SLA-based control system. Here, the call blocking rate limitation for the gold and silver customer classes is set up as part of the SLA requirement as mentioned in §5.1. The maximum acceptable blocking rate for gold and silver customers has been set to 0.03 and 0.05 respectively, since the gold customer pays the highest rate for the least elasticity of service level.



**Figure 6.3 Comparison between conventional system and the one with SLA-based control**

As shown by the solid lines, the call blocking rate for all customer classes increases after the increase in traffic load. At the point where the level of call blocking rate of gold customer exceeds the maximum level, policy 2 is applied which allocates a short buffering time to call requests from gold and silver customers (buffering time of 4s and 2s respectively, with that for bronze customers still being set to zero).

The results show that the call blocking rates of gold and silver customers stabilises, but that for gold does not fall below the maximum value specified by the SLA. After waiting for a short period (here the monitoring time is set to two minutes) to ensure the trend is stable, a further change in policy is applied; this gives longer buffering time for gold customers (8s) and slightly longer buffering time for silver customers (4s), so increasing still further the probability of gold and silver customers (especially gold) being accepted at the expense of bronze. This improves the QoS to meet the SLA and shows the importance of applying the right planning layer policy to the reactive layer when using an intelligent agent to maintain the SLA.

## 6.2 Determining Steady-State of the System

As explained, it is important to be able to assess when the system has reached a steady state after starting from a zero initial state. This section discusses how an assessment is made of when the system has reached that steady state.

The simulation is set up with a normal traffic load and the throughput value is collected throughout the simulation time. These simulations omit the data service because of the inordinate simulation time required if that is included, as discussed in §4.1. Indeed, for the rest of this chapter, the traffic model only includes voice and video service; in addition the system considers only 9 hexagonal cells since the interest is on one cell and the neighbours, but the wrap around technique is used to avoid the border effect. [SBP03]

The simulation time is set to 200s as for the results shown earlier in Figure 6.1 The simulation has been run 50 times to collect 50 samples and the data analysed and plotted as shown in Figure 6.4. The solid line is the average value from the 50 samples and the histograms with distribution curves are calculated from all data in each point.

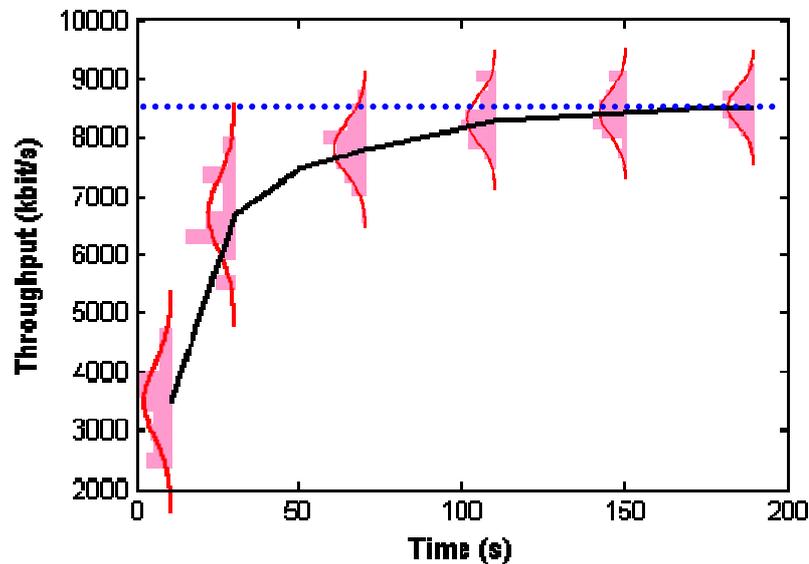


Figure 6.4 Throughput plot over simulation time

It can be seen that as the simulation time increases, the throughput value converges to a steady value represented by the dotted line; in addition, the variation becomes

smaller as the simulation progresses. This shows that at simulation time 200s the system is stabilised and it can then be modified to change the system environment.

## **6.3 Simulation of the SLA-Based Control System Using CBR Approach**

This section presents the generation of cases for the CBR model development followed by results from the simulation of SLA-based control system with the implementation of CBR approach.

### ***6.3.1 Generating Cases***

In this part, the generation of cases for the CBR model will be given (Detail on the case library structure was given in §4.6 and chapter 5). There are two main scenarios tested here: random overload cases and hotspot cases as explained in §5.1. Several methods have been taken into account and the configurations used to recover from congestion have been chosen based on experiments. Therefore, the solution for each case is achieved by selecting the solution that provides the best result. The following sections, §6.3.1.1 and §6.3.1.2 illustrate the solution for each case generated for the case library and also the results of implementing the solution for each congestion pattern.

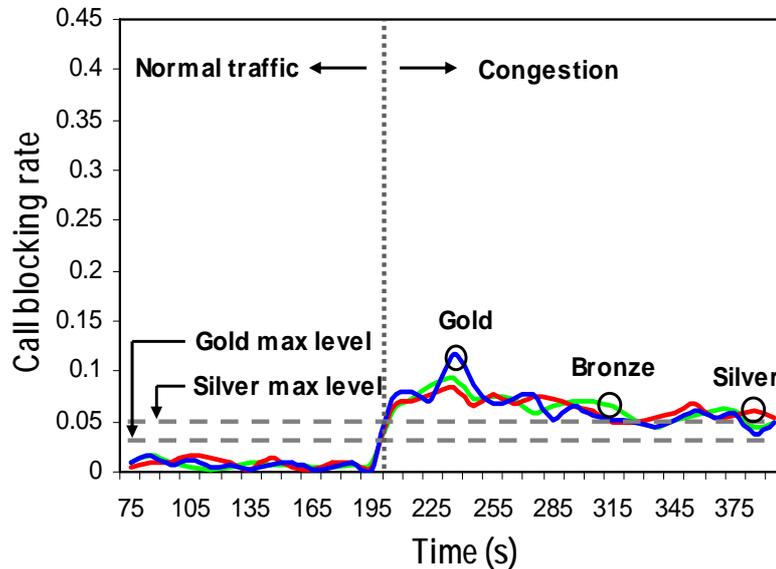
#### ***6.3.1.1 Random Overload Cases***

For this part, the simulation repeated the previous work. Figure 6.5 shows the simulation result of the call blocking rate over simulation time as the traffic load increases in a conventional system that does not change the policy. The call buffering time for all classes of customers have been set to zero to give immediate accept or reject decisions.

Figure 6.6, Figure 6.7 and Figure 6.8 show the effect of changing the reactive layer policy to the chosen one.

It might be thought that these results are simply the normal result of applying priorities, but the technique is more powerful. In many SLAs, it is not short-term violations that are important: an SLA might specify for instance that the blocking rate must not exceed a certain value during a day or a month.

The new policy has been applied to the reactive layer as soon as the system recognises congestion.

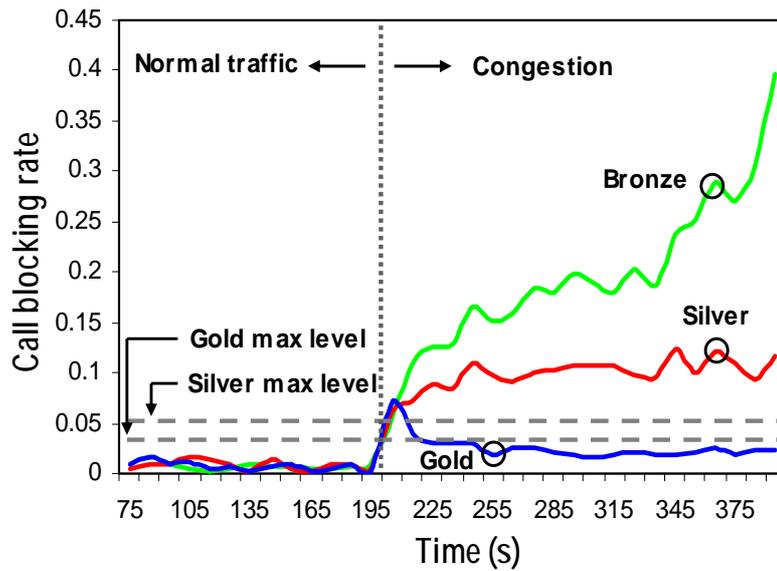


**Figure 6.5 Simulation result from the conventional system for random overload cases**

The implementation here again uses a buffer mechanism to give a short buffering time to call requests that cannot be served immediately, especially for the higher priority customers. The buffering time is configurable. It can be seen from the results that this application keeps the call blocking rate for gold and/or silver customers within the SLA bounds.

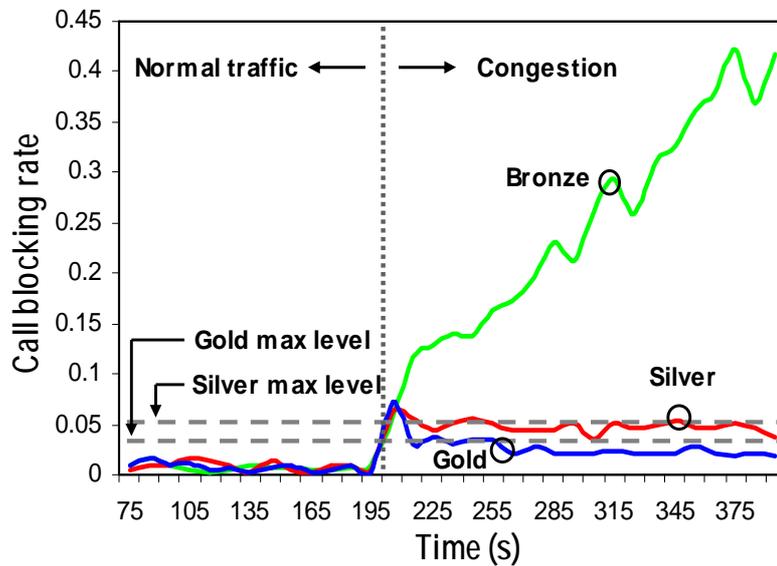
In Figure 6.6, the traffic reaches overload when the accumulative call blocking rate for gold customers exceeds the limit; at that point silver customers is still in an acceptable range. In this case the chosen policy gives the highest buffering time to gold (4s) and lower value for silver (2s) with that for bronze still at zero.

It can be seen that the system detected the overload situation at the point where the traffic load increased and the new policy was applied. As the new policy gives priority to gold, the call blocking rate for gold customer is maintained within an acceptable range at the expense of both silver and bronze.



**Figure 6.6 Simulation results showing the effect of applying a new reactive layer policy for the first random overload case**

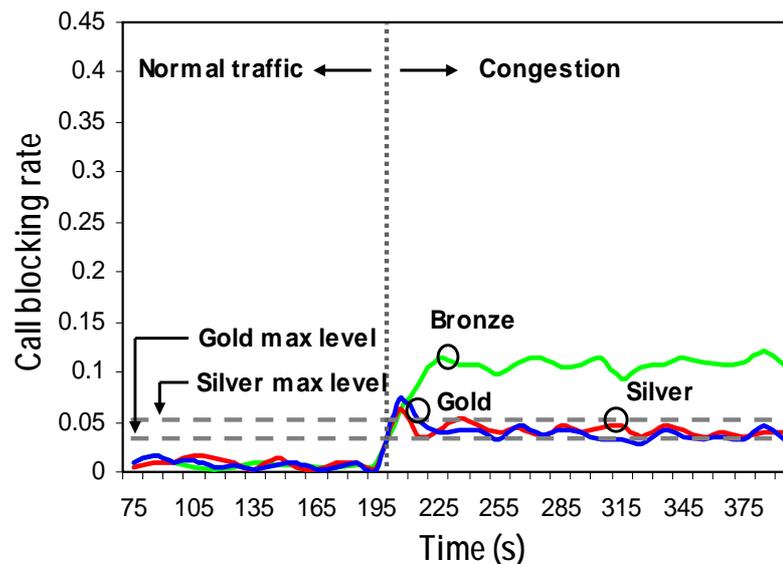
Figure 6.7 shows the result from the second case, where the traffic is overloaded with the accumulative call blocking rate for gold and silver exceeding the maximum value.



**Figure 6.7 Simulation results showing the effect of applying a new reactive layer policy for the second random overload case**

In this case, both silver and gold QoS need to be handled. By giving the highest buffering time to silver and slightly lower for gold, the blocking for both can be kept within the range. As the buffer in this implementation uses the priority arrangement, gold customers are always in the top of the queue, so, in order to also give some priority to silver customers, their buffering times have to be higher. (The buffering time for gold, silver and bronze customer have been set as 2s, 4s, and 0s respectively)

In Figure 6.8 the situation is that the long-term value for gold customers has been met comfortably, but that for silver is at the limit. When congestion occurs, silver customers have to be given priority in order that their long-term blocking is not exceeded, but gold customers can be allowed to have worse service since there is still “slack” in their SLA. The SLA monitoring here is looking at the long-term blocking, has detected that silver needs priority and has applied that priority.



**Figure 6.8 Simulation result showing the effect of applying a new reactive layer policy for the third random overload case**

In this case, the buffering times for gold, silver and bronze customer have been set as 0s, 1s, and 10s respectively. According to the implementation of priority buffering, gold customer are always placed in the front of the queue. Therefore, in order for bronze to also be maintained, the highest buffering time was given to bronze and a smaller amount to silver, so maintaining silver customers within acceptable range.

These results show the flexibility of the control system, which assigns different policies to different scenarios and also shows that the highest priority traffic can also be a sacrifice in order to maintain a long-term SLA for customers who would normally have lower priority.

In fact any SLA that can be evaluated numerically can be used as the basis for controlling the policy: the system is that flexible.

### 6.3.1.2 Hotspot Cases

The six hotspot cases have been described in §5.1.2. In this section, the simulation results will be classified by the solution mechanism or policy offered by the case generating process and the case number refers to the notation as shown in Figure 5.1.

Figure 6.9 shows an example of a simulation result obtained from the conventional system. In this figure, the scenario is hotspot case 1 where the hotspot area covers the whole centre cell. The comparison between the call blocking rate achieved from the hotspot cell with solid lines and the one from the normal traffic cells with dashed lines is shown in the Figure 6.9.

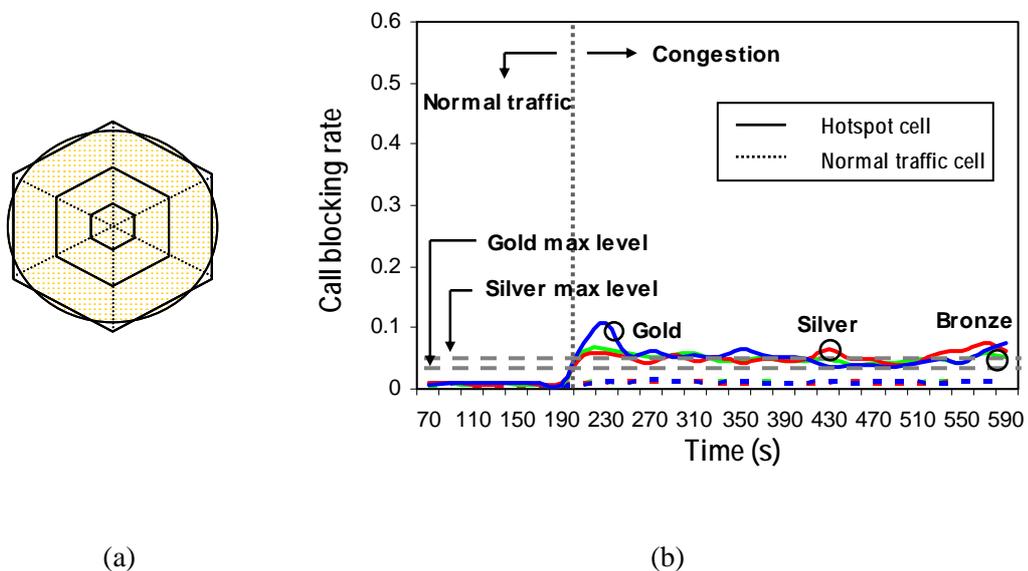


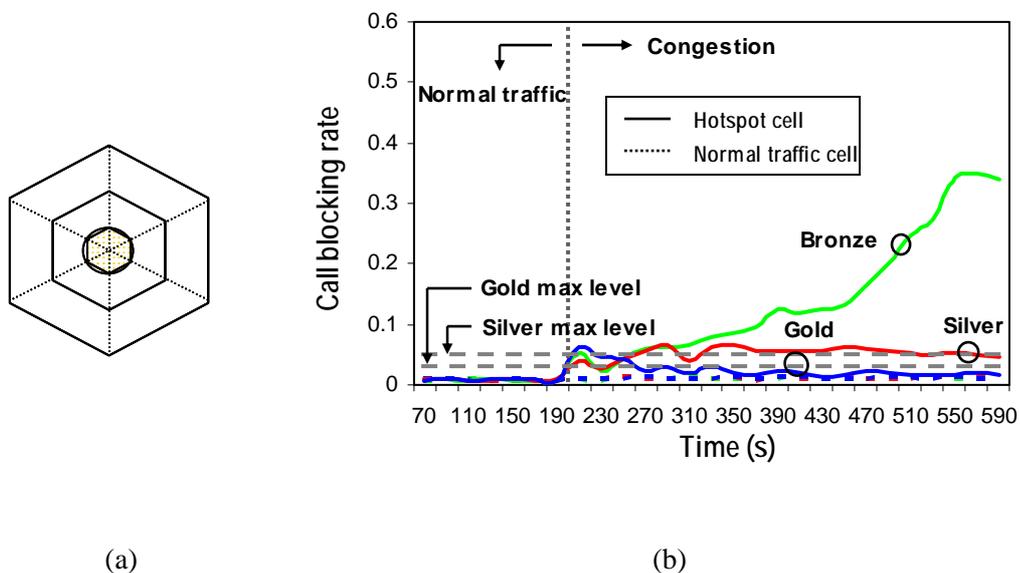
Figure 6.9 (a) Hotspot case 1 layout (b) Simulation result from the conventional system for the hotspot case 1

The graph shows that as the congestion occurs in the hotspot area of the hotspot cell, the call blocking rate for all classes of customer increases and results in the gold

blocking rate exceeding the limit, while the other cells with normal traffic still have an acceptable level of blocking rate. Because the congestion occurs in the centre of the cell, normal CDMA mechanisms have no real effect. This is the only example of result from a conventional system to be presented as the other hotspot cases have a similar trend.

In §5.3, it has been mentioned that the monitoring time of the system is increased from 10s to 60s to give more confidence in congestion pattern matching. In this section, the system is under new monitoring time. Hence, the simulation time is increased to observe longer results after congestion occurs.

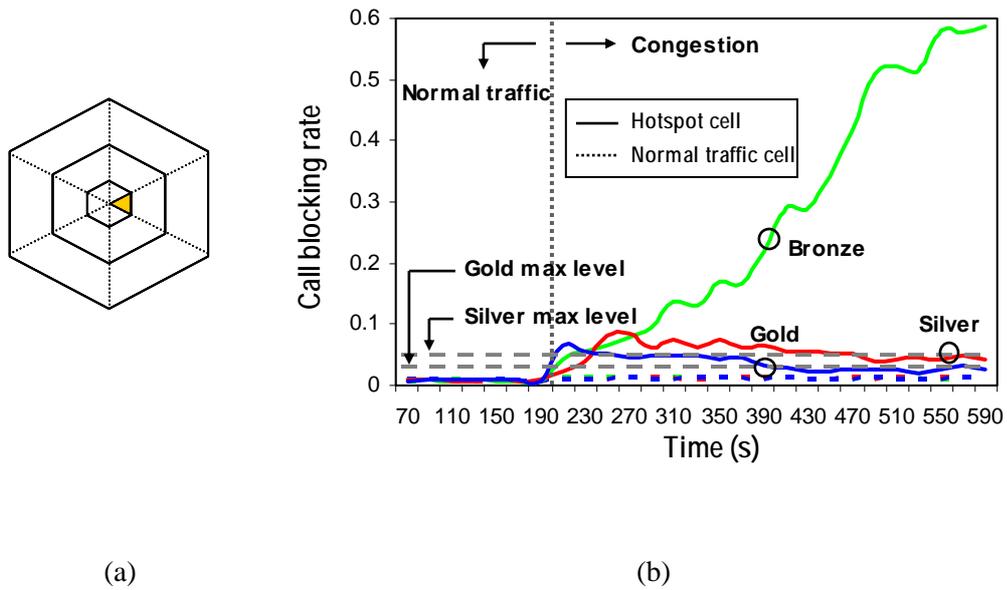
Figure 6.10 illustrates the result of using the chosen policy for hotspot case 2. In this situation, the chosen policy applies the buffer mechanism same as for the random overload case because shrinking the cell will have no effect with the users in the centre. Hence, different classes of customer can be prioritised. The buffering time for gold, silver and bronze customer were set as 4s, 4s, and 0s correspondingly.



**Figure 6.10 (a) Hotspot case 2 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 2**

This method of congestion control is the most suitable of the two used in this work for all forms of hotspot confined to the centre of the cell, so it is also used in hotspot case 4 (Figure 6.11).

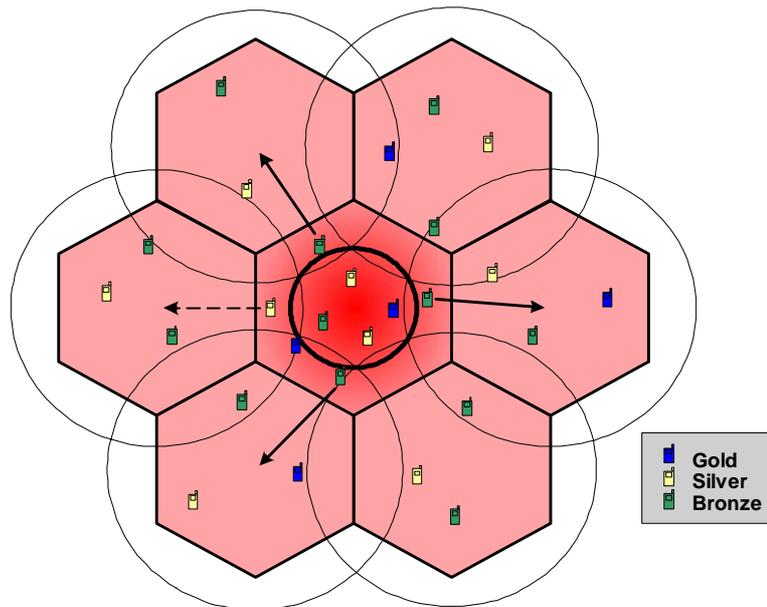
Note that although the same *mechanism* has been applied to two different hotspot cases, in setting up the case library, experiments were done to find the most suitable configuration for each case and in this case different buffering times ( $G=4s$ ,  $S=6s$ , and  $B=0s$ ) were found to most suitable. Again, with no buffering time, the bronze customers have no control over their blocking rate.



**Figure 6.11 (a) Hotspot case 4 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 4**

The next case is the hotspot area covering the whole hotspot cell. When this congestion pattern is detected, a new policy is applied and shrinks the size of the cell by transferring its users near the cell edge to the neighbouring cells; it also only accepts new connection requests within this area until the situation is back to normal and the default policy is applied. This process is shown in Figure 6.12.

Considering the buffer mechanism for this congestion pattern, customers could also be prioritised to maintain quality of service of higher priority customers. However, the method applied here, which transfers excessive customers to neighbouring cells, does keep all customer classes within the SLA requirements without the necessitation of any sacrifice.

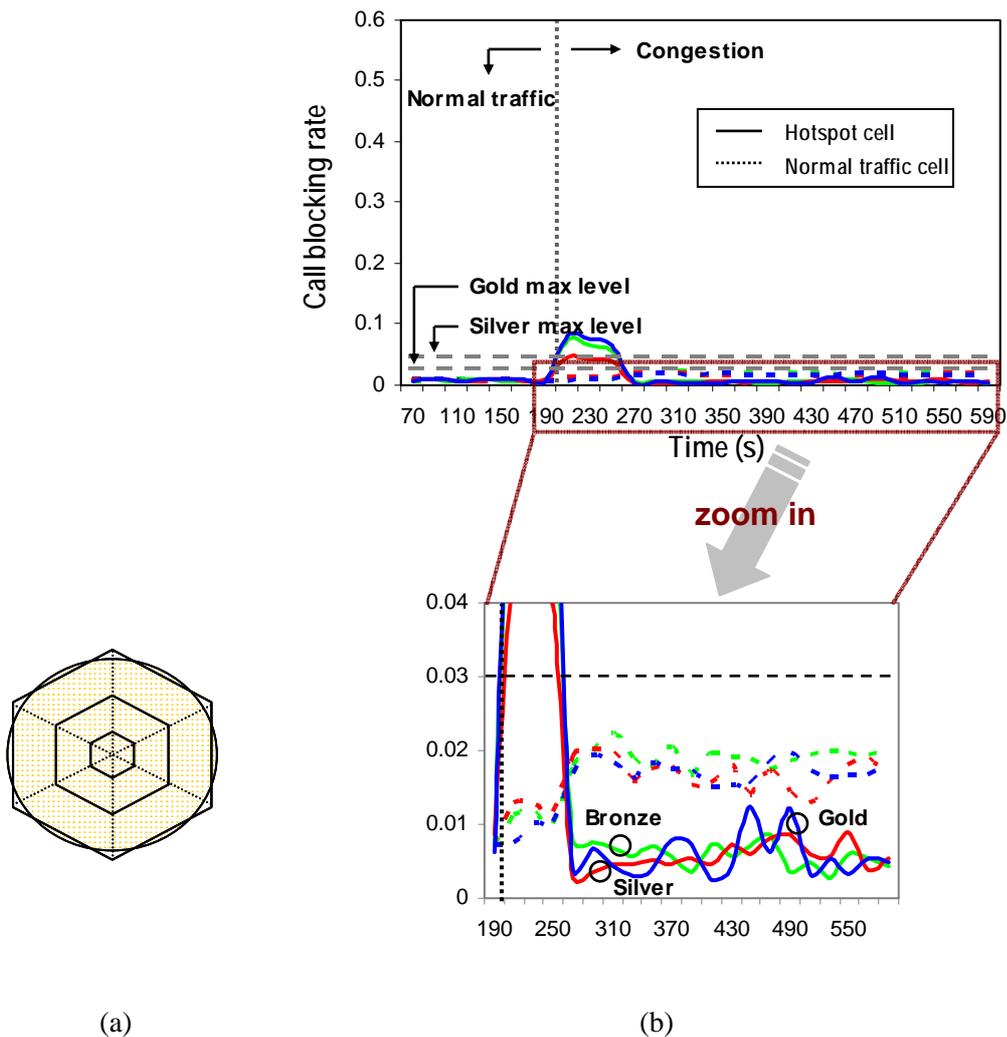


**Figure 6.12 Transferring method as centre cell shrinks**

The result from the SLA-based control system is shown in Figure 6.13. In this case, the hotspot cell is shrunk to radius 400m. The existing bronze and silver customers outside this area are transferred to the available neighbouring cells. The available neighbouring cell means the cell that has enough bandwidth to support the handover connections. The new connection request locates outside the shrunken area will also be made connection with the available neighbouring cell.

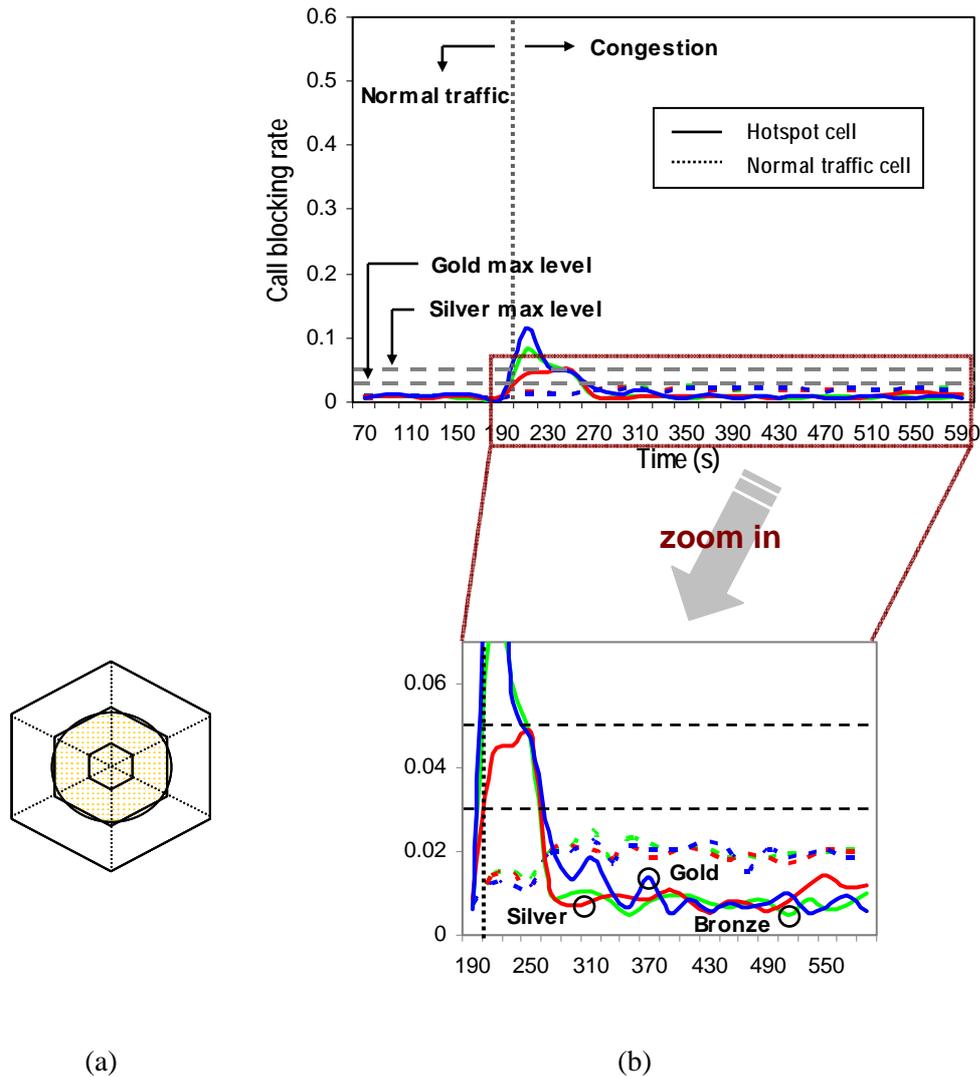
From Figure 6.13, the call blocking rate for all customer classes is kept within the limit as displayed by the solid lines with a small compensation from the neighbouring cells represented by the dashed lines. Comparing with the dashed lines in Figure 6.9, the values are higher but still acceptable.

As mentioned before, the slightly increase in transient occurred as smaller amount of data being monitored. In order to give credibility to the simulation results, at least 5 repeated simulations have been run for each case and the results are plotted from the mean value.



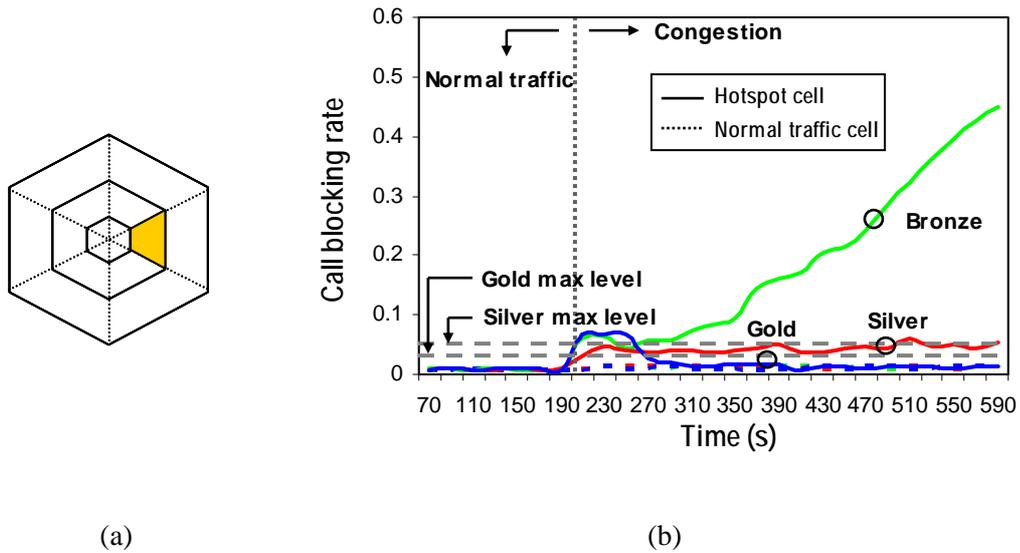
**Figure 6.13 (a) Hotspot case 1 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 1**

This method is also been found to be suitable for hotspot case 3 and the result is shown in Figure 6.14 again experiments have been done to find the most suitable configuration for this particular case, indexed as part of the case library, which is different from case1. For this case the centre cell is shrunk to radius 300m.



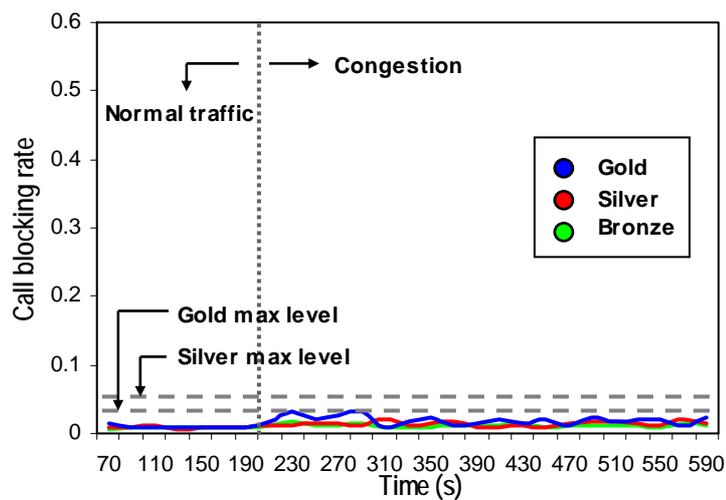
**Figure 6.14 (a) Hotspot case 3 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 3**

The next simulation result comes from the hotspot situation case 5. The method for controlling this case is a combination between cell shrinking and the buffer mechanism and the result is presented in Figure 6.15. The centre cell is shrunk to cell radius 600m. Therefore, some connections are transferred to the neighbours. Also the buffering time for gold and silver customers are set as 4s and 6s respectively which prioritised the remaining customers in the shrunken centre cell.



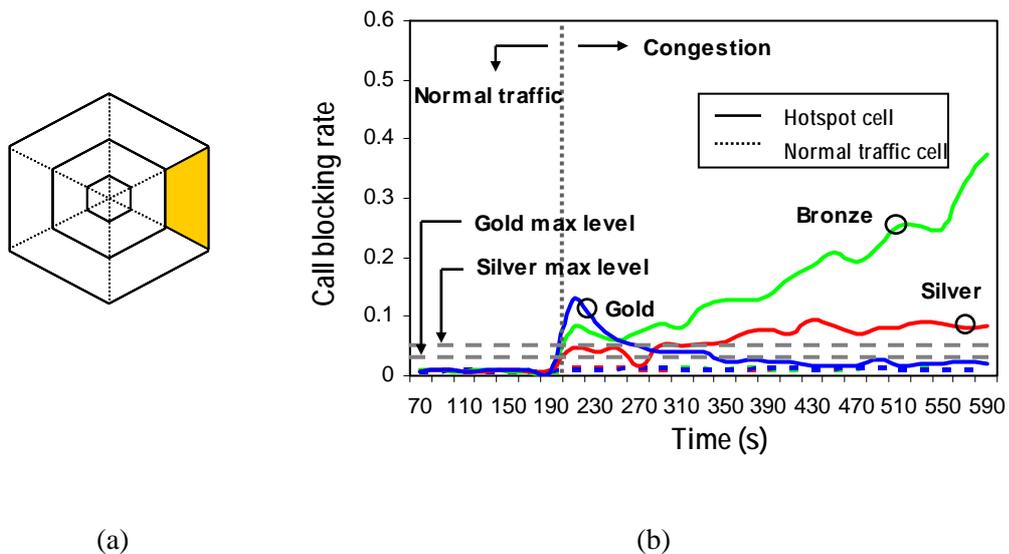
**Figure 6.15 (a) Hotspot case 5 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 5**

Figure 6.15 gives a similar result to that seen in the previous case as the system prioritises different classes of users. In this case, the nearest neighbouring cell to the hotspot area will be primarily affected as most of excessive connections will be transferred to. Hence, while testing the policy achieved from the case matching, that neighbouring cell must also be monitored to ensure that it does not overload by this transfer: moving congestion from one cell to another is no solution! Figure 6.16 shows the result monitored from the most effected neighbouring cell.

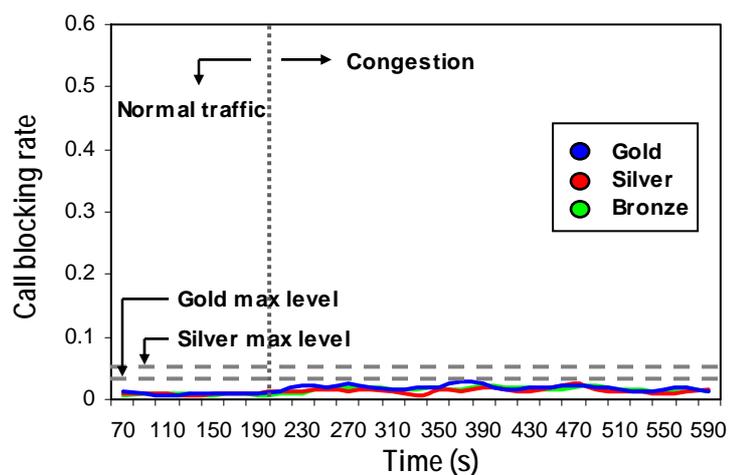


**Figure 6.16 Simulation result of applying the new reactive layer policy to the neighbouring cell for hotspot case 5**

This method is found to be suitable also for the hotspot case 6 where the congestion occurs in the area near the edge of the hotspot cell. The centre cell is shrunk to radius 900m with the buffering time of 5s for gold and 13s for silver customers. Figure 6.17 shows the comparison between the result from hotspot cell and the one from normal traffic cells as CBR model offered the new policy to solve congestion and Figure 6.18 shows the result of the most effected neighbouring cell as part of the monitoring process mentioned earlier.



**Figure 6.17 (a) Hotspot case 6 layout (b) Simulation result showing the effect of applying a new reactive layer policy for hotspot case 6**



**Figure 6.18 Simulation result of applying the new reactive layer policy to the neighbouring cell for hotspot case 6**

### 6.3.2 Result from the matching of the same case as existing ones in the library

From the generation of cases for the case library shown previously, the CBR model was built up. To check that it operates correctly, the simulation was run with the same conditions as used to generate the case to ensure it does match exactly.

Figure 6.19 shows one of the results, which is from the hotspot case 1. This time as the congestion was detected, the CBR model was used to match the pattern with cases in the library and the solution from case 1 was proposed. Comparing this result from the hotspot cell, solid lines, with the result from the simulation done while generating the case, dotted lines (from Figure 6.13), both show the same trend as the same method was implemented for the same congestion pattern.

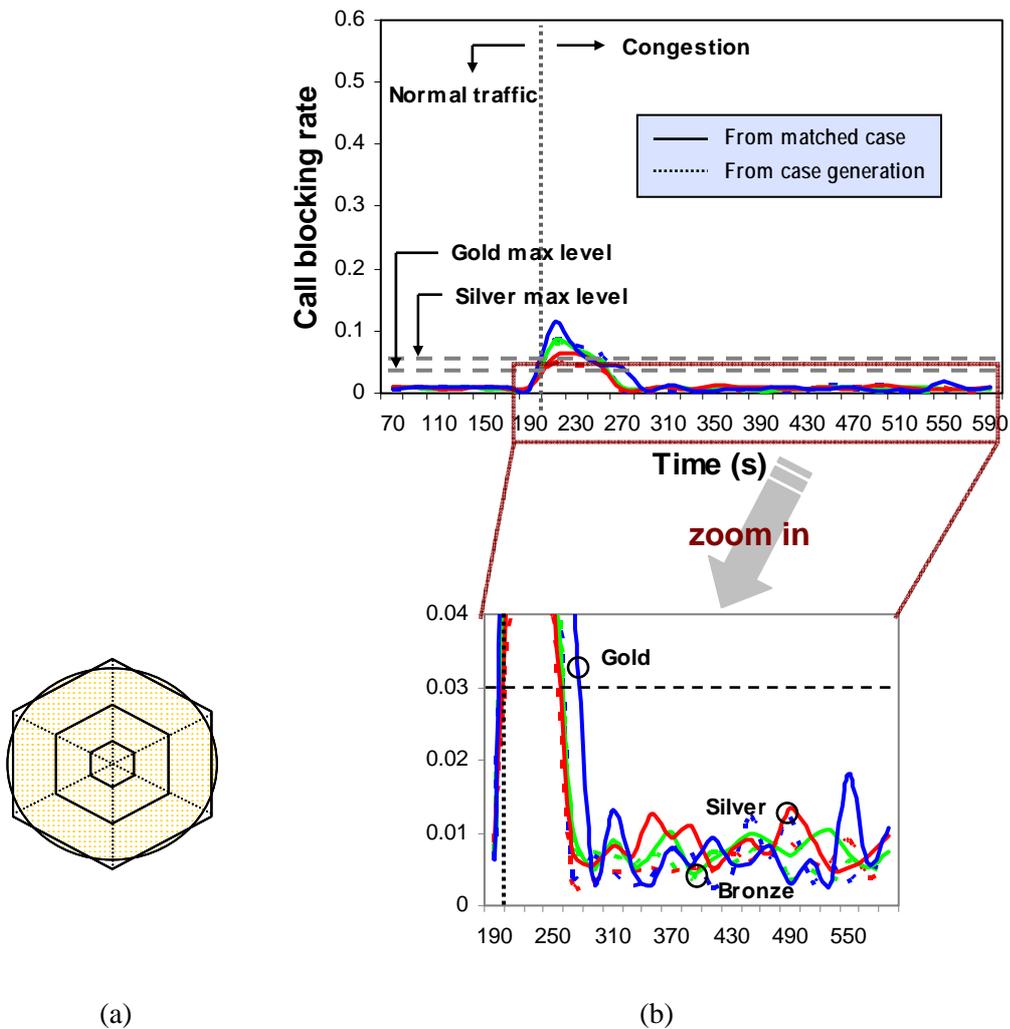


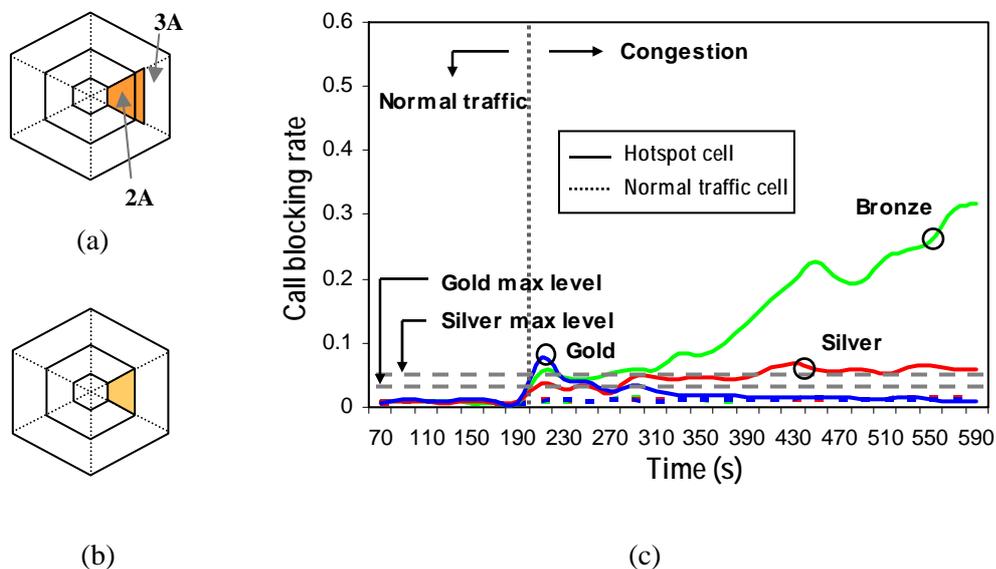
Figure 6.19 (a) Hotspot case 1 layout (b) Simulation result of the hotspot cell for the case matching compared with previous result

## 6.4 Examining the System Performance under Similar Congestion Pattern to the Existing Case

In this section, two experiments were carried out for two different congestion patterns in the hotspot cell, which are similar to the existing cases but not exactly the same. It was mentioned in §3.3.2 that in the first step, CBR retrieves the nearest matched case from the library and reuses its solution. This means that where an exact match is not available the solution from a solved case that is close enough could also be effective if the new event is sufficiently similar. Here, the two experiments are designed to observe the system in this aspect.

### 6.4.1 First experiment for the similar case that has additional congestion area across bands

This first experiment is for the congestion in the hotspot cell as shown by the cell layout in Figure 6.20(a). As can be seen, the congestion pattern is similar to that existing in the case library (case 5), shown in Figure 6.20(b). Figure 6.20(c) shows that simulation result as this congestion pattern was applied.



**Figure 6.20** (a) Hotspot cell layout (b) Hotspot cell layout of the matched case (c) result for the similar case with additional congestion area across bands

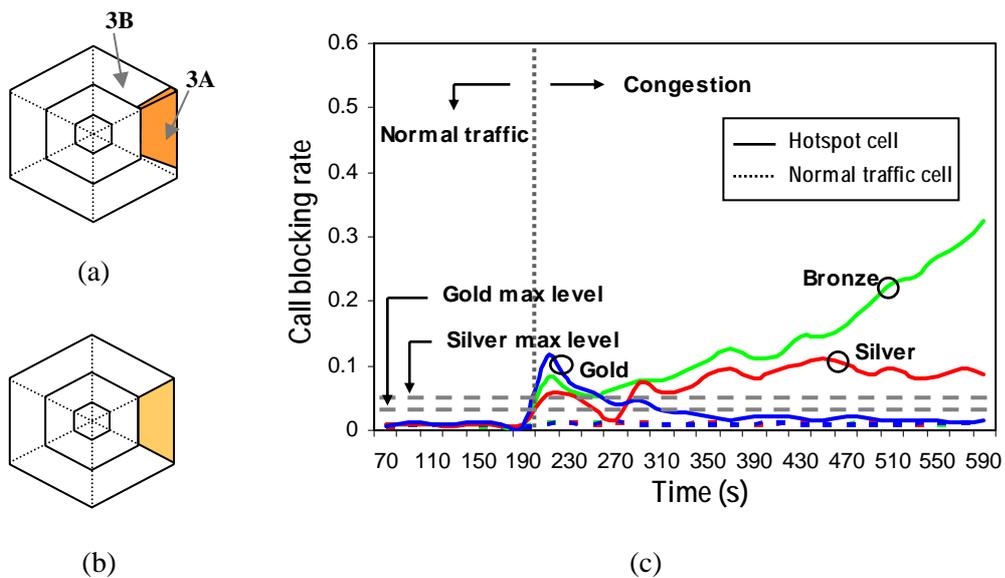
This same solution as the previous case 5 shown in §6.3.1.2 was offered by shrinking the centre cell to cell radius 600m and altering the buffering time for gold and silver customers to 4s and 6s respectively. The reason is because the main congestion area

(segment 2A according to Figure 5.6) was recognised by the monitoring process while the other congestion area (part of segment 3A) was not significant enough and therefore, did not cause the offered throughput in band 2 to exceed the threshold set at the CBR model.

The simulation result shows that the proposed method from the closest matched case solve the problem by transferring parts of connections to neighbouring cells and also prioritising the rest of customers according to their classes.

#### 6.4.2 Second experiment for the similar cases that has additional congestion area across segments

A more sensitive case is investigated in this section. The layout of the hotspot cell examined here is shown in Figure 6.21(a). The main congestion area is now in segment 3A with a small congestion area as part of segment 3B. (This area covers 5% of the whole segment area)



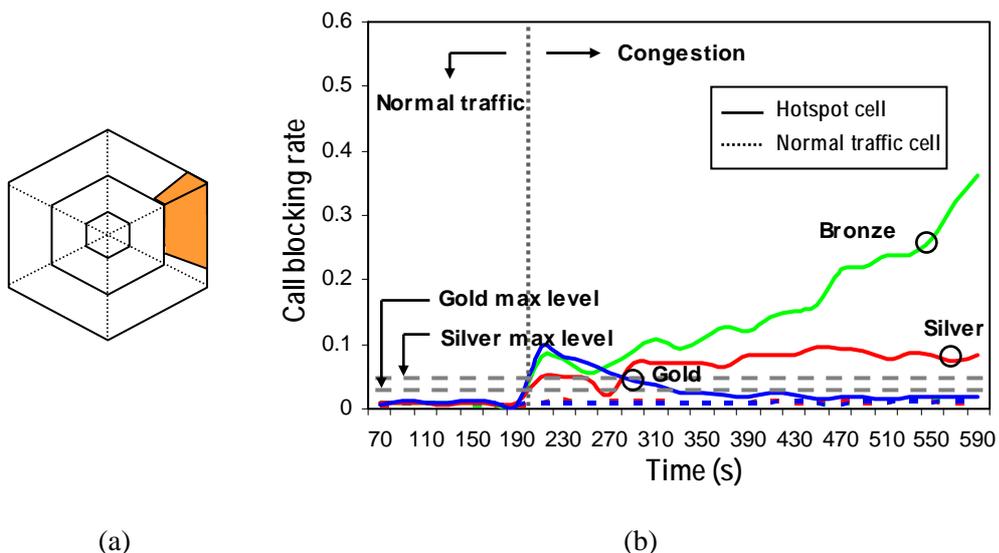
**Figure 6.21 (a) Hotspot cell layout (b) Hotspot cell layout of the matched case (c) Result for 5% of segment is an additional congestion area across segments**

Again, this time the congestion in the segment 3A is recognised, however not that within segment 3B. As congestion was detected, the solution from previously solved case 6 (Figure 6.21(b) shows the hotspot cell layout) was proposed. The centre cell is shrunk to radius 900m with the buffering time of 5s for gold and 13s for silver customers. The simulation result is shown in Figure 6.21(c).

The next step is to increase the congestion area within segment 3B to cover 25% of the whole segment. This time as congestion is reported, the CBR model could not find the match, hence used the calculation method to propose the solution.

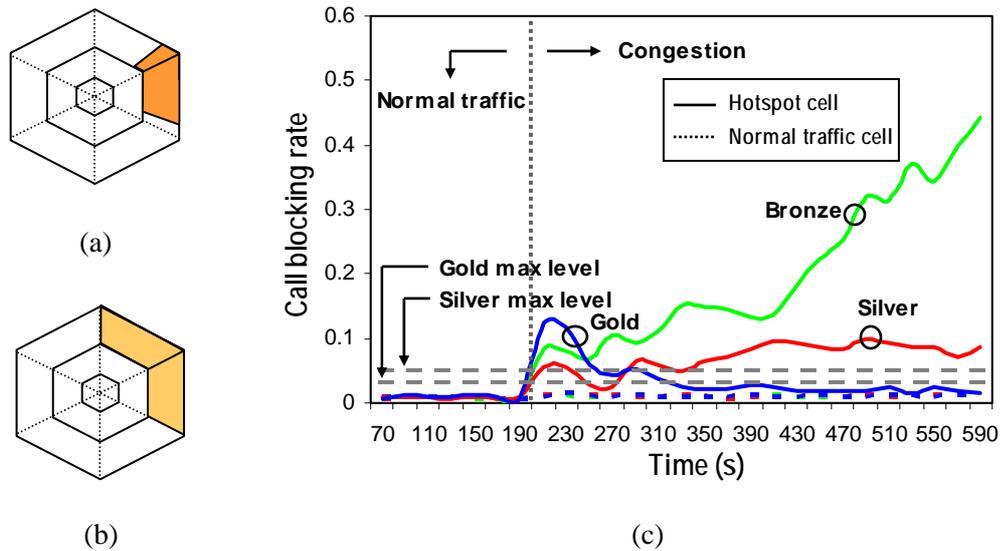
To complete the CBR model, the final step is to develop the calculation method in order to cope with an event that cannot be matched with the existing cases. The previous solved cases and methods were used to give an idea on developing the calculation method, which consists of a set of rules. More detail will be given in the next section, where the investigation on unfamiliar cases is presented.

Here the system was not able to find matched case, hence used the calculation method. Nonetheless, this case has been offered the same solution as before and the simulation result is illustrated in Figure 6.22(b).



**Figure 6.22 (a) hotspot cell layout (b) Simulation result for 25% of segment is an additional congestion area across segments**

As a result of this test, the new case has been indexed into the library as one of the hotspot cases, which hotspot area includes segment 3A and 3B, with its solution. After that, the same test was done again to investigate whether the system can recognise this congestion pattern. The simulation result can be seen in Figure 6.23(b) and it shows the obvious similarity between Figure 6.22(c) and Figure 6.23(c) as the same method has been proposed. The only difference is that this time the system did find the best match and straight away offered its solution.



**Figure 6.23** (a) Hotspot cell layout (b) Hotspot cell layout of the matched case  
(c) Result for a repeated simulation by matching with the new case

## 6.5 Examining the System Performance under Unfamiliar Congestion Pattern

This section considers how to use the CBR method when unfamiliar cases are detected. The following section will illustrate the process in setting up the rules for the calculation method. After that, §6.5.2, §6.5.3, and §6.5.4 will present the simulation result from the system tested against unfamiliar cases.

### 6.5.1 Setting up rules for the calculation method

In order to test the system when faced with cases very different from those in the case library, a calculation method was developed consisting of a simple set of rules based on the studies done while generating cases for the library and also from the solutions of all cases. Figure 6.24 summarises the existing cases and their solutions.

Figure 6.25 shows the rule-based algorithm for the CBR calculation method.

Case	Method	Shrunk radius	Buffering time (s) – [Bronze, Silver, Gold]
1 	Cell shrinking	400 m	[ 0, 0, 0 ]
2 	Buffer mechanism	Remain same Cell size	[ 0, 4, 4 ]
3 	Cell shrinking	300 m	[ 0, 0, 0 ]
4 	Buffer mechanism	Remain same Cell size	[ 0, 6, 4 ]
5 	Hybrid	600 m	[ 0, 6, 4 ]
6 	Hybrid	900 m	[ 0, 13, 5 ]

Figure 6.24 Existing cases and their solutions

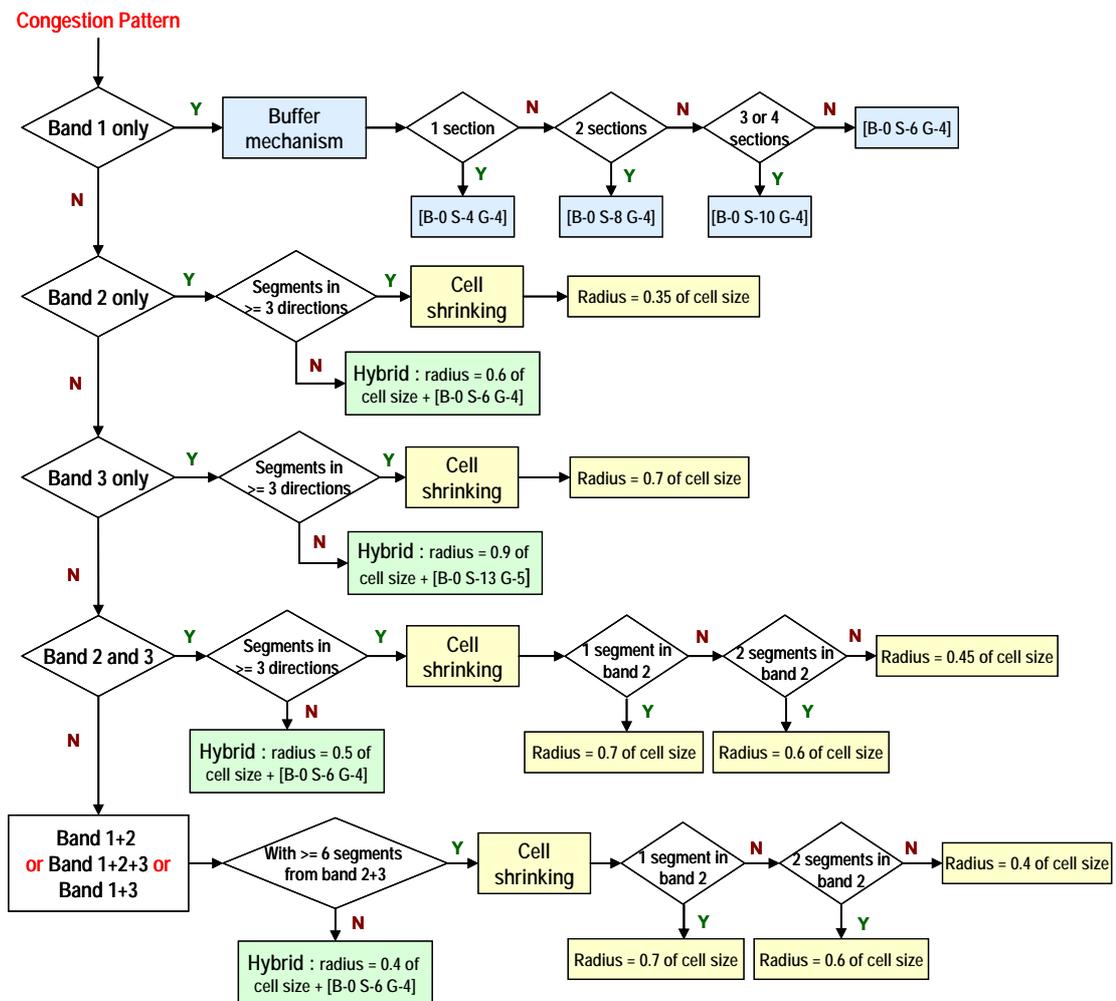


Figure 6.25 Summary of the rule-based algorithm for the calculation method

In this work the main focus is to show that the CBR approach is a viable technique for resource management. When there is no matched case the system needs to calculate a solution and to demonstrate this, a simple set of rules was developed for the calculation method as part of the CBR model. Note that the aim here is to demonstrate the capability of CBR in this application, not to refine the rules, so that the rules have been made deliberately simple, yet effective enough to provide the system with a suitable policy.

From Figure 6.25, as the new congestion pattern occurs, the pattern of load distribution across the bands is first used to differentiate the new event. It was explained previously that the buffer mechanism is only used in cases where congestion is confined to being near the cell centre. Hence, in the rules, the buffer mechanism will always be used when there is congestion in band 1: either as a single solution with congestion only in band 1 or as part of a hybrid solution. Buffer configurations were set partly by referring to the solution from the existing cases (cases 2 and 4) and the experiments done.

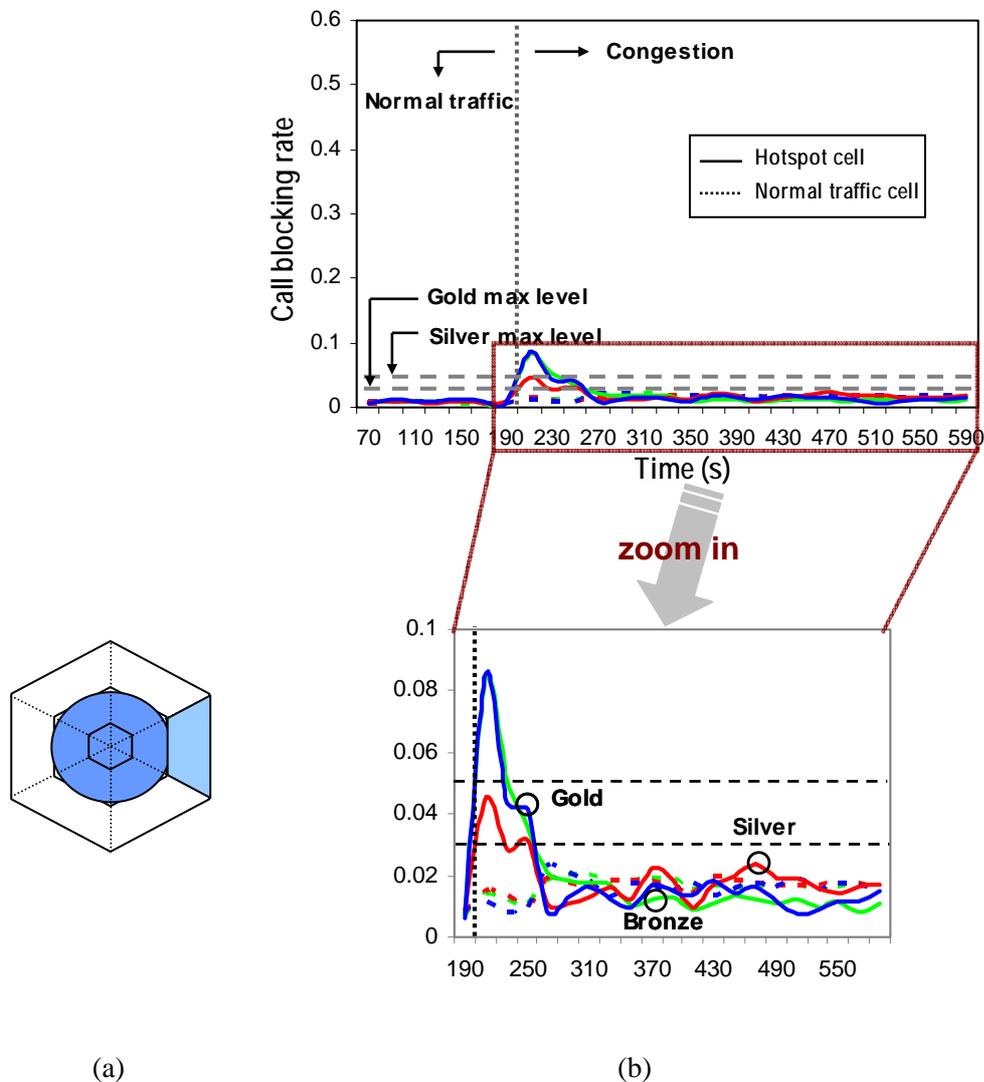
When congestion is located only in band 2 or band 3, both cell shrinking and the hybrid method were considered depending on the distribution of load. From the existing cases 1 and 3, it can be seen that the suitable method is cell shrinking when the congestion covers most of the directions around the cell. On the other hand, case 5 and 6 only have congestion in one direction, and the most suitable method is the hybrid method. Here the distribution of load over the cell was a main factor; the distribution of load is used to determine whether the cell shrinking or hybrid method should be used. If the congestion covers at least three segments of the band the cell shrinking method will be chosen, otherwise the hybrid method will be used. The new radius for each case was set up by considering the layout of congestion itself as well as the solution from existing cases. For the buffer configuration, the figures are mainly generated from the previous solutions.

A similar situation happens for the pattern where congestion areas cover a combination of areas across bands, the directions of load distribution again being used to identify which method should be used: cell shrinking or hybrid.

### 6.5.2 Cell Shrinking Method

As stated, this method is appropriate for the hotspot pattern where the congestion area is near the edge of the cell. This is because the excess connections can easily be transferred to the neighbours in all directions.

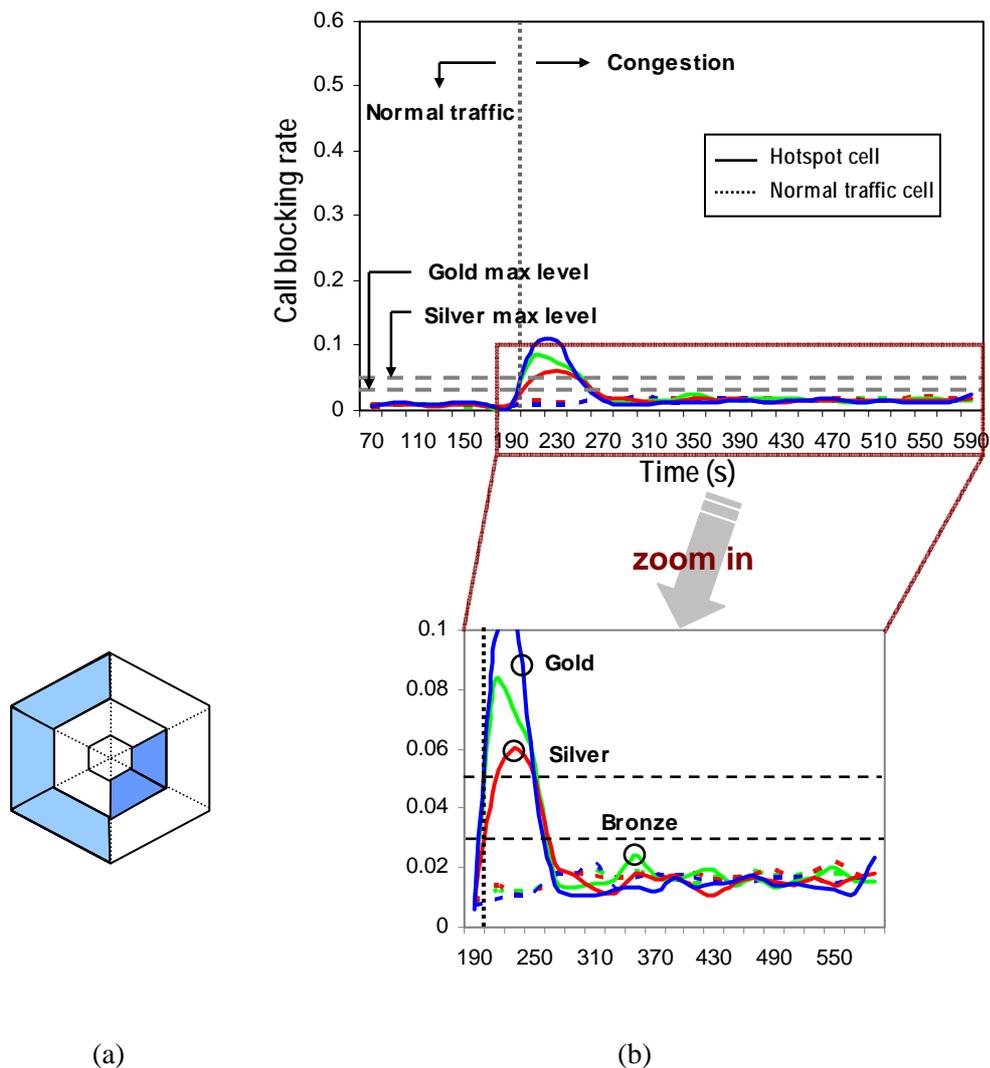
Figure 6.26, Figure 6.27 and Figure 6.28 demonstrate results from three different hotspot patterns in this category. The layout of each hotspot pattern is given on the left hand side of each figure. The graphs on the right hand of each figure show the call blocking rate as the congestion occurs for the hotspot cell with the solid lines and the normal traffic cells with the dash lines.



**Figure 6.26 (a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – cell shrinking method**

Although these three patterns have the same proposed solution from the calculation method, different configurations have been offered to each of them. The calculation method developed here does not simply find the most suitable solution but also offers the right configurations for each particular case. In Figure 6.26, the centre cell is shrunk to a radius of 400m.

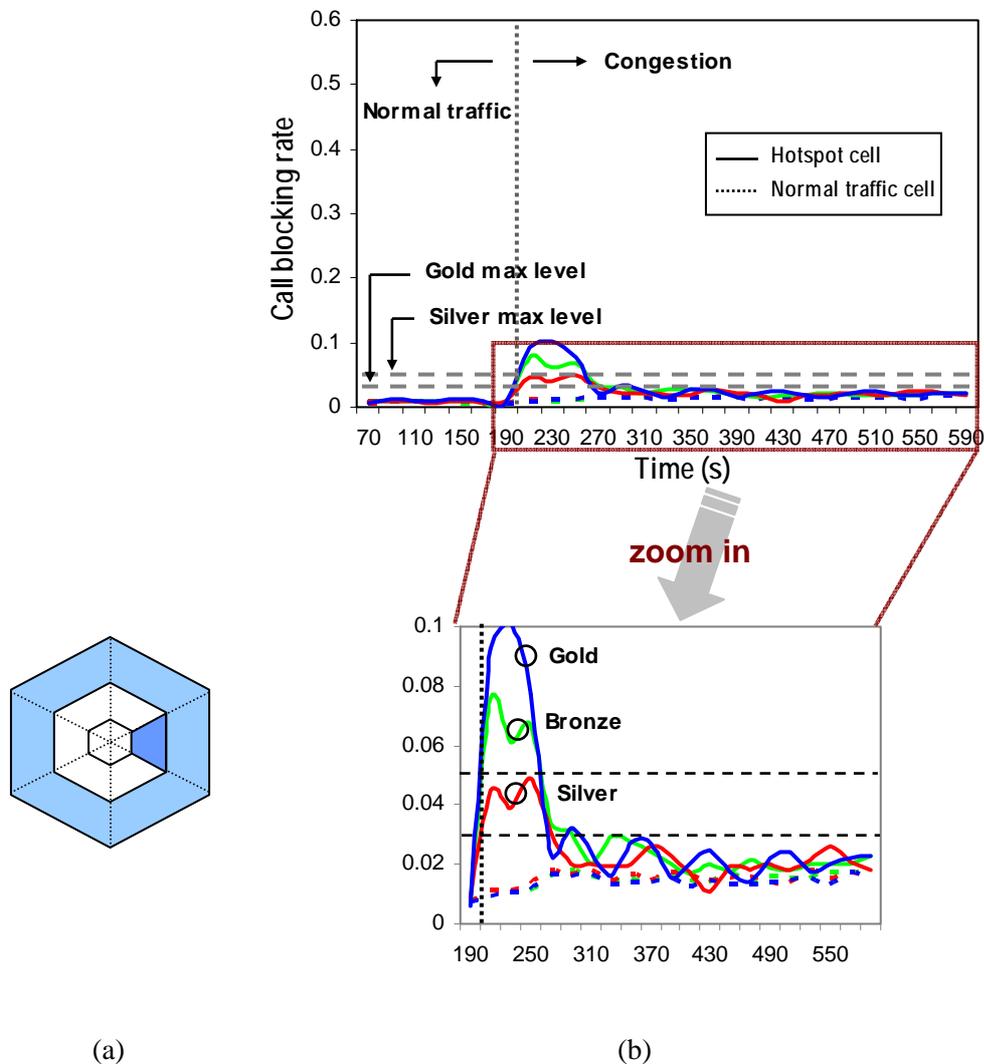
For the hotspot pattern shown in Figure 6.27, the centre cell is shrunk to a radius of 600m; in the pattern displayed in Figure 6.28, the centre cell is shrunk to a radius of 700m.



**Figure 6.27 (a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – cell shrinking method**

These results demonstrate the ability of CBR model to generate a suitable policy for the unfamiliar hotspot pattern: the pattern cannot be matched with ones in the library

so the calculation is used instead. They also show that as soon as the congestion is detected, a newly calculated policy is applied and the call blocking rate for hotspot cell has been maintained as well as that for the normal traffic cells.

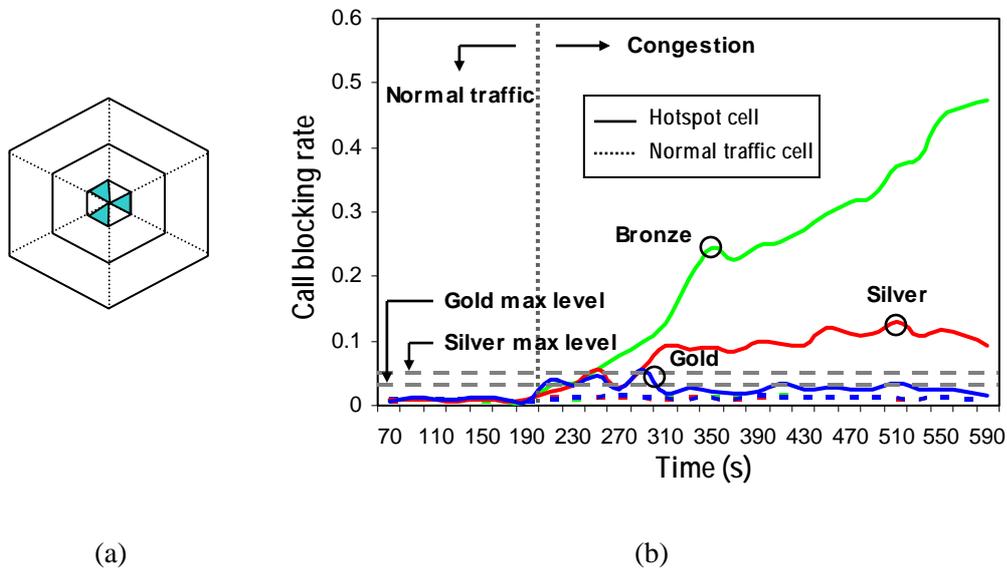


**Figure 6.28 (a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – cell shrinking method**

### 6.5.3 Buffer Mechanism

Figure 6.29 illustrates the result of an example of hotspot case in this category, where the congestion is near the centre of the cell. The hotspot pattern is shown on the left hand side of the result. In this case, the pattern consists of a few congestion areas scattered near the cell centre.

The solution proposed by the calculation algorithm is the buffer mechanism with the buffering time for gold, silver, and bronze set to 4s, 10s, and 0s respectively. Once again, gold as the highest priority class is maintained within its SLA requirement at the expense of bronze and silver customers.

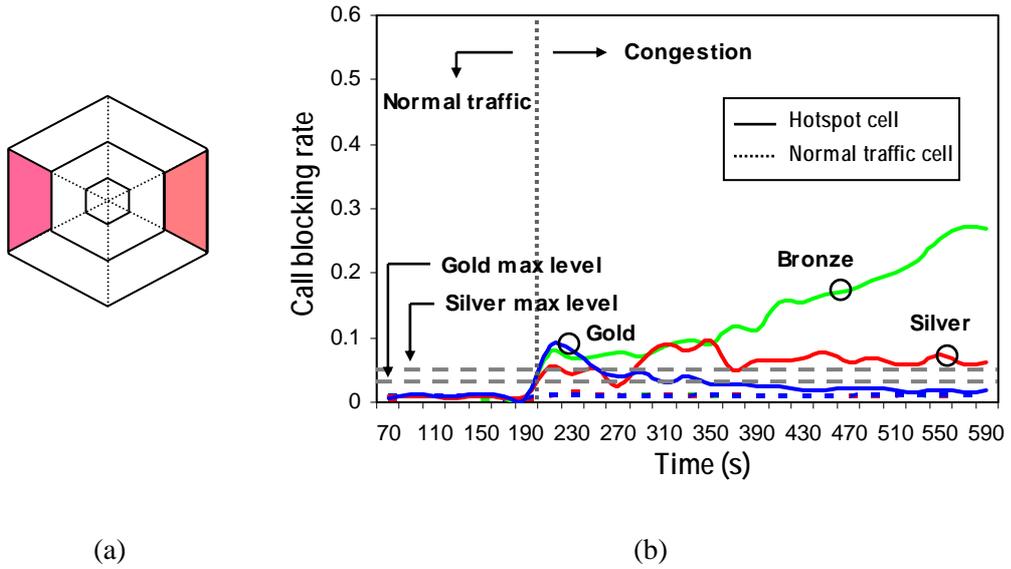


**Figure 6.29 (a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – buffer mechanism**

#### 6.5.4 Hybrid Method

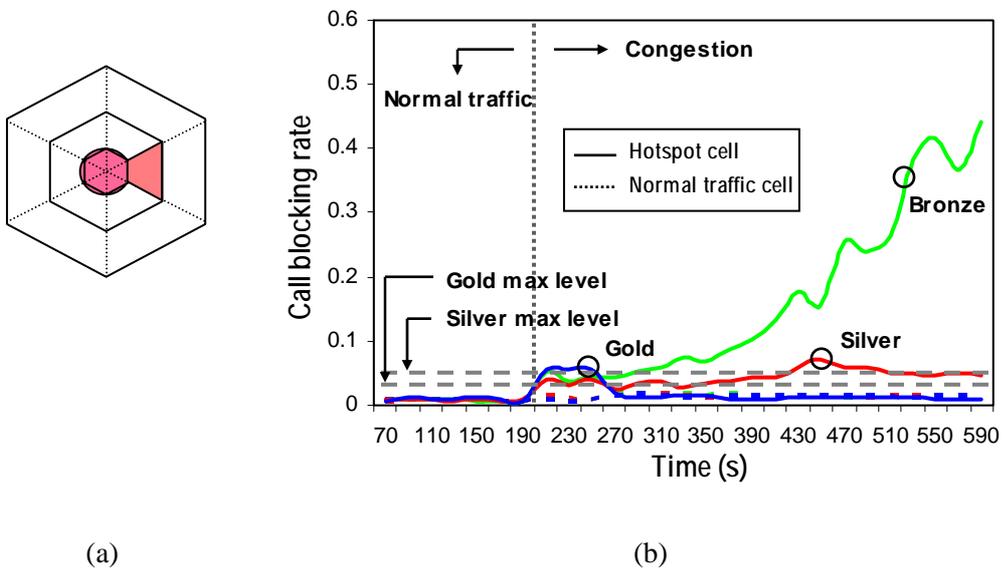
This method is a combination of the cell shrinking method and the buffer mechanism as used before in §6.4. It is suitable for a congestion pattern that has a combination of different areas near and far from the cell centre, or where the areas are near the boundary but not equally spread in all directions.

Three different hotspot patterns were investigated here and the results are illustrated in Figure 6.30, Figure 6.31, and Figure 6.32, again with the hotspot pattern on the left hand side of each result. Firstly the result shown in Figure 6.30: the CBR model proposes the hybrid method with the shrinking of the centre cell to a radius of 900m and buffering times for gold and silver set to 5s and 13s respectively.



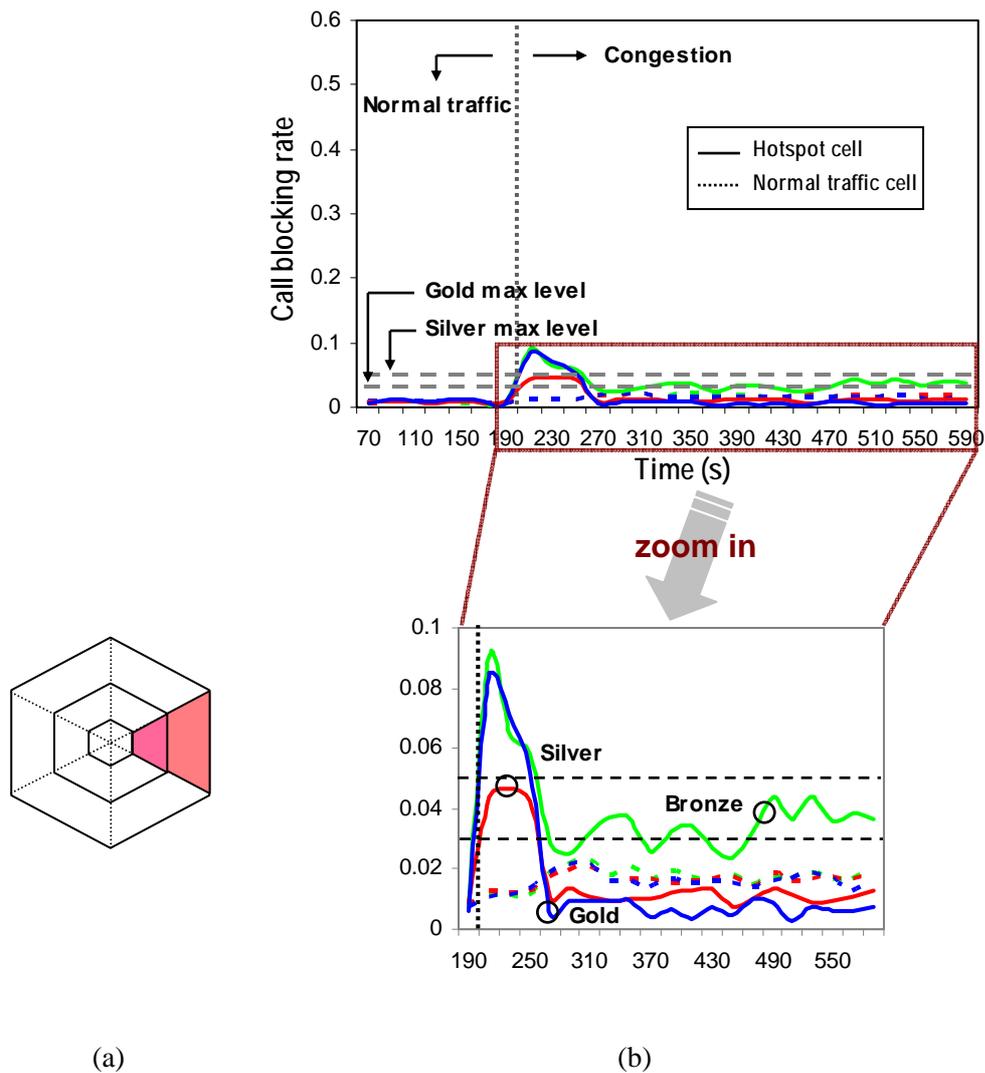
**Figure 6.30 (a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – hybrid method**

The second example is shown in Figure 6.31. In this case, the hotspot cell radius is reduced to 400m and the buffering time for gold and silver customers are set as 4s and 6s respectively.



**Figure 6.31 (a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – hybrid method**

The last example of an unfamiliar case for which the CBR model offers this method is shown in Figure 6.32. The same configuration as in the previous case for buffering time is also used in this case, but with a different cell radius of 500m. The graph in Figure 6.32 illustrates the success of using this solution which maintains the call blocking rate for all customer classes within an acceptable range.



**Figure 6.32 (a) Unfamiliar hotspot cell layout (b) Simulation result for the unfamiliar case – hybrid method**

## 6.6 Summary

This chapter provides a sequence of simulation results starting from the initial stage of the investigation on the impact of changing the reactive layer policy to overcome the system congestion environment. Later in the chapter, results from the study of

implementing the CBR as part of the agent system to propose the best solution for each congestion pattern were presented. Results were given for:

- the results from the case generation process, which showed the benefit of using the proposed method in recovering different congestion patterns. The result from the same scenario used to populate the case library was also given to ensure that the system can immediately propose the right policy to the congestion pattern, which it experienced before
- the results from simulation under similar hotspot patterns, which showed the flexibility and learning ability of the system
- the results from simulation under unfamiliar hotspot patterns, which showed that the system was able to learn from experience and adapt the knowledge to give the right policy to recover these congestion patterns.

# Chapter 7 Conclusions

## 7.1 Conclusions

This work has shown that there is value in using Case Based Reasoning to manage the reactive layer policies in an agent-based RRM system. Because the target of the management system can be expressed in any form the system is extremely flexible. The results here have been applied to both instantaneous call blocking and accumulative call blocking, but any measurable key performance indicator could be used.

Moreover, any form of control strategy could be used – the two here (cell shrinking and connection request buffering) were two examples that demonstrated that different schemes could be applied in different types of congestion: because CBR uses matching of previously-solved cases from a library, it automatically selects the best approach (or combination of approaches) for a particular situation.

The results in Chapter 6 also show that unfamiliar cases can also be dealt with and the system as implemented in the simulator can make use of the previous cases to generate a strategy, even though the situation being dealt with does not appear in the case library.

## 7.2 Further Work

Further work could evaluate the performance of CBR in more complex scenarios:

- Wider variety of traffic conditions and mobility
- Faster moving mobiles where handover call dropping may be more of an issue.
- Geographical constraints (for example shadowing of buildings).

Also, since this work focuses on one single agent within the multi-agent system, the study could be continued to a situation where there are different agents to investigate the collaboration of agents to deal with congestion as well as to support the elaboration of case database.

At a higher level, study could be carried out by including the agents in the negotiation plane to investigate how interaction between agents affects the underlying concept.

## Author's Publications

- [CC03a] S. Chantaraskul and L.G. Cuthbert, "Intelligent, Dynamic Resource Management of 3G Networks", in Proc. of the 5<sup>th</sup> European Personal Mobile Communications Conference, 2003 (EPMCC 2003), Glasgow, Scotland, April 2003.
- [CC03b] S. Chantaraskul and L.G. Cuthbert, "SLA Control for Congestion Management in 3G Networks", in Proc. of The 3<sup>rd</sup> IASTED International Conference on Wireless and Optical Communications, 2003 (WOC 2003), Banff, Canada, July 2003.
- [CC04a] S. Chantaraskul, L.G. Cuthbert, "Introducing Case-Based Reasoning in SLA Control for Congestion Management in 3G Networks", in Proc. of the IEEE Wireless Communications and Networking Conference, 2004 (WCNC 2004), Atlanta GA, USA, March 2004.
- [CC04b] S. Chantaraskul, L.G. Cuthbert, "Using Case-Based Reasoning in Traffic Pattern Recognition for Best Resource Management in 3G Networks", in Proc. of the 7<sup>th</sup> ACM/IEEE International Symposium on Modelling, Analysis and Simulation of Wireless and Mobile Systems Conference, 2004 (ACM/IEEE MSWiM 2004), Venice, Italy, October 2004.
- [CC05a] S. Chantaraskul, L.G. Cuthbert, "Congestion Pattern Matching in Case-Based Reasoning Control for 3G Networks" accepted for presentation in the IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, 2005 (WiMob'2005), Montreal, Canada, August 2005.

## References

- [3GPP01] 3GPP Technical Specification, “UTRAN Overall Description”, 3GPP TS 25.401, version 5.1.0, 2001.
- [3GPP03] 3GPP, Technical Specification Group Radio Access Network, “Physical Layer Procedures (FDD)”, 3GPP TS 25.214, version 4.6.0, 2003.
- [3GPP05] 3GPP, Technical Specification Group Radio Access Network, “Radio Resource Management Strategies (Release 6)”, 3GPP TS 25.922, version 6.1.0, May 2005.
- [3GPP99a] 3GPP, Technical Specification Group Services and System Aspects, “General UMTS Architecture (Release 99)”, 3GPP TR 23.101, version 3.0.0, 1999.
- [3GPP99b] 3GPP, Technical Specification Group Services and System Aspects, “QoS Concept”, 3GPP TR 23.907, version 1.3.0, 1999.
- [Alcatel03] Alcatel, “Service Level Agreement Solutions for Backbone Carriers”, Technical paper in the backbone network and business solutions series, 2003.
- [Ahmed05] M.H. Ahmed, “Call Admission Control in Wireless Networks: A Comprehensive Survey”, IEEE Communication Surveys & Tutorials, vol. 7, no. 1, 2005, pp.50-69.
- [ABGC03] P. Agin, B. Bouffaut, D. Grillo, and G. Colorni, “Soft Handover Performance for UMTS Operations”, In Proc. of the Fifth European Personal Mobile Communications Conference, 2003 (Conf. Publ. No. 492), April 2003, pp. 597-602.
- [AKKC02] E.S. Angelou, N.Th. Koutsokeras, A.G. Kanatas and Ph. Constantinou, “SIR-Based Uplink Terrestrial Call Admission Control Scheme with Handoff for Mixed Traffic W-CDMA Networks”, in Proc. of the 4<sup>th</sup> International Workshop on Mobile and Wireless Communications Network, 2002, pp. 83-87.

- [AP94] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches", *AI Communications*. IOS Press, the European Journal of Artificial Intelligence, Vol. 7:1, 1994, pp. 39-59.
- [Bod00] E.L. Bodanese, "A Distributed Channel Allocation Scheme for Cellular Networks using Intelligent Software Agents", London, UK: University of London PhD Thesis, 2000.
- [BC00a] E.L. Bodanese and L.G. Cuthbert, "Application of Intelligent Agents in Channel Allocation Strategies for Mobile Networks", In *Proc. of the ICC 2000*, New Orleans, LA, 2000, pp.181-185.
- [BC00b] E.L. Bodanese and L.G. Cuthbert, "A Multi-Agent Channel Allocation Scheme for Cellular Mobile Networks", In *Proc. of the ICMAS 2000*, Boston, MA, 2000, pp. 63-70.
- [BHC00] N. Binucci, K. Hiltunen, and M. Caselli, "Soft Handover Gain in WCDMA", in *Proc. of the IEEE 52<sup>nd</sup> Vehicular Technology Conference, 2000 (IEEE VTC-Fall 2000)*, vol. 3, Sept. 2000, pp. 1467-1472.
- [BIVK00] R.D. Bernardi, D. Imbeni, L. Vignali, and M. Karlsson, "Load Control Strategies for Mixed Services in WCDMA", In *Proc. of the IEEE 51<sup>st</sup> Vehicular Technology Conference (IEEE VTC 2000-Spring)*, vol. 2, May 2000, pp. 825-829.
- [BM00] M.P.J. Baker and T.J. Mouldsley, "Power Control in UMTS Release'99", *3G Mobile Telecommunication Technologies*, Conference Publication No.471, IEE 2000.
- [Chen03] Y. Chen "Soft Handover Issues in Radio Resource Management for 3G WCDMA Networks" PhD Thesis Nov 2003, University of London.
- [CC97] C.-C. Chao and W. Chen, "Connection Admission Control for Mobile Multiple-Class Personal Communications Networks", *Selected Areas in Communications*, IEEE Journal on, vol. 15, issue 8, Oct. 1997, pp. 1618-1626.

- [CH95] P. Caulier and B. Houriez, "A Case-Based Reasoning Approach in Network Traffic Control", In Proc. of the IEEE International Conference on Systems, Man and Cybernetics, 1995. 'Intelligent Systems for the 21st Century', vol. 2, October 1995, pp. 1430-1435.
- [CKCN00] C. Jihyuk, K. Taekyoung, Y. Choi, and M. Naghshineh, "Call Admission Control for Multimedia Services in Mobile Cellular Networks: A Markov Decision Approach", in Proc. of the Fifth IEEE Symposium on Computers and Communications, 2000 (ISCC 2000), July 2000, pp. 594-599.
- [CR98] M.A. Centeno and M.F. Reyes, "So you have your model: What to do next, A tutorial on simulation output analysis" In Proc. of the 1998 Winter Simulation Conference, 1998, pp.23-29.
- [CR01] A. Capone and S. Redana, "Call Admission Control Techniques for UMTS", in Proc. of the IEEE Vehicular Technology Conference, 2001, pp.925-929.
- [CRTBB01] L.G. Cuthbert, D. Ryan, L. Tokarchuk, J. Bigam and E. Bodanese, "Using intelligent agents to manage resource in 3G Networks", Journal of IBTE, 2001, vol. 2 part 4.
- [CV01] N. Chandran and M.C. Valenti, "Three generations of cellular wireless systems" IEEE Potentials, Volume: 21 Issue: 1, Feb.-March. 2001, pp. 32-35.
- [DJM96] Z. Dziong, M. Jia, and P. Mermelstein, "Adaptive Traffic Admission for Integrated Services in CDMA Wireless-Access Networks", Selected Areas in Communications, IEEE Journal on, vol. 14, issue 9, Dec. 1996, pp. 1737-1747.
- [ES00] B.M. Epstein and M. Schwartz, "Predictive QoS-Based Admission Control for Multiclass Traffic in Cellular Wireless Networks", IEEE Journal in Selected Areas in Communications, vol. 18, issue 3, March 2000, pp. 523-534.
- [Far96] S. Faruque, "Cellular Mobile Systems Engineering", Mobile Communications Series, Artech House Publishers, 1996.

- [Fer92] I.A. Ferguson, "TouringMachines: An Architecture for Dynamic, Rational, Mobile Agents", Ph.D. thesis, Clare Hall, University of Cambridge, England, 1992.
- [Fer95] I.A. Ferguson, "Integrated Control and Coordinated Behaviour: A Case for Agent Models", Intelligent Agents: Theories, Architectures, and Languages, M. Wooldridge and N. Jennings, (eds.), ECAI-94 Workshop on Agent Theories, Architectures and Languages, vol. 890 of LNCS New York: Springer, 1995, pp. 203-218.
- [Fer99] J. Ferber, "Multi-Agent Systems an Introduction to Distributed Artificial Intelligence", Addison-Wesley, 1999.
- [FSW03] I. Forkel, M. Schinnenburg, and B. Wouters, "Performance Evaluation of Soft Handover in a Realistic UMTS Network", In Proc. of the 57<sup>th</sup> IEEE Semiannual Vehicular Technology Conference, 2003 (VTC 2003-Spring), vol.3, April 2203, pp. 1979-1983.
- [GSM04] GSM MOU Association statistics: from [www.gsmworld.com](http://www.gsmworld.com)
- [GSM05] GSM Association 2005, GSM Technology: from <http://www.gsmworld.com/technology/spectrum/frequencies.shtml>
- [Heine98] G. Heine, "GSM Networks: Protocols, Terminology, and Implementation", Mobile Communications Series, Artech House Publishers, 1998.
- [HA1-MA1-Z01] H. Hassanein, A. Al-Monayyes and M. Al-Zubi, "Improving Call Admission Control in ATM Networks Using Case-Based Reasoning", In proc. of the IEEE International Conference on Performance, Computing, and Communications, 2001, pp. 120-127.
- [HNPR01] S. Hamiti, E. Nikula, J. Parantainen, T. Tantalainen, B. Sébire, and G. Sébire, "GSM/EDGE Radio Access Network (GERAN)-Evolution of GSM/EDGE towards 3G Mobile Services", In proc. of the IEEE International Conference on Telecommunications, 2001.

- [HRM03] T. Halonen, J. Romero, and J. Melero, "GSM, GPRS and EDGE Performance: Evolution towards 3G/UMTS", 2<sup>nd</sup> Edition, Wiley, 2003.
- [HT02] H. Holma and A. Toskala, "WCDMA for UMTS: Radio Access for Third Generation Mobile Communication", Wiley & Sons, Ltd., 2002.
- [HY96] C.Y. Huang and R.D. Yates, "Call Admission in Power Controlled CDMA Systems", In Proc. of the IEEE Vehicular Technology Conference, 1996, pp.227-231.
- [IU97] Y. Ishikawa and N. Umeda, "Capacity design and performance of call admission control in cellular CDMA systems" IEEE Journal on Selected Areas in Communications, vol. 15, issue 8, Oct. 1997, pp. 1627 – 1635.
- [JK99] R. Jain and E.W. Knightly, "A Framework for Design and Evaluation of Admission Control Algorithms in Multi-Service Mobile Networks", in Proc. of the 18<sup>th</sup> Annual Joint Conference of the IEEE Computer and Communications Societies, (INFOCOM '99), vol. 3, March 1999, pp. 1027-1035.
- [Kelton97] W.D. Kelton, "Statistical Analysis of Simulation Output" In Proc. of the 1997 Winter Simulation Conference, 1997, pp.23-30.
- [Kol93] J. Kolodner, "Case-Based Reasoning", Morgan Kaufmann Publishers, Inc., 1993.
- [KALNN01] H. Kaaranen, A. Ahtiainen, L. Laitinen, S. Naghian, V. Niemi, "UMTS Networks Architecture, Mobility and Services", Wiley & Sons, Ltd., 2001.
- [KK04] S.A. Kyriazakos and G.T. Karetzos, "Practical Radio Resource Management in Wireless Systems", Artech House, Inc. 2004.
- [KM99] J. Kuri and P. Mermelstein, "Call Admission on the Uplink of a CDMA System based on Total Received Power", in Proc. of the IEEE International Conference on Communications, 1999, vol. 3, pp. 1431-1436.

- [KSL00] Il-M. Kim, B.-C. Shin, and D.-J. Lee, "SIR-Based Call Admission Control by Intercell Interference Prediction for DS-CDMA Systems" IEEE Communications Letters, vol. 4, issue 1, January 2000, pp. 29-31.
- [LI-E02] R. Lloyd-Evans, "QoS in Integrated 3G Networks", Mobile Communications Series" Artech House, 2002.
- [LAd'I04] M. Luck, R. Ashri and M. d'Inverno, "Agent-Based Software Development", Artech House, Inc., 2004.
- [LAN97] D. Levine, I. Akyildiz, and Am Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept", IEEE/ACM Transactions on Networking, vol. 5, issue 1, February 1997, pp.1-12.
- [LBCH99] Z. Luo, J. Bigham, L.G. Cuthbert and A.L. Hayzelden, "Traffic Control and Resource Management Using a Multi-Agent System", IFIP Broadband Communications'99 Conference, Hong Kong and in "Broadband Communications: Convergence of Network Technologies", Tsang, D. H. K. and Kuhn, P. J. (eds). Kluwer Academic Publishers, USA 11/99, pp. 597-606.
- [LC01] Y-B. Lin, I. Chlamtac, "Wireless and Mobile Network Architecture", Wiley, 2001.
- [LK91] A.M. Law and W.D. Kelton, "Simulation Modelling & Analysis", 2nd Edition, McGraw-Hill International Series, 1991
- [LR99] L. Lewis and P. Ray, "Service Level Agreement Definition, Architecture, and Research Challenges", In Proc. of the Global Telecommunications Conference, 1999, vol. 3, pp. 1974-1978.
- [LWN02] J. Laiho, A. Wacker, T. Novosad, "Radio Network Planning and Optimisation for UMTS", John Wiley & Sons, Ltd., 2002.
- [LZ94] Z. Liu and M.E. Zarki, "SIR Based Call Admission Control for DS-CDMA Cellular System", IEEE Journal on Selected Areas in Communications, 12(4), 1994, pp. 638-644.

- [Mobile03] Mobile Pipeline News on 03<sup>rd</sup> January 2005, [www.mobilepipeline.com](http://www.mobilepipeline.com)
- [MB01] J. Muckenheim and U. Bernhard, "A Framework for Load Control in 3<sup>rd</sup> Generation CDMA Networks", in Proc of the IEEE Global Telecommunications Conference, 2001 (GLOBECOM '01), vol. 6, Nov. 2001, pp. 3738-3742.
- [MBP02] J. Mueckenheim, U. Bernhard, and H. Pampel, "Application of Load Control in 3G CDMA Networks for Improved System Level Modelling and Performance Analysis", in Proc. of the Third International Conference on 3G Mobile Communication Technologies, 2002 (Conf. Publ. No. 489), May 2002, pp. 297-301.
- [MKSAR88] B. Maglaris, D. Anastassiou, P. Sen, G. Karlson, and J.D. Robbins, "Performance models of statistical multiplexing in packet video communications", IEEE Transaction on Communications, vol. 36, no. 7, pp. 834-844, July 1988.
- [MMPG02] E. Marilly, O. Martinot, H. Papini and D. Goderis, "Service Level Agreements: A Main Challenge for Next Generation Networks", In Proc. of the European Conference on Universal Multiservice Network, 2002, pp. 297-304.
- [MP92] M. Mouly and M.-B. Pautet, "The GSM System for Mobile Communications", Published by authors, 1992.
- [MP94] J.P. Miller and M. Pischel, "Modelling Interacting Agents in Dynamic Environments", In Proc. of the 11<sup>th</sup> European Conference on Artificial Intelligence, 1994, pp. 709-713.
- [MPT95] J.P. Miller, M. Pischel, and M. Thiel, "Modelling Reactive Behaviour in Vertically Layered Agent Architectures", Intelligent Agents: Theories, Architectures, and Languages, Lecture Notes in Artificial Intelligence 890, M. Wooldridge and N. Jennings, (eds.), New York: Springer, 1995, pp.261-276.
- [Nakayama03] M.K. Nakayama, "Analysis of Simulation Output", In Proc. of the 2003 Winter Simulation Conference, 2003, pp.49-58.

- [Nwa96] H.S. Nwana, "Software Agents: An Overview", Knowledge Engineering Review, vol.11, No. 3, 1996, pp. 205-244.
- [NTT00] S. Nourizadeh, P. Taaghool, and R. Tafazolli, "A Novel Close Loop Power Control for UMTS", In Proc. of the First International Conference on 3G Mobile Communication Technologies, 2000 (Conf. Publ. No. 471), March 2000, pp. 56-59.
- [OfTel04] UK market – Q4 2002/3 from OfTel market reports: from [www.oftel.gov.uk](http://www.oftel.gov.uk)
- [Parry02] R. Parry, "Overlooking 3G", IEEE Potentials, Volume: 21 Issue: 4, Oct.-Nov. 2002, pp. 6 –9.
- [PS01] J.M. Pitts and J.A. Schormans, "Introduction to IP and ATM Design and Performance" John Wiley & Sons, 2001.
- [Rap96] T.S. Rappaport, "Wireless Communications: Principles & Practice", Prentice Hall, Inc., 1996.
- [RKVF01] W. Rave, T. Kohler, J. Voigt, and G. Fettweis, "Evaluation of Load Control Strategies in an UTRA/FDD Network", In Proc. of the IEEE 53<sup>rd</sup> Vehicular Technology Conference, (IEEE VTC 2001-Spring), vol. 4, May 2001, pp. 2710-2714.
- [RNT96] R. Ramjee, R. Nagarajan, and D. Towsley, "On Optional Call Admission Control in Cellular Networks", in Proc. of the 15<sup>th</sup> Annual Joint Conference of the IEEE Computer Societies, Networking the Next Generation (INFOCOM '96), vol. 1, March 1996, pp. 43-50.
- [RSA05] J. Pérez-Romero, O. Sallent, and R. Agustí, "Radio Resource Management Strategies in UMTS" John Wiley & Sons, Ltd., 2005.
- [SB99] R.S. Sutton and A.G. Barto, "Reinforcement Learning: An Introduction", A Bradford Book, The MIT Press, 1999.
- [SBP03] A.H. Solana, A.V. Bardají, and F.C. Palacio, "Uplink Call Admission Control Techniques for Multimedia Packet Transmission in UMTS WCDMA System", in Proc. of the 5<sup>th</sup> European Personal

- Mobile Communications Conference, 2003 (EPMCC 2003), Glasgow, Scotland, April 2003.
- [SLG01] R. Steele, C-C. Lee, and P. Gould, “GSM, cdmaOnce and 3G Systems”, John Wiley & Sons, 2001.
- [SC02] J-W. So and D-H. Cho, “Access Control of Data in Integrated Voice/Data/Video CDMA Systems”, In Proc. of the IEEE 5<sup>th</sup> Vehicular Technology Conference, Spring 2002, vol. 3, pp.1512-1516.
- [SGSS99] M.L. Sim, E. Gunawan, B.-H. Soong, and C.-B. Soh, “Performance Study of Closs Loop Power Control Algorithms for a Cellular CDMA System”, Vehicular Technology, IEEE Transactions on , vol. 48 , issue 3 , May 1999, pp. 911-921.
- [SPRAC03] O. Sallent, J. Perez-Romero, R. Agusti, and F. Casadevall, “Provisioning Multimedia Wireless Networks for Better QoS: RRM Strategies for 3G W-CDMA”, Communication Magazine, IEEE, vol. 41, issue 2, Feb. 2003, pp. 100-106.
- [Tri01] N.D. Tripathi, “Simulation Based Analysis of the Radio Interface Performance of IS-2000 Systems for Various Data Services”, In Proc. of the IEEE Vehicular Technology Conference, Fall 2001, vol. 4, pp.2665-2669.
- [TB02] W.S. Thong and J. Bigham, “Hierarchical management of CDMA network resources”, In Proc. of the Third International Conference on 3G Mobile Communication Technologies, 2002 (Conf. Publ. No. 489), May 2002, pp. 216 – 220.
- [TNCLD02] S.N. Tehranj, K. Najarian, C. Curescu, T. Lingvall, and T.A. Dahlberg, “Adaptive Load Control Algorithms for 3<sup>rd</sup> Generation Mobile Networks”, in Proc. of the Fifth International Workshop on Modelling Analysis and Simulation of Wireless and Mobile Systems, 2002, pp. 104-111.
- [UMTS04] “3G Frequencies”, from [www.umtsworld.com](http://www.umtsworld.com)

- [VL02] S. Valaee and B. Li, "Distributed Call Admission Control for Ad-Hoc Networks", in Proc. of the 56<sup>th</sup> Vehicular Technology Conference, 2002 (IEEE VTC 2002-Fall), vol. 2, Sept. 2002, pp. 1244-1248.
- [Weiss99] G. Weiss, "Multiagent System: A Modern Approach to Distributed Artificial Intelligence", The MIT Press, Cambridge, Massachusetts, London, England, 1999.
- [Wat97] I.D. Watson, "Applying Case-Based Reasoning: Techniques for Enterprise Systems", Morgan Kaufmann Publishers, Inc., 1997.
- [Wool02] M.J. Wooldridge, "An introduction to Multiagent Systems", John Wiley & Sons, 2002.
- [WJ95] M.J. Wooldridge and N.R. Jennings, "Agent Theories, Architectures, and Languages: A Survey" in Wooldridge and Jennings Eds., Intelligent Agents, vol. 890 of LNCS, Springer, 1995.
- [YGM00] L. Nuaymi, P. Godlewski, and C. Mihailescu, "Call Admission Control Algorithm for Cellular CDMA Systems Based on Best Achievable Performance", in Proc. of the 51<sup>st</sup> IEEE Vehicular Technology Conference, 2000 (IEEE VTC 2000-Spring), vol. 1, May 2000, pp. 375-379.
- [YHP02] J. Ye, J. Hou and S. Papavassiliou, "A comprehensive resource management framework for next generation wireless networks"; Mobile Computing, IEEE Transactions on , vol. 1 , issue 4 , Oct.-Dec. 2002, pp. 249 – 264.
- [ZM04] L. Zhao and J.W. Mark, "Multistep Closed-Loop Power Control Using Linear Receivers for DS-CDMA Systems", Wireless Communications, IEEE Transactions on, vol. 3, issue 6, Nov. 2004. pp. 2141-2155.
- [ZSM00] D. Zhao, X. Shen, and J.W. Mark, "Call Admission Control for Heterogeneous Services in Wireless Networks", in Proc. of the IEEE International Conference on Communications, 2000 (ICC 2000), vol. 2, June 2000, pp. 964-968.