RAT Selection Algorithms for Common Radio Resource Management

---

A Thesis presented to

the Faculty of Engineering and IT

at the University of Technology, Sydney

In accordance with

the requirements for the Degree of

Doctor of Philosophy

---

by

LEIJIA WU

Supervisors: A/Prof. Kumbesan Sandrasegaran, Mr. Anthony Kadi, and Dr. Maged Elkashlan

July 2011

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged with the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate: LEIJIA WU

Production Note:
Signature removed prior to publication.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to A/Prof. Kumbesan Sandrasegaran, who has directed and mentored my research work greatly. A/Prof. Sandrasegaran helped me in many ways that supported, encouraged and motivated me during the pursuit of my PhD degree. His endless efforts made sure that both the flexibility of choosing research topics and the correct research direction.

I would like to give my special thanks to the Commonwealth Scientific and Industrial Research Organization (CSIRO) Information and Communication Technologies (ICT) center, who partially sponsored my research work.

I would like to give my special thanks to my thesis co-supervisors Mr. Anthony Kadi and Dr. Maged Elkashlan, who guided and helped me on my research work.

I would like to thank the University of Technology, Sydney and the Faculty of Engineering and IT for their support through the APA scholarship and sponsoring my conference trips.

I would like to extend my appreciation to my fellow officemates: Mo Li, Rachod Patachaianand, Xiaoan Huang, Prashant Gami, Huda Adibah Mohd Raml, Minjie Xue, Chris Lin, and Lu Chen.

Finally, I would like to give grateful thank to my parents, Jiaming and Guopei, who supported and encouraged me to achieve as high education as they can.

Contents

CHAPTER

List of Tables

List of Figures

ABBREVIATIONS

1G	First Generation
2G	Second Generation
3G	Third Generation
3GPP	3rd Generation Partnership Project
4G	Fourth Generation
AC	Admission Control
APC	Access Point Controllers
ATLB	Adaptive Threshold Load Balancing
BLER	BLock Error Rate
BLJRRME	Base Layer Joint Radio Resource Management Entity
BS	Base Station
BSC	Base Station Controller
CA	Collision Avoidance
CBR	Constant Bit Rate
CC	Congestion Control
CN	Core Network
CRRM	Common Radio Resource Management
CSMA	Carrier Sense Multiple Access
DR	Direct Retry
FDMA	Frequency Division Multiple Access
FSD	Fuzzy Selected Decision
GERAN	GSM/EDGE Radio Access Network
GPRS	General Packet Radio Service

GSM	Global System for Mobile Communication
HC	Handover Control
HO	Handover
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSUPA	High Speed Uplink Packet Access
JRRM	Joint Radio Resource Management
IN	Indoor
LB	Load balancing
LTE	Long Term Evolution
MADM	Multiple Attribute Decision Making
MCDM	Multi-Criteria Decision Making
MCS	Modulation and Coding Scheme
MODM	Multiple Objective Decision Making
MRRM	Multi-access Radio Resource Management
MS	Mobile Station
NCCB	Network Controlled Cell Breathing
NRT	Non-Real Time
OSM	Operator Software Module
PC	Power Control
PS	Packet Scheduling
QoS	Quality of Service
RAT	Radio Access Technology
RNC	Radio Network Controller
RRM	Radio Resource Management
RRME	RAT Resource Management Entity
RRU	Radio Resource Unit

RT	Real Time
SIR	Signal to Interference Ratio
SMD	Semi-Markov Decision
TDMA	Time Division Multiple Access
UE	User Equipment
ULJRRME	Upper Layer Joint Radio Resource Management Entity
UMTS	Universal Mobile Telecommunications System
USaBS	User Satisfaction Based Selection
USaLOR	User Satisfaction with Low Resources Selection
USM	User Software Module
UT	User terminal
UTRAN	Universal Terrestrial Radio Access Network
VG	Voice GERAN
VHO	Vertical Handover
VoIP	Voice over IP
VU	Voice UTRAN
WCDMA	Wideband Code Division Multiple Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network
WWAN	Wireless Wide Area Network

ABSTRACT

The future wireless network is expected to be a heterogeneous network, which integrates different Radio Access Technologies (RATs) through a common platform. A major challenge arising from the heterogeneous network is Radio Resource Management (RRM) strategy. Common RRM (CRRM) has been proposed in the literature to jointly manage radio resources among a number of overlapped RATs in an optimized way. RAT selection algorithm is one of the key research areas in CRRM. In the literature, a number of RAT selection algorithms have been proposed and some performance evaluations have been conducted. However, this area still has many challenges. Some performance metrics still have not been evaluated well and the existing algorithms can be further improved.

In this thesis, some performance evaluations on a number of RAT selection algorithms have been carried out. The effects of load threshold setting on Load Balancing (LB) based RAT selection algorithm's performance are evaluated. It is found that setting a proper load threshold can achieve a more balanced load distribution among overlapped cells. However, it will also cause higher Direct Retry (DR)/Vertical Handover (VHO) probability and in turn higher overhead and blocking/dropping probability.

This thesis evaluates the performance of three RAT selection algorithms, LB based using maximum resource consumption, LB based using minimum resource consumption, and service based algorithms, in terms of traffic distribution, blocking probability, throughput, and throughput fairness for a co-located GERAN/UTRAN/WLAN network. Simulation results show that in terms of blocking probability, the service based algorithm is the worst one when the traffic load is high. In terms of data throughput, the LB based using maximum resource consumption algorithm performs better than the other two when the traffic load is low. However, the service based algorithm outperforms the other two when the traffic load is high. In terms of throughput fairness, the service based algorithm achieves the best performance.

The relationship among overall downlink data throughput, user satisfaction rate, and path loss threshold is studied in this thesis. It is found that in some cases, an optimum path loss threshold value can be found to achieve better performance in terms of both overall throughput and user satisfaction rate. However, in other cases, a tradeoff has to be made between them.

This thesis studies policy based RAT selection algorithms for a co-located UMTS-/GSM network. A three-complex policy based algorithm called IN*VG*Load algorithm is proposed based on improvements on the existing IN*VG algorithm. The simulation results show that the IN*VG*Load algorithm can optimize the system performance in highly loaded co-located UMTS/GSM networks. A Proposed Policy Based Algorithm 2 is found to be suitable for low to medium loaded UMTS/GSM networks.

In order to support the conceptual development of RAT selection algorithms in heterogeneous networks, the theory of Markov model is used. This thesis proposes both user level and network level Markov models for a co-located GERAN/UTRAN/WLAN network. The proposed Markov models are not only extensions of the existing two co-located RATs models but more complex with more state transitions. The performance of two basic RAT selection algorithms: LB based and service based algorithms are evaluated in terms of call blocking probability. The numerical results obtained from the proposed network level Markov model are validated by simulation results.

Chapter 1

Introduction

1.1 Evolution of wireless networks

Wireless networks have become an important part of our everyday life. People enjoy the great convenience of wireless communications, for both personal and business purposes. Due to the explosive growth in the usage of wireless communications, radio spectrum has become a scarce and expensive commodity. Network operators need to obtain a license before transmitting on a licensed frequency band. In order to establish compatibility and inter-operability between different networks and network operators, standards are developed to specify the information transferred on all interfaces.

Each user in a wireless network has to be allocated an appropriate Radio Resource Unit (RRU) for communication in the uplink (user to network) and downlink (network to user) direction. A RRU may have many dimensions such as frequency, time, code, and power dependent on the wireless technology being used. The amount of RRUs allocated to a user may vary with time and the type of service currently being used. Higher data rate services, such as video streaming, will require more RRUs compared to lower data rate services such as voice. The method of allocation of RRUs is referred to as multiple access technique.

A number of Radio Access Technologies (RATs) have been developed over the last 30 years. RATs can be classified by generations (1G, 2G,...4G), multiple access technology, coverage, etc. In terms of coverage, wireless networks can be classified into Wireless Personal Area Network (WPAN), Wireless Local Area Network (WLAN), Wireless Metropolitan Area Network (WMAN), and Wireless Wide Area Network (WWAN).

First Generation (1G) mobile networks are based on analogue technology and offered speech services only. The multiple access technique used in 1G mobile networks is Frequency Division Multiple Access (FDMA). The RRU allocated to each user connecting to a 1G wireless network is a fixed narrow frequency band for the entire call duration. This is not an efficient method for usage of available spectrum.

Second Generation (2G) mobile networks use circuit switching and digital transmission technologies, which allowed the use of more efficient multiple access techniques, such as Time Division Multiple Access (TDMA). A good example of a 2G mobile network is the Global System for Mobile Communication (GSM) system which has been the most successful mobile communication system implemented to date. In GSM, the available frequency band is divided into several sub-channels and the RRU allocated to each user is a timeslot on a sub-channel.

2.5G mobile networks, e.g. General Packet Radio Service (GPRS), were designed to offer packet switching services only with minimal changes to the radio interface of GSM networks. The RRU allocated to a user in GPRS is a radio block which refers to partial usage of a timeslot.

Third Generation (3G) mobile networks improved the bottleneck of the radio interface in 2G mobile networks and are able to offer circuit and packet switching technologies. Universal Mobile Telecommunication System (UMTS), relying on Wideband Code Division Multiple Access (WCDMA) techniques, is one of the most successful 3G technologies. It was standardised by the 3rd Generation Partnership Project (3GPP) to provide high data rate applications, which 2G technologies (i.e. GSM) could not support. At present, UMTS users can obtain data rates up to 384 kbps, which is much greater than 14.4 kbps provided by the earlier GSM technology. A new radio access network called UMTS Terrestrial Radio Access Network (UTRAN) was deployed by network operators and the UTRAN was connected to the core networks inherited from GSM and GPRS. There has been a significant growth in the number of 3G/UMTS subscribers worldwide. Many network operators continue to operate GSM, GPRS, and UMTS networks today. In a UMTS network, all users communicate on the same 5 MHz bandwidth and at the same time. A RRU is defined by a carrier frequency, a code sequence, and a power level.

The maximum data rates in 3G networks have been enhanced to 14 Mbps in downlink with the use of High Speed Downlink Packet Access (HSDPA) and 5 Mbps in the uplink with the use of High Speed Uplink Packet Access (HSUPA). A good example is the Telstra's billion dollar Next G High Speed Packet Access (HSPA) network in Australia, which was launched in October 2006.

Wireless Local Area Networks (WLAN), such as IEEE 802.11, are now an effective means of public wireless access using the 2.4 GHz and 5 GHz unlicensed bands. The IEEE 802.11 standard, also known as Wi-Fi, can provide high speed data services with a link rate up to 54Mbps within a 200m radio range [1]. The small radio coverage of WLAN technologies is due to the limitations of transmitting power on unlicensed frequency bands and the use of Carrier Sense Multiple Access (CSMA) with Collision Avoidance (CA). More information about the technical details of the above mentioned wireless networks can be found in [2, 3].

Today, a number of RATs coexist and overlap in the same geographical area. For example, users in a building may be within the coverage area of GSM, UMTS, HSPA, and WLAN at the same time. Furthermore, wireless terminals that can communicate with multiple RATs have become available today. At present, each RAT operates independently as a homogenous network. The future Fourth Generation (4G) network is expected to be a heterogeneous wireless network that integrates a number of RATs, e.g. GSM/EDGE Radio Access Network (GERAN), UTRAN, and WLAN through a common platform.

A challenge arising in the heterogeneous network is how to allocate a particular user to the most suitable wireless network. An effective solution for this problem can bring significant benefits to both end users and service providers, such as efficient radio resource utilization, better system performance, better Quality of Service (QoS), overall network stability, enhanced user satisfaction, and increased operator's revenue.

1.2 Common radio resource management

Radio Resource Management (RRM) refers to a group of mechanisms that are collectively responsible for efficiently utilizing RRUs within a RAT to provide services with an acceptable level of QoS. RRM mechanisms contain Power Control (PC), Handover Control (HC), Packet Scheduling (PS), Congestion Control (CC), and Admission Control (AC).

At present, Radio Resource Management (RRM) strategies are implemented independently in each RAT. None of the RRM strategies is suitable for the heterogeneous network, because each RRM strategy only considers the situation of one particular RAT. The Common RRM (CRRM) strategy, also known as Multi-access RRM (MRRM) or Joint RRM (JRRM), has been proposed in the literature to coordinate

RRU utilization among a number of RATs in an optimized way. One of the earliest work in CRRM [4] shows that networks using CRRM outperform those without CRRM for both real time (RT) and non-real time (NRT) services in terms of call blocking probability and capacity gain.

1.2.1 CRRM operation

The CRRM concept is based on a two-tier RRM model [5], consisting of CRRM and RRM entities as shown in Fig. 1.1. The RRM entity is located at the lower tier and manages RRUs within a RAT. The CRRM entity is at the upper tier of the two-tier RRM model. It controls a number of RRM entities and can communicate with other CRRM entities. Based on the information gathered from its controlling RRM entities, the CRRM entity is able to know the RRU availability of multiple RATs and allocate a user to the most suitable RAT.

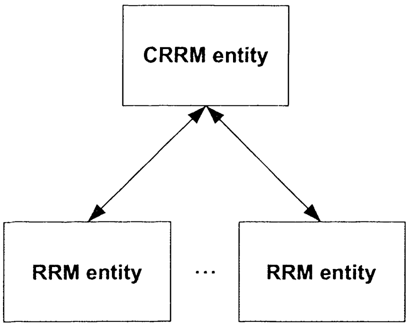


Figure 1.1: Two-tier RRM model

The interactions between RRM and CRRM entities support two basic functions. The first function is referred to as the information reporting function, which allows RRM entities to report relevant information to their controlling CRRM entity. The information reporting can be performed either periodically or be triggered by an event. The reported information contains static cell information (cell relations, capabilities, capacities, QoS, maximum bit rate for a given service, and average buffer delay, etc.) and dynamic cell information (cell load, received power level, transmit power level, and interference measurements, etc.) [6]. The information reporting function is also used for information exchange and sharing between different CRRM entities as shown in Fig. 1.2.

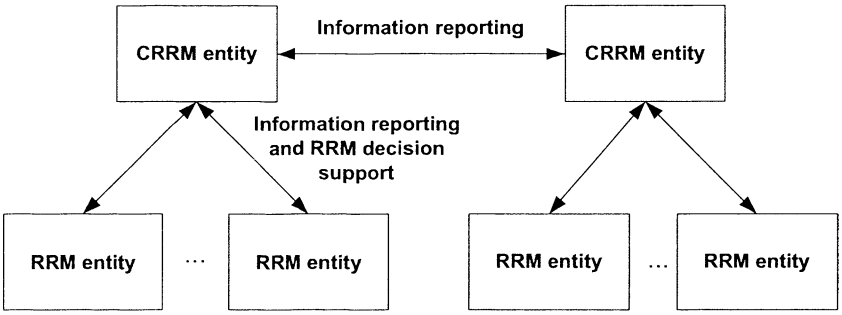


Figure 1.2: CRRM interaction model

The second function is RRM decision support function, which describes the way that RRM and CRRM entities interact with each other to make decisions as shown in Fig. 1.2. There are two RRM decision-making methods. One is CRRM centered decision making, in which the CRRM entity makes decisions and informs RRM entities to execute them. The second is local RRM centered decision-making, where the CRRM entity only advises RRM entities but the final decision is made by the RRM entities rather than the CRRM entity.

A number of interaction degrees exist between CRRM and RRM entities according to the split of functionalities. Pérez-Romero et al. [7] introduced four interaction degrees, which are summarized in Table 1.1. The first column of the table shows the four possible interaction degrees: Low, Intermediate, High, and Very High. Low interaction degree means that the majority of RRM functions are performed in the local RRM entities whereas the Very High interaction degree means that the majority of functions are performed in the CRRM entities. The second column (the interaction time scale) in the table indicates how often the CRRM entities need to communicate with RRM entities. A higher interaction degree between RRM and CRRM entities can achieve a more efficient radio resource management, because more functions are performed at the CRRM level, and the interaction time scale between RRM and CRRM entities is shorter. However, a higher interaction degree means more interaction activities, therefore leads to higher amount of overhead. Currently, most of the research work in CRRM, including the work described in this thesis, is focusing on the intermediate interaction degree level.

Table 1.1: Interaction degrees between RRM/CRRM entities

Interaction degree	Interaction time scale	Functions in CRRM entities	Functions in local RRM entities
Low	Hours/days	Policy translation and configuration	Initial RAT selection, VHO, Admission control, Congestion control, Horizontal handover, Packet scheduling, Power control
Intermediate	Minutes	Policy translation and configuration, Initial RAT selection, VHO	Admission control, Congestion control, Horizontal handover, Packet scheduling, Power control
High	Seconds	Policy translation and configuration, Initial RAT selection, VHO, Admission control, Congestion control, Horizontal handover	Packet scheduling, Power control
Very High	Milliseconds	Policy translation and configuration, Initial RAT selection, VHO, Admission control, Congestion control, Horizontal handover, Packet scheduling	Power control

1.2.2 CRRM topologies

In the previous section, CRRM was introduced from the functional point of view. From the network point of view, the implementation of CRRM has a number of alternatives. RRM entities are usually integrated into Base Station Controllers (BSCs) in GERAN, Radio Network Controllers (RNCs) in UTRAN, and Access Point Controllers (APCs) in WLAN. The CRRM entity can be implemented in a number of ways.

CRRM server topology

In [8, 9], a CRRM server topology as shown in Fig. 1.3 is proposed. A new logical node referred to as the CRRM server is added in the Core Network (CN). It contains all CRRM functions and is connected with a number of RRM entities. The CRRM server topology is centralized so that it can achieve high scalability. However, the introduction of a new network element will increase the cost of network implementation. The communication between RRM entities and the CRRM server introduces additional signalling delays.

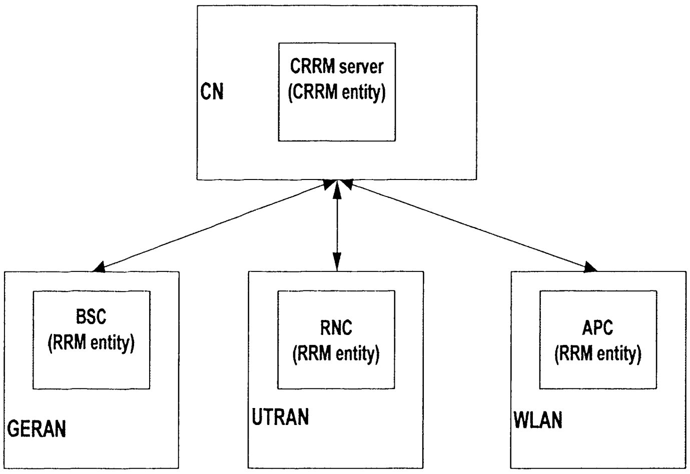


Figure 1.3: CRRM server approach network topology

Integrated CRRM topology

In [8, 9, 10], an integrated CRRM topology has been proposed (as shown in Fig. 1.4). Unlike the centralized CRRM server topology, the integrated CRRM topology distributes CRRM functionalities into existing network nodes (BSCs, RNCs, and APCs), which requires minimum infrastructure changes. The execution of CRRM functions can be performed directly between RATs rather than through the CN, so that no additional delay will be incurred. However, the distributed nature of this approach causes a scalability problem. With the increase of the number of RRM entities, the number of connections between the RRM entities will grow exponentially.

In the integrated CRRM topology, CRRM entities may be located either within every BSC, RNC, and APC nodes, or only in some of them [7, 11]. In the first case, the RRM decision support function does not need to be standardized because decision-making processes between CRRM and RRM entities are performed locally in the same physical entity. However, in the latter case, the RRM support function needs to be standardized because some RRM entities are not co-located with the CRRM entity and open interfaces exist between them.

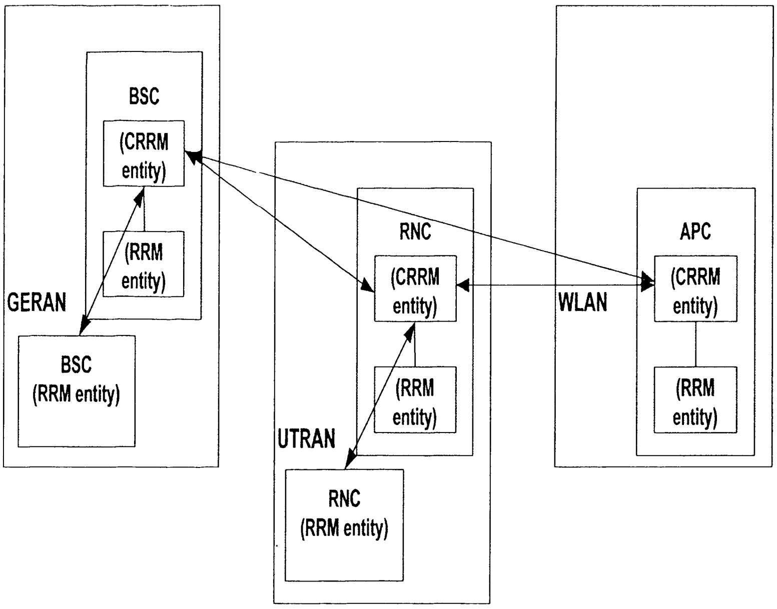


Figure 1.4: Integrated CRRM approach network topology

Hierarchical CRRM topology

In [12], a hierarchical CRRM topology, which is a tradeoff between the centralized and distributed topologies is proposed. As shown in Fig. 1.5, the hierarchical CRRM topology has four layers. The BS is located at the lowest layer, the RAT Resource Management Entity (RRME) manages BSs belonging to the same RAT, the Base Layer Joint Radio Resource Management Entity (BLJRRME) coordinates a number of RRMEs and the Upper Layer Joint Radio Resource Management Entity (ULJRRME) controls a number of BLJRRMEs. When a new call arrives, RRMEs will select available cells for it, and subsequently BLJRRMEs will choose the best RAT under its control, finally, the ULJRRME will allocate the call to the most suitable RAT among the RATs recommended by BLJRRMEs.

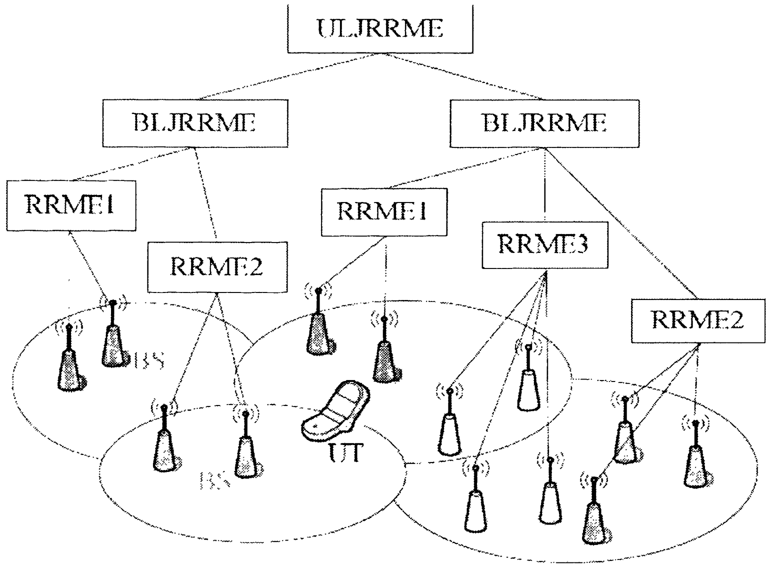


Figure 1.5: Hierarchical CRRM approach network topology [12]

CRRM functions in UT topology

Magnusson et al. [13] proposed a CRRM functions in User Terminal (UT) topology as shown in Fig. 1.6. This topology allows the end user, rather than the network operator to make the RAT selection decisions.

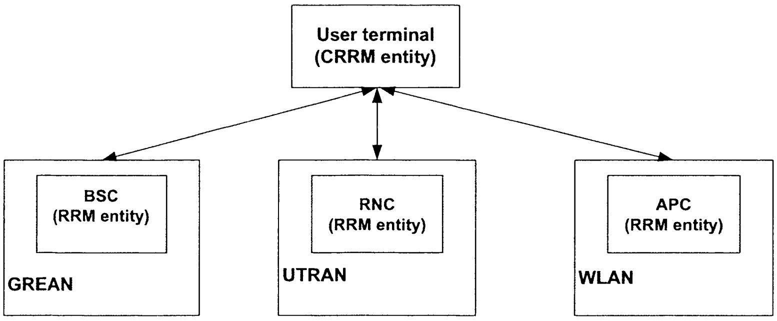


Figure 1.6: CRRM functions in UT topology

All CRRM topologies given above have their pros and cons. The CRRM server topology is best suited for long-term RRM functions, such as overall load balancing. The integrated CRRM approach combined with the CRRM functions in UT works well for dynamic RRM handling, which requires frequent signal exchanges. The hierarchical CRRM topology is a tradeoff between the two. In the work of this thesis, the CRRM server topology is used.

1.3 Problem statement and research questions

1.3.1 Problem statement

A number of RAT selection algorithms have been proposed for the heterogeneous network in the literature, however, the existing algorithms have their disadvantages and can be further improved. The RAT selection algorithms for a two co-located

RAT scenario were studied relatively well in the literature, however, there is not sufficient research work focusing on the RAT selection algorithms for a three RAT scenario. The theory of Markov Chain is used in the RAT selection area to provide an analytical way to analyze the system performance, however, most of the current work only focuses on a two co-located RAT scenario too. Solutions are required to deal with the above issues.

1.3.2 Research questions

The following research questions are defined and solved in this thesis:

Based on the currently known load threshold knowledge, can one evaluate the effects of load threshold setting on the performance of Load Balancing (LB) based RAT selection algorithm for real time traffic?

Based on the currently known RAT selection algorithms, can one evaluate the performance on RAT selection algorithms for a co-located GERAN/UTRAN/WLAN network?

Based on the currently known NCCB algorithm knowledge, can one find the relationship between overall downlink data throughput, user satisfaction rate, and path loss threshold in the NCCB algorithm?

Based on the currently known policy based RAT selection algorithms, is there a more optimal RAT selection algorithm for a co-located UMTS/GSM network?

Based on the currently known Markov models for a two co-located RAT scenario, can one build Markov models for a three co-located RAT scenario?

1.4 Contributions

The major contributions of this thesis are as follows:

1.4.1 Critical surveys

A critical survey of existing RAT selection algorithms for heterogeneous wireless networks is presented. RAT selection algorithms are classified in terms of selection criteria. Advantages and disadvantages of these RAT selection algorithms are analyzed, and different RAT selection algorithms are compared. In this thesis, a critical survey of existing Markov models for RAT selection algorithms is conducted. Both user level and network level Markov models in the literature are reviewed. Strengths and weaknesses of these models are presented.

1.4.2 Performance evaluations

In this thesis, a number of performance evaluations on RAT selection algorithms are carried out. The effects of load threshold setting on LB based RAT selection algorithm performance are evaluated.

This thesis evaluates the performance of three RAT selection algorithms: LB based using maximum resource consumption, LB based using minimum resource consumption, and service based algorithms, for a co-located GERAN/UTRAN/WLAN network in terms of traffic distribution, blocking probability, overall throughput, and throughput fairness.

The setting of a proper path loss threshold is a key issue in the NCCB algorithm. In this thesis, the relationship between overall downlink data throughput, user satisfaction rate, and path loss threshold is studied.

1.4.3 Improved policy based RAT selection algorithms

This thesis studies policy based RAT selection algorithms in co-located UMTS/GSM networks. A three-complex algorithm, IN*VG*Load, is proposed based on improvements on the existing IN*VG algorithm. Simulation results show that the IN*VG*Load algorithm can optimize the system performance in highly loaded co-located UMTS/GSM networks. A Proposed Policy Based Algorithm 2 is found to be suitable for low to medium loaded UMTS/GSM networks.

1.4.4 Markov models

This thesis proposes new user level Markov models for a three co-located RATs network based on an extension from existing two co-located RATs Markov models in the literature. The LB based and service based RAT selection algorithms have been analyzed using the proposed Markov models. By using the proposed user level Markov model, it can be known which RAT a user will be allocated to, given related information, such as the environment, RAT selection algorithm, call type, etc.

This thesis proposes a new network level three-dimensional Markov model for a co-located GERAN/UTRAN/WLAN network. Compared to the existing network level Markov models, the proposed model considers both service differentiation and mobility issues. Numerical results obtained from the proposed Markov model are validated by simulation results. The performance of two basic RAT selection algorithms: LB based and service based algorithms are evaluated in terms of call blocking probability using the proposed model.

1.5 Thesis outline

The rest of this thesis is organized as follows. In Chapter 2, a comprehensive literature review of RAT selection algorithms is presented. Chapter 3 presents simulation models, performance parameters, and RAT selection algorithms that will be simulated. In Chapter 4, simulation results are presented to evaluate the performance of a number of RAT selection algorithms. In Chapter 5, background knowledge of Markov models and existing user level Markov models are presented, and new user level Markov models are proposed. Chapter 6 reviews existing network level Markov models and proposes a new network level Markov model. Chapter 7 concludes this thesis and discusses future work.

1.6 Related publications

The following publications have been produced based on the contributions included in this thesis.

L. Wu and K.Sandrasegaran, "A Survey on Common Radio Resource Management", The Second Australia Conference on Wireless Broadband and Ultra Wideband Communications (Auswireless07), Sydney, Australia, 27-30 Aug. 2007, pp. 66.

L. Wu and K.Sandrasegaran, "A Study on RAT Selection Algorithms in Combined UMTS/GSM Networks", The 7th International Symposium on Communications and Information Technologies (ISCIT 2007), Sydney, Australia, 16-19 Oct. 2007, pp 421-426.

L. Wu and K.Sandrasegaran, "A Study on RAT Selection Algorithms in Combined UMTS/GSM Networks," ECTI Transaction on Electrical Engineering, Electronics and Communications, Vol.6, No.2, pp. 86-92, Aug. 2008.

L. Wu, K. Sandrasegaran, and H. A. M. Ramli, "A Study on Load Threshold Setting Issue in Load Based Common Radio Resource Management," in The 4th International Conference on Information Technology and Multimedia at UNITEN (ICIMU' 2008), Bangi, Malaysia, 18-19 Nov. 2008.

L. Wu, K. Sandrasegaran, and M. Elkashlan, "A System Level Simulation Model for Common Radio Resource Management," in The 15th Asia-Pacific Conference on Communications (APCC 2009), Shanghai, China, 8-10 Oct. 2009, pp. 686-689.

L. Wu, K. Sandrasegaran, and M. Elkashlan, "Tradeoff between Overall Throughput and Throughput Fairness in Network Controlled Cell Breathing Algorithm," in The 15th Asia-Pacific Conference on Communications (APCC 2009), Shanghai, China, 8-10 Oct. 2009, pp. 708-712.

L. Wu, A. Sabbagh, K. Sandrasegaran, and M. Elkashlan, "A User Level Markov Model for Load Balancing Based RAT Selection Algorithm", in The 8th International Information and Telecommunication Technologies Symposium (I2TS 2009), 09-11 Nov. Florianpolis, Brazil, 2009.

L. Wu, A. Sabbagh, K. Sandrasegaran, and M. Elkashlan, "A User Level Markov Model for Service Based CRRM Algorithm", in The International Conference on Multimedia Computing and Information Technology (MCIT-2010), Sharjah, UAE, 2-4 Mar. 2010, pp. 41-44.

L. Wu, A. Sabbagh, K. Sandrasegaran, and M. Elkashlan, "Performance Evaluation on Common Radio Resource Management Algorithms," in The 5th International Workshop on Performance Analysis and Enhancement of Wireless Networks (PAEWN-2010) of the The International Conference on Advanced Information Networking and Applications (AINA 2010), Perth, Australia, 20-23 Apr. 2010, pp. 491-495.

L. Wu, K. Sandrasegaran, and M. Elkashlan, "A Markov Model for Performance Evaluation of CRRM Algorithms in a co-located GERAN/UTRAN/WLAN scenario," in The 5nd International Workshop on Planning and Optimization of Wireless Communication Networks (PlanNet 2010) of the IEEE Wireless Communications & Networking Conference (WCNC 2010), Sydney, Australia, 18-21 Apr. 2010, pp. 1-6.

L. Wu and K. Sandrasegaran "A Survey on RAT Selection Algorithms for Common Radio Resource Management," submitted to the IEEE surveys and tutorials, 2011.

Chapter 2

RAT selection algorithms

Research in CRRM has many directions, e.g. policy translation and configuration, RAT selection, admission control, congestion control, horizontal handover, and packet scheduling. RAT selection algorithm is a key research area of CRRM at present. A suitable RAT selection algorithm can manage radio resources among multiple RATs more efficiently, enhance system performance, and provide better QoS to users. The RAT selection algorithm contains two parts: initial RAT selection and vertical handover (VHO). The former is used to allocate new calls to a suitable RAT and the latter is about transferring an ongoing call from its current serving RAT to a more suitable RAT. A number of RAT selection algorithms have been studied in the literature [14, 15, 16]. These algorithms use one or more RAT selection criteria. These RAT selection criteria are based on user's perspective, operator's perspective or both. From the user's perspective, the serving RAT should meet one or more of the following requirements: low service price, low delay, high data rate, large coverage area, low battery power consumption, and high network security. From the operator's perspective, a preferred RAT selection algorithm should meet one or more of the following requirements: load balancing, high revenue, low call blocking and dropping probabilities, and efficient radio resource utilization. An algorithm using one of the criteria is called a single criterion based algorithm, while an algorithm using two or more criteria is called a multiple criteria based algorithm.

It is possible that users and operators may have different perspectives on the same criterion. For example, for the user's perspective, the service price should be as low as possible while for the operator's perspective, the service price should be high enough to provide a good revenue [17, 18]. A tradeoff is required between user's and operator's preferences. In this chapter, existing RAT selection algorithms using different criteria are discussed.

2.1 Performance parameters

The following parameters are usually used to evaluate the performance of RAT selection algorithms:

1)Load deviation - It refers to a measure of difference for RAT loads between the observed value and the mean. The value of a load deviation is between 0 to 1. A lower value of load deviation means less difference between the observed value and the mean. The load deviation at time interval t, σ_t, is calculated by:

$$\sigma_t = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{it} - \bar{x_t})^2}, \tag{2.1}$$

where N is the number of overlapped cells, x_{it} is the load of cell i at time interval t and $\bar{x}_t$ is the average load of all overlapped cells at time interval t.

2)Blocking probability - It refers to the statistical probability that a new call connection cannot be established due to insufficient resources in the network, usually expressed as a percentage or decimal equivalent of calls blocked by network congestion during the busy hour. A mathematical expression of the blocking probability, P_b, is,

$$P_b = \frac{\frac{E^m}{m!}}{\sum_{i=0}^{m} \frac{E^i}{i!}}, \tag{2.2}$$

where m is the number of resources and E is the total amount of traffic offered in Erlangs.

3)Dropping probability - It refers to the statistical probability that an ongoing call is terminated due to insufficient transmission resources in the network.

4)Throughput - It refers to the average rate of successful message delivery over a communication channel. The throughput is usually measured in bits per second (bit/s or bps). The throughput is calculated by:

$$Throughput = R(1 - BLER), \tag{2.3}$$

where R is the bit rate and $BLER$ is the block error rate.

5)Block Error Rate (BLER) - It is a ratio of the number of erroneous blocks to the total number of blocks received.

)Erlang capacity - It is the amount of offered traffic in Erlangs a system can allow.

6)Fairness index - It refers to a parameter to determine whether users or applications are receiving a fair share of system resources. The well known Jain's fairness index equation is given by [67]:

$$Fairness\ index = \frac{(\sum_{i=0}^{n} F_i)^2}{n \sum F_i^2}, \tag{2.4}$$

where F_i is the throughput for every user and n is the total number of users sharing the resources. The value of the fairness index is between (0,1]. A higher fairness index means users are treated more equally. A fairness index of one means all users are treated equally. In this thesis, the fairness index is used to measure the throughput fairness.

7)User satisfaction rate - It refers to a parameter used to measure the ratio of served users who satisfy the service provided. In this thesis, a user satisfaction rate is defined as,

$$S = n/N, \tag{2.5}$$

where n is the number of users who satisfy the service provided and N is the total number of users served in the network.

2.2 Random selection based algorithm

The random selection based algorithm is the simplest RAT selection algorithm, which can be referred to as a "No CRRM" algorithm. In this algorithm, no CRRM mechanisms are performed. When a new or VHO call arrives, one of the available RATs is randomly selected as the target RAT. The probability of a RAT to be selected P is:

$$P = 1/N, \tag{2.6}$$

where N is the number of available RATs. If the selected RAT does not have sufficient capacity to serve the call, the call will be blocked or re-directed to another randomly selected RAT. This algorithm is simple and easy to implement. However, it will cause high blocking and dropping probabilities and inefficient usage of radio resources.

2.3 Single criterion based algorithms

In this section, a number of single criterion based RAT selection algorithms are discussed.

2.3.1 Load Balancing based algorithms

The concept of load balancing initially appeared in the distributed computing area [19]. In the context of wireless networks, it refers to evenly distributing traffic load among a number of cells or nodes to optimize radio resource utilization, maximize throughput, minimize delay, and avoid overload. Load balancing (LB) based algorithms have been studied to improve the performance of a homogenous network, where the coverage area of a number of Base Stations (BSs) are overlapping [20]. In this case, a new call is directed to the least loaded BS. In a heterogeneous network, under the LB based RAT selection algorithm, a call is always allocated to the least loaded RAT. The probability of the i^{th} RAT to be selected under the LB based algorithm P_i is:

$$P_i = \begin{cases} 1 & L_i = min(L_1, L_2, ...L_N) \ \& \ L_i \leq L_{imax}, \\ 0 & \text{if else}, \end{cases} \tag{2.7}$$

where L_i is the load of the i^{th} RAT, N is the number of available RATs, and L_{imax} is the the maximum allowed load of the i^{th} RAT. A number of LB based algorithms have been studied in the literature. They are discussed in the following sub-sections.

Fixed load threshold algorithms

A fixed load threshold RAT selection algorithm is proposed in [4, 9]. It is assumed that multiple RATs have exactly the same coverage area, network topology, and capacity. Cells sharing the same coverage area but belonging to different RATs are defined as overlapped cells. A predetermined load threshold (e.g. 80% of the maximum cell load) is set for each cell. When a new call arrives at or an ongoing call moves into a cell, if the load of the current cell is below the load threshold, the call will be processed in the current cell. If the load of the current cell is above the load threshold, a target cell will be selected. The target cell is the least loaded overlapped cell known to the CRRM entity. If the load of the target cell is lower than the load threshold, a Direct Retry (DR) for new call or a VHO for ongoing call is triggered and the call will be directed to the target cell (a DR refers to the process of transferring a new call or data session from its current cell to another [21]). If the load of the target cell is above the load threshold too and the load of the current cell is not full, the call will stay in the current cell. If the load of the current cell is full, the call will be directed to the target cell if the target cell has free capacity to serve it, otherwise, the call will be blocked or dropped.

In [4, 9], the fixed load threshold algorithm is compared with a "No CRRM" algorithm, which refers to the random selection algorithm discussed above. The comparison results illustrate that the CRRM algorithm outperforms the "No CRRM" algorithm in terms of blocking probability and user throughput.

In the fixed load threshold algorithm discussed above, the traffic load is continuously balanced. An alternative method is to balance the load at regular intervals of time [22]. This method can reduce the amount of overhead but it is not as efficient as the continuously balanced method due to the reason of using out-of-date information.

In LB based algorithms, the load threshold value should be high enough to reduce unnecessary handovers (HOs). However, it should not be too high, otherwise, the load balancing purpose will not be achieved. If the traffic load in the heterogeneous network is fixed, it is possible to find an optimized load threshold. However, as we know, the wireless environment changes frequently, therefore, adaptive load threshold algorithms have been studied in the literature to provide a better solution.

Adaptive load threshold algorithms

Tolli et al. [23] proposed an adaptive load threshold algorithm. In this algorithm, the load threshold of a cell is adjusted periodically according to the average load of its overlapped cells. The load threshold of a cell should always be higher than the loads of its overlapped cells in order to reduce the number of HO failures, which equals to the number of HO attempts minus the number of load reason HOs [23]. Therefore, the higher the average load in the overlapped cells, the higher the load threshold.

In [23], three important parameters are used in the adaptive load threshold algorithm: tuning step, minimum load threshold, and maximum load threshold. The load threshold of a cell is increased or decreased by one tuning step periodically between the minimum and maximum load thresholds according to the variation of the average load of the overlapped cells. Simulation results in [23] show that the adaptive load threshold algorithm performs better than the fixed load threshold algorithm in terms of reducing HO failures. Challenges in this algorithm is that optimized minimum and maximum threshold values need to be worked out and the ping-pong effect (the threshold keeps going up and down) needs to be alleviated.

An Adaptive Threshold Load Balancing (ATLB) algorithm has been proposed in [24]. This algorithm looks at the load gap, which is the difference between the least and most loaded overlapped cells, rather than the load of each individual cell. In this algorithm, a load gap threshold between the most and least loaded overlapped cells is predefined. Load balancing activities will only be performed when the measured load gap is higher than the load gap threshold in order to minimize unnecessary VHOs. If the load gap is larger than the threshold, new calls will be directed to the least loaded overlapped cell and a portion of users served by the most loaded overlapped cell will be reallocated to the least loaded overlapped cell. Simulation results in [24] show that this algorithm performs better than a fixed load threshold algorithm in terms of call blocking probability.

Dynamic pricing algorithm

A dynamic pricing algorithm was proposed in [25]. This algorithm achieves load balancing by adjusting the price of a service in each overlapped cell rather than directly moving users to the least loaded cells. In this algorithm, a high load threshold and a low load threshold are set. The cell load information is updated periodically. If the cell load is between the high and low load thresholds, the price of serving a user in the cell is fixed at the initial value, however, if the cell load is higher than the high load threshold, the price will be increased by ΔP or 2ΔP dependent on how much the cell load exceeds the high load threshold (ΔP is a predefined amount of increased price). If the cell load is lower than the low load threshold, the price of a service of the cell will be decreased by ΔP or 2ΔP dependent on the difference between the cell load and the low load threshold. This algorithm assumes that users will always select the cheapest cell so that load balancing can be achieved by adjusting the price of a serving cell.

Simulation results in [25] illustrate that the dynamic pricing algorithm outperforms the one without pricing in terms of uplink BLER and revenues when a suitable price updating period is set. However, an assumption made in this algorithm is that all overlapped RATs provide the same QoS to the user, which may not be true in the real world.

Ding's algorithm

In the above algorithms, the occupied load of a RAT is defined as the quotient obtained by dividing the present traffic by the maximum traffic that a RAT can serve. However, Ding et. al [26] proposed an algorithm, in which the RAT load is expressed at a deep level. For example, if it is assumed that a WCDMA network can serve 8 voice calls, 40 video calls, and one 384 kbps data call simultaneously, and there are currently 3 voice and 19 video calls being served, the WCDMA network load situation in Ding's algorithm is then expressed as voice call 3/8, video call 19/40, and 384 kbps data call 0/1. Simulation results in [26] proves that using Ding's algorithm can reduce the call blocking probability compared to a LB based algorithm, in which the current RAT load is represented by a percentage of the maximum RAT load. The reason is that in Ding's algorithm, the CRRM entity not only knows the load situation but also the resource and traffic distribution. The more information known by the CRRM entity, the better decision it can make. However, in Ding's algorithm, a challenge is how to decide the numbers of different types of calls that can be served simultaneously. For example, it can be said that a RAT can serve 8 voice calls and 40 video calls simultaneously but if we reduce the number of served video calls, the number of voice calls that can be served will increase.

In this section, a number of LB based algorithms have been discussed. A shortcoming of these LB based algorithms is that they only consider the load balancing aspect, which is insufficient to provide an optimized solution. In LB based algorithms, it is assumed that multiple RATs have exactly the same coverage area, network topology, and capacity, which is not true in the real world.

2.3.2 Coverage based algorithms

In [27], an “Always WWAN” algorithm is proposed for co-located WWAN/WLAN networks. WWAN is selected as the default RAT for any types of call, because it has a larger coverage area. This algorithm can minimize the number of VHOs, however, it is inefficient in terms of radio resource usage due to the limited capacity of WWANs. In [28], the authors proposed an algorithm that allocates users to the RAT with the smaller coverage first so that more users outside its coverage area can be served by the RAT with larger coverage area and the call blocking probability can then be reduced. However, compared to the "Always WWAN" algorithm, this algorithm may cause more VHOs.

2.3.3 "WLAN if coverage" algorithm

A "WLAN if coverage" algorithm for integrated WWAN/WLAN networks has been proposed in [29], in which calls within a hotspot area (an area where both WWAN and WLAN have coverage) should always be connected to WLAN due to its higher bandwidth and cheaper cost. Compared to the "Always WWAN" algorithm, the "WLAN if coverage" algorithm can achieve higher user throughput and reduce the service cost, however, it will cause a larger amount of VHOs, especially for high mobility users.

In [30], the "WLAN if coverage" algorithm is compared with the "Always WWAN" algorithm. The results show that the "Always WWAN" algorithm performs better than the "WLAN if coverage" algorithm when most of the users are outdoor, while the "WLAN if coverage" algorithm is better on the contrary case.

2.3.4 Service based algorithms

Service based algorithms allocate a call to a particular RAT based on user service types and RAT properties. A number of service based algorithms are discussed in this section.

In [31], Koo et al. proposed a service based algorithm for a co-located GERAN/W-CDMA network. Two types of calls, voice and data are considered. A new call is allocated to the RAT with the smallest expected relative resource consumption for the service class of the call. Simulation results in [31] show that Koo's algorithm can improve the Erlang capacity compared to the random selection algorithm.

Song et al. [32] proposed a service based algorithm for a co-located UMTS/WLAN network. In Song's algorithm, voice calls are allocated to UMTS (unless there is not enough capacity in UMTS) while data calls are allocated to WLAN. Song's algorithm is compared with a "WLAN if coverage" algorithm (both new voice and data calls are allocated to WLAN in the double-coverage area) in [32] and simulation results show that Song's algorithm can reduce the number of HOs for voice calls, because UMTS has larger coverage than WLAN.

The above service based algorithms only consider a two co-located RATs scenario. In [33], the service based algorithm has been extended to be suitable for a co-located GERAN/UMTS/WLAN scenario. Voice calls are allocated according to the following orders: GERAN, UMTS, and WLAN and data calls are allocated in the inverse order. Video calls are allocated in the order of UMTS, WLAN, and GERAN. Simulation results in [33] prove that by using this RAT selection algorithm, the system performance can be improved in terms of blocking probability and downlink average throughput compared to the random RAT selection algorithm. However, this algorithm assumes video calls can be served by GERAN, which is impractical in the real world.

Abuhaija and Al-Begain [34] proposed an algorithm for a scenario where three RATs, GSM/GPRS, WCDMA, and HSDPA are co-located. In Abuhaija and Al-Begain's algorithm, voice calls are allocated to WCDMA first and then GSM/GPRS, streaming services, such as Voice over IP (VoIP), streaming video, and mobile TV, are allocated to HSDPA first and then WCDMA, best effort services are allocated to HSDPA first and then GSM/GPRS. Simulation results in [34] show that the system throughput for voice and streaming services can be increased by using this algorithm compared to the random selection based algorithm.

2.3.5 Path loss based algorithm

In [35, 36], a Network Controlled Cell Breathing (NCCB) algorithm has been proposed for co-located GERAN/UTRAN networks. In FDMA/TDMA systems, the intra-cell interference is minimal. However, in CDMA systems, every user transmitting data in a CDMA cell is a source of interference to all other users served in the same cell. The higher the path loss, the higher the required transmission power and the higher the interference level generated.

The basic idea of the NCCB algorithm is to allocate high path loss users to FDMA/TDMA networks and low path loss users to CDMA networks. For the initial RAT selection, the path loss is measured at a regular interval and averaged. If the path loss of a call is higher than a given threshold, it will be directed to GERAN; otherwise, it will be directed to UTRAN. If there is not enough capacity in the preferred RAT, another RAT will be selected. If both RATs are fully loaded, the call will be blocked.

For VHO, the procedure is similar, however, in order to avoid the ping-pong effect, a hysteresis threshold margin is introduced and a number of consecutive samples will be measured before making a decision. Simulation results in [35, 36] illustrate that by setting an appropriate path loss threshold, the NCCB algorithm performs better than a LB based algorithm in terms of BLER, required BS transmission power, user throughput, blocking and dropping probabilities. Detailed practical implementation issues of the NCCB algorithm are discussed in [37].

There are several weaknesses of the NCCB algorithm. First of all, the setting of the path loss threshold is a challenge. If the threshold is too high, the radius of the CDMA cell will be too large. In this case, the NCCB algorithm will even cause higher blocking and dropping probabilities than the LB based algorithm due to the high intra-cell interference level in CDMA networks [37]. However, if the threshold is too low, the CDMA cell radius will be too small. The users inside the CDMA network coverage area will get better QoS, however, the QoS of the users outside the area will degrade [38]. Secondly, the NCCB algorithm does not consider the penetration loss for indoor users, which may increase the required transmission power and in turn the intra-cell interference level.

2.3.6 User satisfaction based algorithm

In [39], Delicado and Gozalvez proposed a User Satisfaction Based Selection (USaBS) algorithm for co-located GPRS/EDGE/HSDPA networks. In this algorithm, a call's "demand" is defined as the throughput necessary to guarantee a pre-established satisfaction level of a user dependent on the requested service and the user contract. An "offer" is defined as an estimated throughput of a call in a RAT using previous transmission information. For a new call, all RATs providing an "offer" higher than the "demand" are the candidate RATs. The most suitable RAT from a set of candidate RATs is the one providing the lowest "offer". The purpose of this algorithm is to avoid the situation that low "demand" users occupy unnecessary radio resources. Simulation results in [39] show that the USaBS algorithm can guarantee the satisfaction levels for different users, independent of their service and contract types.

In [40], a variant of USaBS, User Satisfaction with Low Resources Selection (USaLoR) algorithm is proposed. In this algorithm, the most suitable RAT from a set of candidate RATs is the one that can use the least amount of radio resources to satisfy a user's "demand". The purpose of this algorithm is to prevent the situation that a low performance RAT uses a large amount of resources for a single call. Simulation results in [40] show that compared to USaBS, the USaLoR algorithm can reduce the probability to saturate low performance RATs, however, it decreases the user satisfaction rate.

2.4 Multiple criteria based algorithms

The algorithms introduced above use a single criterion to make RAT selection decisions. In this section, more complicated algorithms using a number of RAT selection criteria are discussed.

2.4.1 Policy based algorithms

The details of policy based algorithms will be discussed in Chapter 3. Three basic RAT selection policies for initial RAT selection are proposed in [7, 41]: Voice GSM/GERAN (VG), Voice UMTS/UTRAN (VU), and Indoor (IN). The VG policy allocates voice calls to GERAN and interactive calls to UTRAN, the VU policy, on the contrast, allocates voice calls to UTRAN and interactive calls to GERAN, and the IN policy allocates indoor users to GERAN and outdoor users to UTRAN. In [7, 41, 42], three algorithms based on VG, VU, and IN policies respectively, are compared. The simulation results prove that the VG based algorithm performs better than the VU based algorithm in terms of data user throughput when the cell radius is larger than 1km. The main reasons are two-fold. From the voice users' point of view, if the cell radius is larger than 1km, UTRAN users at the cell edge will experience more transmission errors due to the power limitations and the interference-limited nature of WCDMA technology. From the interactive users' point of view, they can get a higher bit rate in UTRAN than in GERAN. The simulation results in [7, 41] demonstrate that a IN based algorithm outperforms the random selection algorithm in terms of uplink BLER. This is because indoor users cause higher interference levels than outdoor users in WCDMA systems due to the additional penetration loss, which will degrade the system performance [43].

The above policy based algorithms are simple, because they only use one policy. However, they have an obvious shortcoming. For example, for the VG based algorithm, when the capacity of GERAN is full, even though there are free resources available in UTRAN, voice calls will still be blocked. In order to solve this problem, complex RAT selection algorithms have been studied in the literature.

Pérez-Romero et al. [41] proposed three two-complex policy based algorithms: VG*VU, IN*VG, and VG*IN. The general format is Policy 1*Policy 2. A new call will be allocated using Policy 1 first. If the capacity of the preferred RAT is full, then the call will be assigned using Policy 2. For example, in the VG*IN algorithm, an outdoor voice call will be allocated to GERAN according to the VG policy. If the capacity of GERAN is full, the call will be assigned to UTRAN according to the IN policy. In these algorithms, a service request is only blocked when both of the two policies are violated so that the blocking probability can be significantly reduced.

Table 2.1 compares the differences among the three two-complex algorithms in terms of RAT selection priority.

Table 2.1: Comparison between VG*IN, IN*VG, and VG*VU algorithms

Service type	VG*IN	IN*VG	VG*VU
Voice and indoor	Select GERAN only	Select GERAN only	Select GERAN first and then UTRAN
Voice and outdoor	Select GERAN first and then UTRAN	Select UTRAN first and then GERAN	Select GERAN first and then UTRAN
Data and indoor	Select UTRAN first and then GERAN	Select GERAN first and then UTRAN	Select UTRAN first and then GERAN
Data and outdoor	Select UTRAN only	Select UTRAN only	Select UTRAN first and then GERAN

Simulation results in [41] prove that VG*IN and VG*VU algorithms outperform the IN*VG algorithm when the number of data calls are much higher than the number of voice calls. This is because in the IN*VG algorithm, a higher number of indoor interactive calls are allocated to GERAN, which causes higher delay and lower throughput. However, when the number of voice calls increases, the IN*VG algorithm becomes better because the IN policy can reduce the number of high interference level users in UTRAN.

A performance comparison between the VG*VU and a LB based algorithms has been carried out in [44]. The simulation results demonstrate that compared to the LB based algorithms, the VG*VU algorithm can reduce the average weighted packet delay for interactive users. The reason is that in the VG*VU algorithm, voice users are allocated to GERAN first so that more UTRAN radio resources are available for data users. However, the LB based algorithm works better in terms of total uplink aggregated throughput because in the VG*VU algorithm, more voice users will be served by GERAN, which in turn causes higher load conditions in GERAN and higher dropping probability. Therefore, the throughput contribution of the voice users decreases.

2.4.2 Variations of NCCB algorithm

Two variations of the NCCB algorithm were introduced in [45]. One is called NCCB-voice, where the NCCB policy is only applied for voice users while interactive users follow the VG policy. Another is called VG-NCCB, where the low path loss users follow the VG policy and high path loss users are always allocated to GERAN first. Simulation results in [45] prove that the NCCB algorithm performs better than NCCBvoice and VG-NCCB algorithms in terms of uplink BLER. However, it also has the highest uplink delay for interactive users among the three algorithms.

2.4.3 Utility/cost-function based algorithms

In [46, 47, 48, 49], a novel fittingness factor based RAT selection algorithm has been proposed for both initial RAT selection and VHO. In this algorithm, every candidate RAT is weighted by a parameter called fittingness factor (ranging from 0 to 1), which is [48]:

$$\Psi_{i,p,s,j} = C_{i,p,s,j} \times Q_{i,p,s,j} \times \delta(\eta_{NF}), \tag{2.8}$$

where, C, Q, and $\delta(\eta_{NF})$ refer to capability, user-centric suitability, and network-centric suitability of the j^{th} RAT for each i^{th} user, who belongs to the p^{th} user profile requesting the s^{th} service respectively. In order to work out the fittingness factor, we need to calculate the values of C, Q, and $\delta(\eta_{NF})$.

The first parameter C reflects both terminal and network capabilities. The value of C is calculated by [47]:

$$C_{i,p,s,j} = T_{i,p,j} \times S_{s,j}, \tag{2.9}$$

where T is the terminal capability and S is the RAT capability. If the terminal of the i^{th} user belonging to the p^{th} profile does not support the j^{th} RAT, $T = 0$, otherwise

$T = 1$. If the s^{th} service is not supported by the j^{th} RAT, $S = 0$, otherwise, $S = 1$.

The second parameter Q reflects the suitability of a RAT to support a particular user service. The calculation of Q varies dependent on different user services and RATs. In GERAN, for voice users [46]:

$$Q_{i,p,voice,GERAN} = \begin{cases} 1 & \text{if } L_i \leq L_{max}, \\ 0 & \text{if } L_i > L_{max}. \end{cases} \tag{2.10}$$

where L_i is the path loss for the i^{th} user and L_{max} is the maximum allowed path loss.

For interactive users [46]:

$$Q_{i,p,interactive,GERAN} = \frac{R_{MCS} L_i}{R_{bmax,s,p}} \cdot min(\varphi_p, M), \tag{2.11}$$

where R_{MCS} is the maximum allowable user bit rate dependent on the Modulation and Coding Scheme (MCS) and path loss L_i, $R_{bmax,s,p}$ is the maximum theoretical bit rate that can be achieved by the p^{th} user profile requesting the s^{th} service among all overlapped RATs, φ_p is the multiplexing factor that reflects the average number of slots per frame allocated to the s^{th} service, and M is the multislot capability, which is the ability for multiplexing and multiple access on the radio path of a network.

For UTRAN, uplink and downlink fittingness factors are calculated separately. For voice and videophone users, the uplink user-centric suitability is calculated in a similar way as the one for GERAN [46]:

$$Q_{UL,i,p,voice,UTRAN} = \begin{cases} 1 & \text{if } L_i \leq L_{max}, \\ 0 & \text{if } L_i > L_{max}. \end{cases} \tag{2.12}$$

The downlink user-centric suitability is computed by [46]:

$$Q_{DL,i,p,voice,UTRAN} = \begin{cases} 1 & \text{if } P_{Ti} \leq \Delta P_{max,p,s}, \\ 0 & \text{if } P_{Ti} > \Delta P_{max,p,s}. \end{cases} \tag{2.13}$$

where P_{Ti} is the required power for the i^{th} user and $\Delta P_{max,p,s}$ is the maximum power available for the i^{th} user with p^{th} user profile.

For interactive users, the uplink user-centric suitability is given by [46]:

$$Q_{UL,i,p,interactive,UTRAN} = \frac{f(R*)}{R_{bmax,s,p}} \varphi_p, \tag{2.14}$$

where $f(R*)$ is the maximum bit rate available for the user, φ_p is the multiplexing factor that refers to the average number of users served with respect to the total number of users of service profile p with data in their buffers.

The calculation of downlink user-centric suitability is the same as the uplink one [46],

$$Q_{DL,i,p,interactive,UTRAN} = \frac{f(R*)}{R_{bmax,s,p}} \varphi_p, \tag{2.15}$$

The third parameter $\delta(\eta_{NF})$ reflects the suitability from an overall RAT perspective. The definition of the network-centric suitability is given by [47]:

$$\delta(\eta_{NF}) = \begin{cases} 1 & \text{if } \eta < 1 - min(\eta_{NF}, D), \\ (\frac{1-\eta}{min(\eta_{NF}, D)})^2 & \text{if } \eta > 1 - min(\eta_{NF}, D) \\ & \text{and traffic is flexible.} \end{cases} \tag{2.16}$$

where η is the normalized load in the RAT and η_{NF} is the non-flexible load in the RAT. The non-flexible load refers to the load from non-flexible traffic, which is the traffic that can only be served by a specific RAT so that it can not provide flexibility to CRRM. For example, video calls can only be served by UTRAN, so they are non-flexible traffic for UTRAN. Parameter min(η_{NF},D) refers to the load reserved for non-flexible traffic in the RAT. From (2.16), we can see that the higher the amount of non-flexible load in a given RAT, the lower the network-centric suitability value for flexible traffic.

After working out the values of capability, user-centric suitability, and network-centric suitability, the uplink and downlink fittingness factors for each RAT can be calculated separately and then be combined as follows [46, 47]:

$$\Psi_{i,p,s,j}(K_j) = \alpha_{p,s}\Psi_{UL,i,p,s,j}(K_j) + (1 - \alpha_{p,s})\Psi_{DL,i,p,s,j}(K_j), \tag{2.17}$$

where $\alpha_{p,s}$ is a weighting factor for candidate cell K_j. $\alpha_{p,s}$ is close to 1 if the uplink is more important and close to 0 when the downlink is more important.

After solving the fittingness factor values of all RATs for a user service, the one with the highest value is selected as the target RAT. Admission control will then be performed to see if the user can be served in the selected RAT. If not, the RAT with the second highest fittingness factor will be selected and so on. If the admission process fails in all RATs, the service request will be blocked.

For on-going calls, the fittingness factor of every candidate RAT is measured at a regular interval. A VHO will be performed if the averaged value of the fittingness factor meets the following condition [46, 47, 48]:

$$\Psi_{i,p,s,j}(K_j) = \Psi_{i,p,s,servingRAT}(servingcell) + \Delta VHO, \tag{2.18}$$

where ΔVHO is a predefined VHO threshold.

According to simulation results in [46, 47, 48, 49], the fittingness factor based RAT selection algorithm is able to reduce both downlink and uplink average packet delay for interactive users compared to a LB based algorithm because the fittingness factor based algorithm considers a number of factors that may influence the performance, rather than only the load factor. However, this algorithm has its shortcomings too. First of all, the equation (Equation (2.17)) being used to calculate the overall fittingness factor is incorrect. For radio communications, a call is accepted only when it meets both uplink and downlink requirements. The fittingness factor should be 0 if either the uplink or the downlink requirements are not satisfied. Equation (2.17) can be modified as follows:

$$\Psi_{i,p,s,j}(K_j) = \begin{cases} 0, \text{if } \Psi_{UL,i,p,s,j}(K_j) \times \Psi_{DL,i,p,s,j}(K_j) = 0, \\ \alpha_{p,s}\Psi_{UL,i,p,s,j}(K_j) + \\ (1-\alpha_{p,s})\Psi_{DL,i,p,s,j}(K_j), \text{ if not.} \end{cases} \tag{2.19}$$

The second problem of this algorithm is that it does not consider RAT load when calculating the fittingness factor. It is a waste of time and resources to calculate the fittingness factor of a RAT that has no free capacity. It is better to integrate the load parameter into the calculation of the fittingness factor. If the load of a RAT is full, its fittingness factor is set to 0 and it will not be considered as a candidate RAT. The third problem is that the RAT selection algorithm for ongoing calls does not consider the handover cost. A RAT with higher fittingness factor may have higher handover cost (such as signaling overhead, handover delay) too. It is better to make a balance between the two. According to [50], a RAT selection for an ongoing call is dependent on the difference between handover gain (the benefits obtained from a VHO, such as increased throughput) and handover cost (such as lost throughput caused by VHO delay).

In [51], the fittingness factor based algorithm and the NCCB algorithm have been compared in terms of voice service performance in a co-located GERAN/UTRAN network. Simulation results in [51] show that if the network load is light, under the two algorithms, especially the fittingness factor based algorithm, most of the voice users are allocated to GERAN. However, when the network is overloaded, most of the users will be served by UTRAN. In terms of call blocking and dropping probabilities, the NCCB algorithm outperforms the fittingness factor based algorithm when the network is low to medium overloaded, however, the fittingness factor algorithm works better when the network is highly overloaded.

A force based RAT selection algorithm is proposed by Pillekeit et al for co-located UMTS/GSM networks [52]. In Pillekeit's algorithm, a "force" is defined for each cell. Every "force" consists of four sub-forces: load force (the available resources in the target cell after a HO), QoS force (the difference of QoS, such as throughput between the source and target cells), migration attenuation force (the time since the last VHO occurred), and handover force (the signaling overhead of VHOs). The load force is an attractive force, the migration and handover forces are repelling while the QoS force can be either attracting or repelling. The importance of each sub-force is described by a weighting factor. The total force of a target cell k for user i $F_{sum,k}(i)$ is the result from the superposition of all sub-forces [52]:

$$F_{sum,k}(i) = C_L F_{L,k}(i) + C_{QoS} F_{QoS,j,k}(i) - C_M F_{M,k}(i) - C_{HO} F_{HO,k}(i), \tag{2.20}$$

where C is the weighting factor, F_L, F_{QoS}, F_M, and F_{HO} represent the load, QoS, migration, and HO cost forces respectively, j is the source cell number.

The overlapped cell with the largest force value will be selected as the target cell. Simulation results in [52] prove that the force based algorithm can achieve a better performance in terms of load balancing, overall traffic capacity, and QoS compared to the random selection algorithm.

In [53], Yu and Krishnamurthy proposed a RAT selection algorithm aimed to maximize the overall network revenue and guarantee QoS constraints in an integrated WLAN/CDMA network. This algorithm is formulated as a Semi-Markov Decision (SMD) problem whose state space is defined by a set of WLAN QoS constraints: throughput, average delay, and CDMA network Signal to Interference Ratio (SIR) outage blocking probability. The optimal solution of the SMD problem is then solved by linear programming techniques. The performance of Yu and Krishnamurthy's algorithm is compared with two reference algorithms, in which the admission control is performed independently in WLAN and CDMA networks. The results show that Yu and Krishnamurthy's algorithm can achieve higher revenue. Yu and Krishnamurthy's algorithm emphasizes on the operator's perspective. A challenge of this algorithm is how to set suitable QoS constraints to balance operator and user requirements.

2.4.4 Adaptive algorithm for co-located WWAN/WLAN networks

An adaptive RAT selection algorithm designed for both initial RAT selection and VHO for co-located WWAN/WLAN networks was proposed by Hasib et al [30]. It decides the serving RAT according to a list of parameters: service type, RAT load, mobility and location prediction information, and service cost. An assumption made by Hasib is that the user location information can be predicted.

The initial RAT selection algorithm works as follows. If a user is predicted to remain in a hotspot area during a session time, WLAN is the preferred RAT. If a user is expected to exit the hotspot during a session time, service type and network load factors are then considered for RAT selection. WWAN is preferred for RT services to avoid VHOs. For NRT services, WLAN is selected if the WWAN is highly loaded, otherwise, a location prediction scheme is used to decide whether a user will move out of the hotspot area soon or not. If yes, WWAN is selected to avoid VHOs, otherwise, WLAN is chosen.

The VHO algorithm works as follows. If a user is moving out of a hotspot and is currently connected to WLAN, a VHO to WWAN is performed. If a user is moving into a hotspot and the service session is long, a VHO is performed to WLAN for NRT sessions. For RT sessions, VHO will be performed if the user is expected to remian within the hotspot.

In [30], the proposed adaptive algorithm is compared with the "Always WWAN" and "WLAN if coverage" algorithms in terms of call blocking probability. Simulation results in [30] prove that the performance of Hasib's algorithm is better than the other two in terms of new call blocking probability because it allocates users according to a number of criteria rather than just allocates users in a predefined order. However, it is more complex and requires more information. A challenge of this algorithm is that it relies on the location prediction information, which may be hard to be obtained in practice. The QoS negotiation framework and detailed signaling procedures for this algorithm are discussed in [54].

2.4.5 Fuzzy logic based algorithms

A number of RAT selection algorithms applying the concept of fuzzy logic have been studied in the literature. A fuzzy-neural based RAT selection algorithm that considers both technical and non-technical aspects (e.g. user demands and operator preferences) is given in [55, 56]. This algorithm contains three main procedures: fuzzy neural, reinforcement learning, and multiple decision-making. The fuzzy neural procedure aims to allocate a numerical indication named Fuzzy Selected Decision (FSD) to each RAT. The value of a FSD is between 0 to 1, which is determined by a set of linguistic variables, such as signal strength, resource availability, and mobile speed. The RAT with the highest FSD value is selected. The reinforcement procedure is used to select and adjust parameters used in the fuzzy-neural algorithm to ensure a target value of a given QoS parameter. The detailed reinforcement procedure can be referred to [57, 58]. Finally, the multiple decision-making procedure is performed to make a final decision on RAT selection using FSD values, user demands, and operator preferences.

A number of RAT selection algorithms using similar concept as the above algorithm but using different RAT selection criteria have been studied in the literature. In [59], Chan et al., presented a RAT selection algorithm using fuzzy Multiple Objective Decision Making (MODM). Chan's algorithm makes RAT selection decisions using seven criteria: signal strength, bandwidth, charging model, reliability, latency, battery status, and priority. Guo et al. [60] proposed a fuzzy multiple objective decision based algorithm using cell type, data rate, coverage, transmission delay, and call arrival rate as RAT selection criteria.

Zhang [61] proposed an algorithm using a Fuzzy Multiple Attribute Decision Making (MADM) method. In this algorithm, fuzzy logic is used to deal with the imprecise information of RAT selection criteria. The imprecise data are first converted to crisp numbers, and then, classical MADM methods are used to determine the ranking of RATs. The RAT with the highest ranking is then selected as the serving RAT.

In [62, 63], Alkhawlani and Hussein proposed an algorithm using fuzzy logic and Multi-Criteria Decision Making (MCDM) for a co-located WWAN/WMAN/WLAN network. Their algorithm contains two modules: User Software Module (USM) in the user terminal and Operator Software Module (OSM) in the CRRM entity. The USM containing a network-assisted terminal-controlled algorithm reflects the user preference. The network-assisted terminal-controlled algorithm contains two components: the fuzzy logic based control component and the MCDM component. The fuzzy logic based control component has four fuzzy logic based subsystems considering four user selection criteria separately: reliability, security, battery power, and price. The inputs of the four subsystems are the user preferred price, user preferred reliability, user preferred security, and the importance of battery power for the user respectively, and each subsystem has three outputs: the probabilities of acceptance for the user in WWAN, WMAN, and WLAN respectively. The MCDM component uses the outputs of the fuzzy logic based control component as inputs and works out ranking values for the three RATs by allocating a weighting factor on each criterion.

The OSM containing a terminal-assisted network controlled algorithm reflects the operator's point of view. The OSM has a fuzzy logic based control component and a MCDM component too. The fuzzy logic based control component has four subsystems considering received signal strength criterion, mobile station speed criterion, service type criterion, and radio resources availability criterion respectively. The inputs of the four subsystems are received signal strengths of the three RATs, Mobile Station (MS) speed, delay limit and required bit rate, and radio resources availability respectively. The outputs are the probabilities of accepting the user in each RAT dependent on each criterion. These outputs and the outputs of USM then becomes inputs of the MCDM component. The final ranking value of the three RATs are solved by allocating weighting factors on each criterion and the user preference. The RAT with the highest ranking value is selected as the serving RAT.

Alkhawlani and Hussein compared their algorithm with three reference algorithms: random selection based, terminal speed based, and service based. The results show that compared to other algorithms, in their algorithm, higher percentages of users can be allocated to their preferred RATs, with better QoS conditions, and lower cost.

2.5 Summary

This chapter reviews a number of RAT selection algorithms for the heterogeneous wireless networks. These algorithms have been grouped into three families: random selection based, single criterion based, and multiple criteria based. The random selection based algorithm is the simplest one but is inefficient. Single criterion algorithms can improve the system performance in some aspects, however, they make RAT selection decision only dependent on one criterion, which may not meet the requirements of both customers and operators in some cases. Multiple criteria algorithms, which make RAT selection decisions after integrating a number of criteria, are more likely to provide an optimal solution. However, they are complicated and sometimes cumbersome to use. A tradeoff needs to be made between the complexity and efficiency of RAT selection algorithms.

Chapter 3

Modeling and simulation

In this chapter, simulation models, performance parameters, and simulation algorithms that will be used in this thesis are discussed. The overall simulation model is shown in Fig. 3.1. This model contains three parts. The first part contains inputs, such as mobility model, radio propagation model, traffic model, and RAT load model, the second part contains RAT selection algorithms, and the third part contains outputs, such as user throughput, blocking probability, dropping probability, and fairness. The three parts will be discussed respectively in the rest of this chapter.

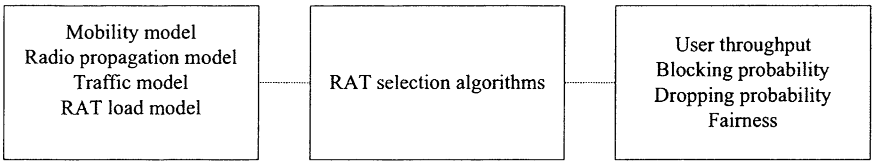


Figure 3.1: Simulation model

3.1 Simulation models

In this thesis, MATLAB is used to implement simulation models and the following models are used.

3.1.1 Mobility model

The mobility model is used to update user positions based on a random user movement pattern. Users are initially distributed randomly within the simulation area. The location of the u^{th} user at time $t+1$, $position_u(t+1)$ is:

$$position_u(t+1) = position_u(t) + v_u(t)e^{i\theta_u(t)}, \tag{3.1}$$

where $v(t)$ and $\theta(t)$ are the user velocity and direction at time t respectively, index u denotes the u^{th} user. $\theta(t)$ is given by:

$$\theta_u(t) = \theta_u(t-1) + 2 \cdot rand \cdot \pi. \tag{3.2}$$

where $rand$ is a random number between 0 and 1.

3.1.2 Topology model

It is assumed that users are only allowed to move within a predefined simulation area. The border effect is alleviated by using the wraparound method. The left and right borders, and top and bottom borders are connected to each other. When a user reaches the border of the simulation area, the user will be wrapped around to the opposite side. Fig. 3.2 shows an example of the wraparound method.

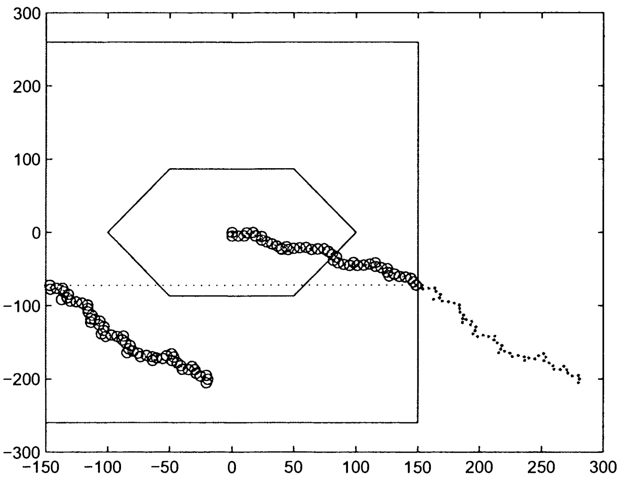


Figure 3.2: An example of the wraparound method

3.1.3 Radio propagation model

In radio transmissions, signal strength decreases with the increase of the distance between the transmitter and the receiver. Path loss (or path attenuation) refers to the reduction in power density (attenuation) of an electromagnetic wave as it propagates through space. The path loss can be modeled as:

$$L_{dB} = \frac{A_l}{d^\gamma}. \tag{3.3}$$

where A_l is a constant which is dependent on the antenna properties, transmission wavelength, environment (rural, suburban, or urban), base station height, etc, d is the distance between transmitter and receiver, and γ is the path loss exponent with typical values ranging from 2 in a free space propagation environment to 5 in a dense urban area.

The above path loss model is a very simple model. A number of complex models have been studied in the literature [64]. In most of the literature, the well known Okumura-Hata propagation model is used:

$$L_{dB} = 46.3+33.9log(f)-13.82log(h_b)-a(h_m)+[44.9-6.55log(h_b)]log(d)+C_m, \quad (3.4)$$

where f is the frequency of the transmission in MHz, h_b is the height of antenna at the base station in meters, h_m is the height of the mobile or receiver in meters , d is the distance between the receiver and the transmitter in km, $a(h_m)$ is the mobile antenna correction factor, and C_m is the correction factor which has a different value for each environment. Usually, a simplified Okumura-Hata propagation model is used in simulations:

$$L_{dB} = A + Blog_{10}(d), \quad (3.5)$$

where A and B are constants computed from a given set of parameters including BS antenna height, mobile station (MS) antenna height, and carrier frequency; d is the distance between BS and MS. In this thesis, the simplified Okumura-Hata propagation model is used.

3.1.4 Traffic model

New calls arrive according to a Poisson processs:

$$f(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \quad (3.6)$$

where k = 0,1,2..., and λ is the average number of arrival calls during a given interval. The holding time of user calls is generated according to an exponential distribution.

The probability density function $f(x,h)$ is given by:

$$f(x,h) = \begin{cases} h \cdot e^{-h \cdot x} & x \geq 0, \\ 0 & x < 0. \end{cases} \tag{3.7}$$

where h is the mean call holding time.

3.1.5 RAT load models

RAT load model 1

RAT load model 1 is a simple model, in which all RATs are assumed to have the same capacity, and there is no service differentiation. The bit rate of all calls is simply assumed to be same.

GSM/GERAN load model

In the GSM model, there are n carrier frequencies in each cell and every carrier frequency contains eight time slots so that a GSM cell has a total number of $8 \times n$ physical channels. Some of the channels are signalling channels and the rest are traffic channels. Each call (voice or data) will be allocated to one traffic channel.

The GERAN model is similar to the GSM model. However, in the GERAN model, if the RAT capacity is insufficient, multiple data users will be forced to share one physical channel. If the capacity is sufficient, each data user can occupy one traffic channel.

There are two ways to measure the RAT load: one is dependent on the maximum resource consumption and another is dependent on the minimum resource consumption. According to the maximum resource consumption method, the GERAN load can be worked out by:

$$L_G = (i + j)/N_c, \tag{3.8}$$

where i is the number of voice users served in GERAN, j is the number of data users served in GERAN, and N_c is the total number of traffic channels in GERAN. In this case, every data user can occupy one GERAN channel.

According to the minimum resource consumption method:

$$L_G = i/N_c + j/(nN_c). \tag{3.9}$$

where n is the maximum number of data users that can share one traffic channel. In the simulations performed in this thesis, it is assumed that voice users have higher service priority than data users in GSM/GERAN network. Data users share the traffic channels not occupied by voice users.

UMTS/UTRAN load model 1

The UMTS/UTRAN load model 1 is discussed in [65]. In UMTS networks, a parameter called load factor is introduced to measure the system load. When a UMTS network is fully loaded, its load factor is one. A safety margin is used because a UMTS system will be unstable if it is fully loaded. Thus a load factor threshold (maximum allowed load factor value) needs to be set. In UMTS networks, the uplink and downlink load factors should be calculated separately. The uplink load factor, η_{UL}, can be calculated as [65]:

$$\eta_{UL} = (1 + f)\sum_{j=1}^{N} \frac{1}{1 + \frac{W}{(E_b/N_0)_j R_j v_j}}, \tag{3.10}$$

where f is the other cells to own cell interference ratio, N is the number of service connections, W is the WCDMA chip rate, E_b/N_0 is the signal energy per bit to noise spectral density ratio, R_j is the bit rate of the j^{th} call, and v_j is the activity factor of the j^{th} call. The downlink load factor, η_{DL}, is given by [65]:

$$\eta_{DL} = \sum_{j=1}^{N} v_j \frac{(E_b/N_0)_j}{W/R_j}(1 - \bar{\alpha} + \bar{f}), \tag{3.11}$$

where $\bar{\alpha}$ is the average orthogonality factor in the cell and $\bar{f}$ is the average other cells to own cell interference ratio. A new service request is accepted if the following conditions is satisfied:

$$New\_\eta_{UL} < \eta_{UL\_threshold}, \tag{3.12}$$

and

$$New\_\eta_{DL} < \eta_{DL\_threshold}, \tag{3.13}$$

where New_η_{UL} and New_η_{DL} are the new uplink and downlink load factors after accepting a new service request, $\eta_{UL\_threshold}$ and $\eta_{DL\_threshold}$ are the uplink and downlink load factor thresholds respectively.

The load of UTRAN, L_U, is given by [65]:

$$L_U = max(L_{U\_UL}, L_{U\_DL}), \tag{3.14}$$

where L_{U_UL}, L_{U_DL} are the uplink and downlink loads of UTRAN respectively:

$$L_{U\_UL} = \frac{\eta_{UL}}{\eta_{UL\_threshold}}, \tag{3.15}$$

$$L_{U\_DL} = \frac{\eta_{DL}}{\eta_{DL\_threshold}}. \tag{3.16}$$

UMTS/UTRAN load model 2

In UMTS/UTRAN load model 1, downlink BS transmission power limitation is not considered. In a UMTS network, part of the BS transmission power is reserved for signalling channels. The downlink transmission power calculation is given by [66]:

$$P_{DL} = \frac{P_N \sum_{j=1}^{N} \frac{(E_b/N_0)_j R_j v_j}{W} L_j}{1 - \eta_{DL}}, \tag{3.17}$$

where P_{DL} is the downlink BS transmission power for traffic channels, P_N is the thermal noise power, and L_j is the loss between BS and the j^{th} UE (including path loss and penetration loss).

A new service request is accepted, if it meets the load factor requirement and the following power requirement:

$$New\_P_{DL} < P_{DL\_max}, \tag{3.18}$$

where New_P_{DL} is the downlink transmission power after accepting the new service request and P_{DL_max} is the maximum BS transmission power allocated to traffic channels. A service request is admitted when it meets all the requirements given in (3.12), (3.13) and (3.18).

In the simulations performed in this thesis, it is assumed that voice users have higher priority than data users in UMTS/UTRAN network. Data users share the resources not used by voice users. Due to the asymmetric property of data services, only downlink load is measured. It is assumed that all data users have the same bit rate requirement. By rearranging (3.11), the downlink data user bit rate is:

$$R = \frac{\eta_{DL\_data} \cdot W}{V_{data}(E_b/N_0)_{data}(1 - \bar{\alpha} + \bar{\bar{f}})}, \tag{3.19}$$

where R is the total downlink bit rate for data users, η_{DL_data} is the downlink load factor of data users, which is given by:

$$\eta_{DL\_data} = \eta_{DL\_threshold} - \eta_{DL\_voice}, \tag{3.20}$$

where η_{DL_voice} is the downlink load factor of voice services. According to (3.17), the total downlink data bit rate can be calculated by:

$$R = \frac{P_{DL\_data} \cdot W(1 - \eta_{DL})}{P_N \cdot (E_b/N_0)_{data} v_{data} \bar{\bar{L}}}, \tag{3.21}$$

where $\bar{L}$ is the average loss of all served data users and P_{DL_data} is the power that can be used by data users. P_{DL_data} can be calculated by:

$$P_{DL\_data} = P_{DL\_max} - P_{DL\_voice}, \tag{3.22}$$

where P_{DL_voice} is the power allocated to voice users.

With the increase of cell size, the path loss will increase and the required BS transmission power will in turn be higher. Usually, when the cell size is large, the maximum available throughput is determined by the BS transmission power, otherwise, it is determined by the load factor.

WLAN load model

A WLAN network has higher capacity than GERAN and UTRAN, however, due to overheads, the available capacity of a WLAN network is much lower than its bandwidth. In a WLAN network, both uplink and downlink traffic share the same bandwidth. The WLAN load calculation is still a research challenge. A simple WLAN load calculation is given by [33]:

$$L_W = \frac{m \times R_v \times 2 + n \times R_d \times 2}{W_{WLAN}}, \tag{3.23}$$

where R_v is the user bit rate for voice calls, R_d is the user bit rate for data calls, m is the number of voice users served in WLAN, n is the number of data users served in WLAN, and W_{WLAN} is the available WLAN capacity. It should be noticed that R_d does not reflect the actual bit rate of a data user served in WLAN. It is only used as a parameter to calculate the WLAN load factor, which need to be used in the LB based algorithm. Data users share the WLAN capacity not used by voice users. A data user can occupy the whole WLAN bandwidth if there are no other users served.

The reason for choosing the above models in this work is that they are common models used by most of the literature in this area. These models can be further improved so that they will be closer to the real world situations. However, due to the time constraints, more complicated models are not considered in this work but they can be included in the future work.

3.2 Simulation algorithms

In this section, a number of simulation algorithms are discussed.

3.2.1 LB based algorithm

The LB based RAT selection algorithm determines the least loaded RAT and allocates new calls to it. A detailed LB based algorithm is as follows:

```
LB based algorithm
t=0
while (t < simulation time), t++
    If a new call arrives
        Check the call location
        Find all RATs that have coverage for the call
        Check the call type
        Find all RATs that have coverage for the call and can serve the call
        Calculate the loads of all available RATs
        Select the least loaded RAT as the target RAT
            If the load of the target RAT is enough to serve the call
                Allocate the call to the target RAT
            Else
                Block the call
```

3.2.2 NCCB algorithm

In the NCCB RAT selection algorithm, high path loss users are allocated to GSM, while low path loss users are allocated to UMTS. A detailed NCCB algorithm is as follows:

```
NCCB algorithm
t=0
while (t < simulation time), t++
    If a new call arrives
        If the path loss of the call > a predefined path loss threshold
            If the GSM capacity is enough to serve the new call
                Allocate the call to GSM
            Else
                If the UMTS capacity is enough to serve the new call
                    Allocate the call to UMTS
                Else
                    Block the call
        Else
            If the UMTS capacity is enough to serve the new call
                Allocate the call to UMTS
            Else
                If the GSM capacity is enough to serve the new call
                    Allocate the call to GSM
                Else
                    Block the call
```

3.2.3 Service based algorithm

In the service based algorithm, a particular type of user is allocated to the RAT that is most suitable to it. A detailed service based algorithm is as follows:

```
Service based algorithm
t=0
while (t < simulation time), t++
    If a new call arrives
        Check the call location
        Find all RATs that have coverage for the call
        Check the call type
        Find all RATs that have coverage for the call and can serve the call
        Find the most suitable RAT for this particular type of call
        Select the RAT as target RAT (1)
            If the load of the target RAT is enough to serve the call
                Allocate the call to the target RAT
            Else
                Remove this RAT from the available RAT list
                Find the second suitable RAT for this particular type of call
                    Loop to (1) until find a RAT that has enough capacity to serve the call
                    If none of the RAT have enough capacity to serve the call
                        Block the call
```

3.2.4 Proposed policy based algorithms

In the last chapter, three two-complex policy based RAT selection algorithms: VG*IN, IN*VG, and VG*VU were discussed. However, for VG*IN and IN*VG algorithms, indoor voice users and outdoor data users will only be assigned to one particular RAT. They can be further improved by allowing them to be allocated to another RAT when the capacity of the preferred RAT is full. Table 3.1 describes the improved VG*IN and IN*VG algorithms.

From Table 3.1, it can be seen that the improved VG*IN algorithm becomes the same as the VG*VU algorithm. The improved IN*VG algorithm becomes a three-complex algorithm: IN*VG*Load, where the Load policy becomes the third policy, which allocates users to the least loaded RAT. For example, if the GSM capacity is full, indoor voice users can be allocated to UMTS according to the Load policy. In a heavily loaded network, the IN*VG*Load algorithm can optimize the system performance by minimizing the number of indoor users in UMTS. However, it does not work well in a lightly loaded network, because allocating indoor data users to GSM is not a good choice when there are sufficient resources in UMTS. The allocation of outdoor voice users to UMTS may also decrease the throughput of data users in UMTS. The details of the reasons will be discussed in the next chapter.

Another new algorithm, called Proposed Policy Based Algorithm 2, is proposed to

Table 3.1: Improved VG*IN and IN*VG algorithms

Service type	Improved VG*IN	Improved IN*VG
Voice and indoor	Select GSM first and then UMTS	Select GSM first and then UMTS
Voice and outdoor	Select GSM first and then UMTS	Select UMTS first and then GSM
Data and indoor	Select UMTS first and then GSM	Select GSM first and then UMTS
Data and outdoor	Select UMTS first and then GSM	Select UMTS first and then GSM

maximize the system performance in low to medium loaded co-located UMTS/GSM networks. The Proposed Policy Based Algorithm 2 aims to minimize blocking and dropping probabilities of voice calls while maintaining a high throughput for data calls in a low to medium loaded network. The details of this algorithm is as follows:

```
Proposed Policy Based Algorithm 2
t=0
while (t < simulation time), t++
    If a new call arrives
        Check the call type
            If it is a voice call
                Check the GSM capacity (1)
                    If the GSM capacity is enough to serve the call
                        Allocate the call to GSM
                    Else
                        Check the UMTS capacity
                            If the UMTS capacity is enough to serve the call
                                Allocate the call to UMTS
                            Else
                                Block the call
            If it is a data call
                Check the call is indoor or outdoor
                    If the call is indoor
                         New UMTS data throughput minus old UMTS data throughput
> 14.4 kbps?
                            If yes
                                Allocate the call to UMTS
                            Else
                                Go to (1)
                    If the call is outdoor
                        Allocate the call to UMTS
```

The Proposed Policy Based Algorithm 2 considers VG, IN, and Load policies. It makes RAT selection decisions based on the difference between new and old UMTS data throughput. In this algorithm, voice calls are allocated to GSM first. If the capacity of GSM is full, they will be served by UMTS. This is because allocation of voice calls into UMTS may decrease the throughput of data users serving by UMTS.

Data and outdoor services are allocated to UMTS, because they satisfy both VG and IN policies. For data and indoor users, if the new data throughput of UMTS (the data throughput of UMTS including the new data user) minus the old throughput of UMTS (the data throughput of UMTS before adding the new data user) is larger than 14.4 kbps, it is allocated to UMTS. Otherwise, if the GSM capacity is not full, it is allocated to GSM. If the GSM capacity is full, it is allocated to UMTS. The rationale is as follows. In GSM, data users can obtain a bit rate of 14.4 kbps. If a new data user is allocated into GSM, the overall data throughput will be increased by 14.4 kbps. If the new UMTS data throughput is larger than the old one plus 14.4 kbps, it means that allocating the new data user into UMTS can obtain a higher overall throughout of data users than allocating it to GSM. So, the new data user should be allocated to UMTS. Otherwise, it should be allocated to GSM.

3.3 Summary

This chapter discusses a number of simulation models, simulation parameters and simulation algorithms. Two improved policy based RAT selection algorithms are proposed in this chapter. The performance of these algorithms will be evaluated using the above simulation models and parameters in the next chapter.

Chapter 4

Performance evaluation

In this chapter, four simulations are carried out to evaluate the performance of RAT selection algorithms.

4.1 Load threshold setting in the LB based algorithm

In Chapter 2, the load threshold of LB based RAT selection algorithm was discussed. However, to the best of our knowledge, existing works did not clearly illustrate the effects of load threshold setting on the performance of LB based RAT selection algorithm. This section focuses on the study of this issue. In this section, the effects of load threshold setting on the performance of load based RAT selection algorithm for real time traffic will be evaluated in terms of load balancing, call blocking/dropping probability, and DR/VHO probability.

It is assumed that there are three overlapped cells (named Cell 1, Cell 2, and Cell 3 respectively), which overlap in the same coverage area but belong to different RATs.

In this simulation, RAT model 1 is used. Simulation parameters are summarized in Table 4.1.

Table 4.1: Simulation parameters

Cell capacity (for Cell 1, 2, and 3)	1 Mbps
User bit rate	100 kbps
Initial load for Cell 1	90% of the maximum cell capacity
Initial load for Cell 2	80% of the maximum cell capacity
Initial load for Cell 1	70% of the maximum cell capacity
Load information update interval	1 time interval
Call arrival rate	1 per time interval (to a randomly selected cell)
Call completion rate	1 per time interval (from a randomly selected cell)

Fig. 4.1 presents the load variation patterns of cells 1, 2, and 3 respectively when a load threshold of 0.8 of the maximum cell capacity is set. Fig. 4.2 presents the load variation patterns when load threshold is not set. It should be noted that the later case equals to set a load threshold of 100% of the cell capacity. It can be seen that the traffic load can be distributed among overlapped cells in a much more balanced manner by setting a proper load threshold.

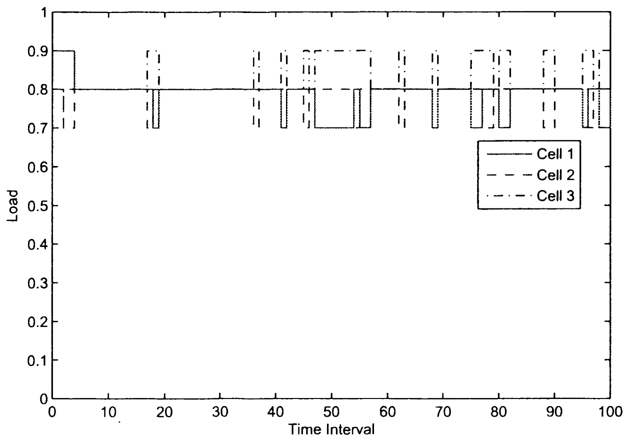
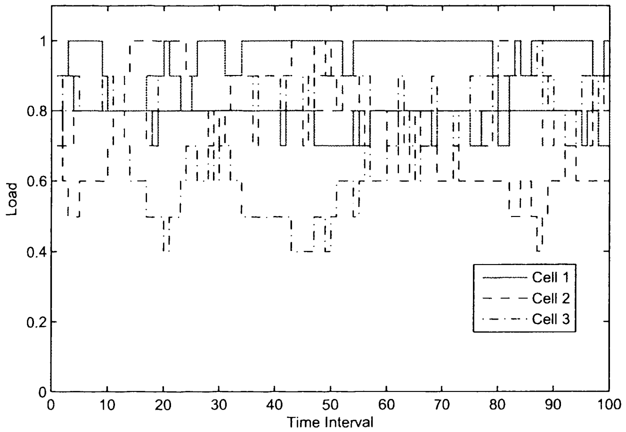


Figure 4.1: Load distribution patterns when a load threshold of 80% of the maximum cell capacity is set and the load information is updated at every time interval



Figurc 4.2: Load distribution patterns when no load threshold is set and the load information is updated at every time interval

In order to get quantitative comparison results, load deviation is used to measure the degree of load balancing. The average load deviation can be obtained by averaging thc sum of load deviation values at all time intervals. In order to get more accurate results, thc simulation is run by a number of times and the load deviation value is further averaged. The results are shown in Table 4.2. When the load information is updated at every time interval, the load deviation value of setting a load threshold is only about 30% of the load deviation value of not setting a load threshold. The effect of setting a load threshold is quite obvious in this case.

Table 4.2: Load deviation values

Load Threshold value	Load deviation value	
	Load information updated at every time interval	Load information updated at every 10 time intervals
0.8	0.0413	0.1415
1	0.1406	0.1895

The above simulation scenario assumes an idea CRRM model, in which the CRRM entity always knows the latest load information of all cells. However, in practice, in order to reduce the amount of overhead, the load information is usually updated in a delayed manner. Thus the CRRM entity sometimes makes decisions based on out-of-date information. The above simulation is redone by keeping all parameters the same as before except that the load information update period is changed from one time interval to ten time intervals. The obtained load deviation values are shown in Table 4.2. It can be seen that the load deviation value of setting a load threshold is about 75% of the load deviation value of not setting a load threshold. In this case, the advantage of achieving load balancing by setting a load threshold is significantly weakened when the load information is updated in a delayed manner.

Fig. 4.3 presents the relationship between call blocking/dropping probability, load information update period, and load threshold. Fig. 4.4 demonstrates the relationship among call DR/VHO probability, load information update period, and load threshold. From Fig. 4.3, it can be seen that when the load information update period is one time interval, the call blocking/dropping probability is always zero regardless of the load threshold value because new calls are always allocated to the least loaded RAT. However, when the load information update period is longer, the call blocking/dropping probability increases with a decrease of the load threshold. From Fig. 4.3, it can be seen when the load information update period is 15 time intervals, the blocking/dropping probability almost doubled when the load threshold is decreased from 100% of the cell capacity to 80% of the cell capacity. This is because in an ideal CRRM model, a call will never be blocked or dropped unless the capacities of all overlapped cells are full. The setting of load threshold or not has no influence on the call blocking/dropping probability. However, in practice, a call may be blocked or dropped even though some of the cells still have sufficient capacities. The reason is that some calls may be directed to wrong cells due to the out-of-date load information in the CRRM entity.

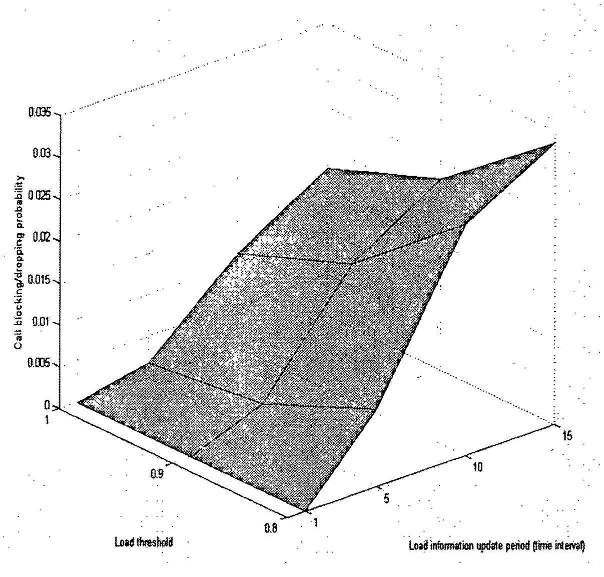


Figure 4.3: Relationship among load information update period, load threshold, and call blocking/dropping probability

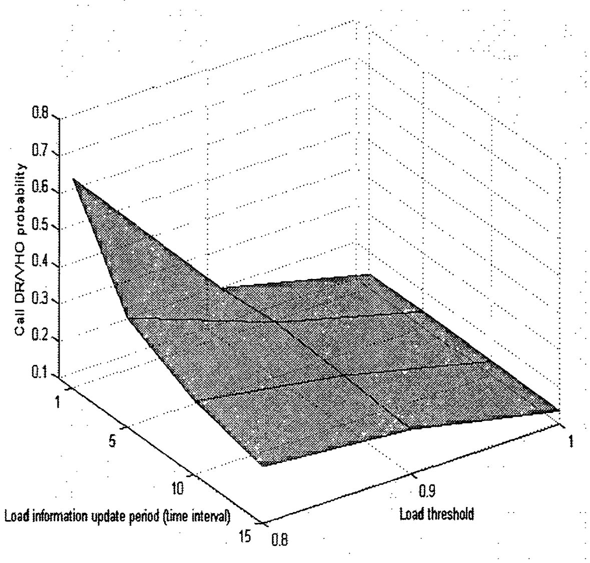


Figure 4.4: Relationship among load information update period, load threshold, and call DR/VHO probability

Let us look at an example. At time t, a call arrives at Cell 1, which is fully loaded. Load information stored in the CRRM entity at time t shows that Cell 2 is the least loaded overlapped cell of Cell 1 and it has free capacity to serve the new call. The call will then be directed to Cell 2. However, due to the out-of-date load information, the actual load of Cell 2 is full at time t. The call will be rejected by Cell 2 and dropped even though Cell 3 has free capacity to serve it. If we set a load threshold, the CRRM entity will start to direct calls to overlapped cells when the load of the current cell is above the threshold. The lower the threshold, the higher the probability that a DR or VHO will occur (shown in Fig. 4.4). For example, the DR/VHO probability tripled when the load threshold decreases from 0.9 of the cell capacity to 0.8 of the cell capacity. As mentioned before, there is a risk of call blocking/dropping for DR/VHO due to the out-of-date load information. So, higher DR/VHO probability will in turn cause higher blocking/dropping probability. If we don't set a load threshold, a call will only perform DR/VHO when the load of the current cell is full. In this case, the probability of the DR/VHO is minimized and in turn the call blocking/dropping probability is minimized. Higher DR/VHO probability will also cause more frequent DR/VHO actions, which in turn causes more overheads.

In summary, setting a proper load threshold may achieve a more balanced load distribution among overlapped RATs. However, it also may cause higher DR/VHO probability and in turn higher overhead and blocking/dropping probability. Tradeoffs need to be made before making decisions.

4.2 Performance comparison of three algorithms

Although in [33], some simulations have been performed to demonstrate that in a co-located GERAN/UTRAN/WLAN network, using CRRM can achieve better performance than the case not using CRRM, to the best of our knowledge, there are no detailed performance evaluation works being carried out. This section evaluates the performance of three RAT selection algorithms in terms of traffic distribution, blocking probability, throughput, and throughput fairness. The fairness index is used to measure the throughput fairness.

It is assumed that three RATs, GERAN, UTRAN, and WLAN co-exist in the same coverage area as shown in Fig. 6.3. In order to simplify the complexity, users are assumed arriving and moving only within the hotspot area. Two different service types are considered, voice and data. Data traffic is assumed to be symmetric.

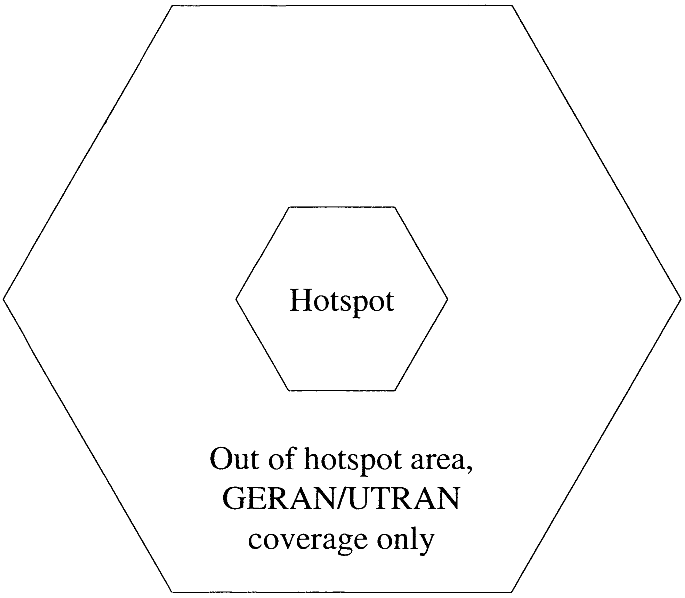


Figure 4.5: Network topology

The following three RAT selection algorithms will be evaluated in this section:
Algorithm 1: It is LB based. RAT loads are calculated based on the maximum resource consumption when the traffic load is low. However, when the traffic load is high (the loads of all the three RATs reaches 1 by using the maximum resource consumption calculation), RAT loads are then calculated based on the minimum resource consumption.
Algorithm 2: It is LB based too, however, RAT loads are directly calculated based on the minimum resource consumption.
Algorithm 3: It is a service based algorithm, in which voice users are allocated to GERAN, UTRAN, and WLAN in order and data users are allocated in an inverse order.

Because WCDMA systems are uplink limited, only uplink load factor is considered in this simulation. GSM/GERAN model, UMTS/UTRAN model 1, and WLAN model are used in this simulation. The detailed simulation parameters are summarized in Table 4.3:

Table 4.3: Simulation parameters

GERAN		
Parameters	Voice	Data
User bit rate	12.2 kbps	Minimum: $16kbps$ Maximum: 59.2 kbps
Number of carrier frequency	3	
Number of data users can share one channel	3	
UTRAN		
Parameters	Voice	Data
Activity factor	0.67	1
E_b/N_0	6dB	5dB
User bit rate	12.2 kbps	Minimum: 16kbps Maximum: 128kbps
Uplink load factor threshold η_{UL}	0.75	
WCDMA chip rate W	3.84 Mcps	
Carrier frequency	1950MHZ	
WLAN		
Parameters	Voice	Data
Type of RAT	802.11b	
Available bandwidth	6 Mbps	
User bit rate	22.8kbps [33]	Minimum: 16kbps Maximum: 128kbps

Table 4.4 works out the load factors for different types of user services in different RATs. For example, a load factor of 0.0417 means that a user occupies 4.17% of the total RAT capacity. Obviously, in a LB based algorithm, the lower load factor of a user for a RAT, the more likely the user will be allocated to that RAT. For example, voice users are more likely to be allocated to WLAN, because its load factor in WLAN is the lowest. In order to compare user load factors in different load calculation schemes, normalized user load factors are listed in Table 4.5. From Table 4.5, it can be seen that data users are more likely to be allocated to GERAN in Algorithm 1 and to WLAN in Algorithm 2.

Table 4.4: Load factors

	GERAN	UTRAN	WLAN
Load factors for voice users	0.0417	0.0084	0.0076
Load factors for data users according to the maximum resource consumption calculation	0.0417	0.0954	0.0427
Load factors for data users according to the minimum resource consumption calculation	0.0139	0.0130	0.0053

Table 4.5: Normalized load factors

	GERAN	UTRAN	WLAN
Load factors for voice users	0.7227	0.1456	0.1317
Load factors for data users according to the maximum resource consumption calculation	0.2319	0.5306	0.2375
Load factors for data users according to the minimum resource consumption calculation	0.4317	0.4037	0.1646

Figs. 4.6 to 4.8 present the user distribution patterns of the three algorithms respectively. The numbers of users will increase with the increase of the traffic load and then start to level out when the network load reach to its maximum value. When the traffic load is low, in Algorithm 1, more voice users are allocated to UTRAN and less voice users are served by WLAN. This is because in Algorithm 1, the load factor of data users for UTRAN is much higher than the load factors of data users for GERAN and WLAN, so that less data users are served by UTRAN. Therefore, more UTRAN capacities are available for voice users and most of the voice users are allocated to UTRAN, which in turn causes less voice users in WLAN. In Algorithm 3, none of the data users is served by GERAN and UTRAN, because all GERAN and UTRAN capacities are occupied by voice users.

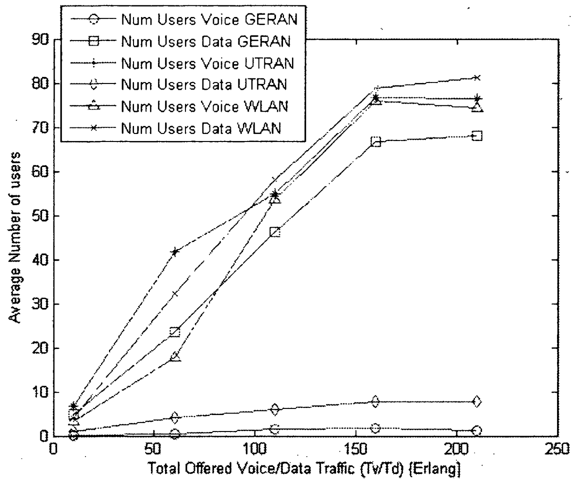


Figure 4.6: User distribution patterns for Algorithm 1

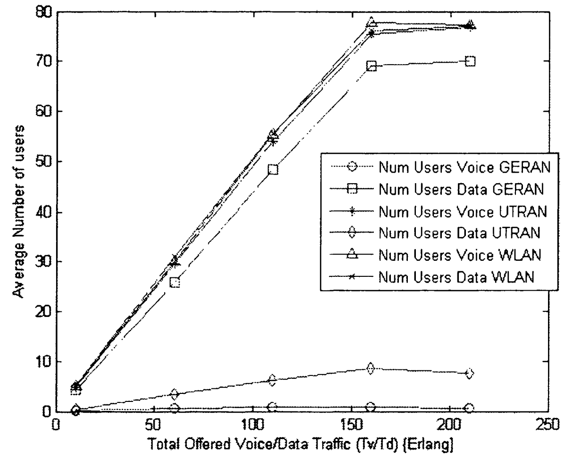


Figure 4.7: User distribution patterns for Algorithm 2

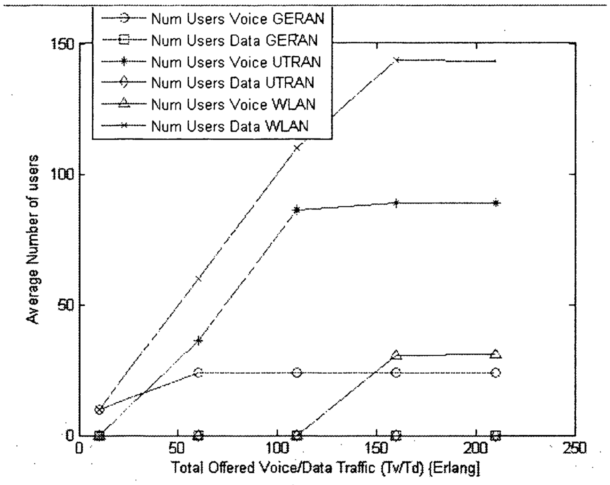


Figure 4.8: User distribution patterns for Algorithm 3

Fig. 4.9 presents the blocking probabilities of the three algorithms. From Fig. 4.9, it can be seen that in terms of blocking probability, Algorithm 3 is the worst one when the traffic load is high. For example, when the offered voice and data traffics are 180 Erlangs, the blocking probability of using Algorithm 3 is around 35% higher than using the other two algorithms. This is because in the two LB based algorithms, some data users are forced to share one channel in GERAN when the traffic load is high. However, in Algorithm 3, because voice users will be served in GERAN first and every voice user occupies one GERAN channel so that the total number of users that can be served is less than Algorithms 1 and 2.

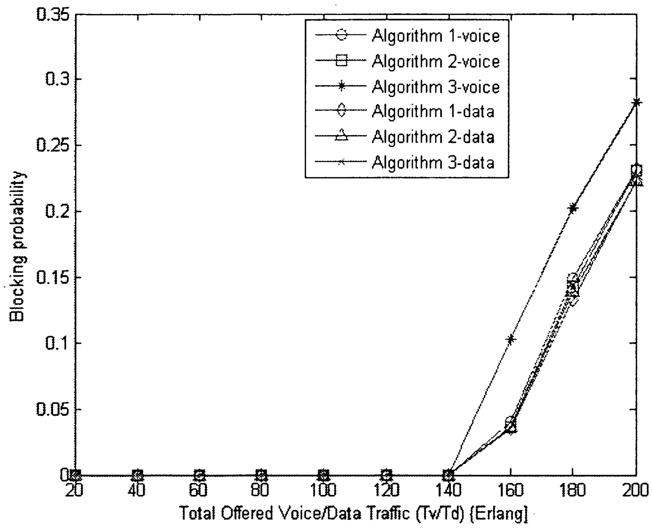


Figure 4.9: Blocking probabilities

Fig. 4.10 illustrates average data throughputs of the three algorithms. From Fig. 4.10, it can be seen that in terms of data throughput, Algorithm 1 performs better than the other two when the traffic load is low (for example, about 13% higher than Algorithm 2 and around 20% higher than Algorithm 3 when the offered voice and data traffics are 40 Erlangs). This is because in Algorithm 1, as discussed before, only a small number of voice users are served by WLAN. Therefore, more WLAN capacity is available for data users. Algorithm 3 is the worst one, because it allocates all voice users to GERAN and UTRAN, and all data users to WLAN so that no GERAN and UTRAN capacities are available for data users. When the traffic load becomes high, Algorithm 3 outperforms the other two algorithms in terms of data throughput (almost 6 times as high as Algorithms 1 and 2), because it can minimize the number of voice users served by WLAN.

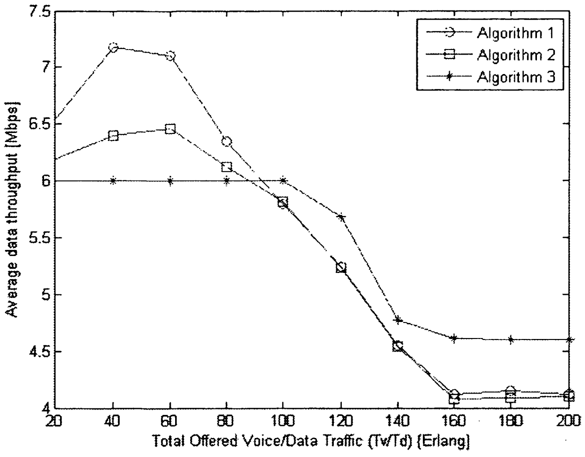


Figure 4.10: Average data throughput

Fig. 4.11 presents the throughput fairness of the three algorithms. Algorithm 3 achieves the best performance, because all data users are served in WLAN and share the same amount of resources so that the fairness index is always 1.

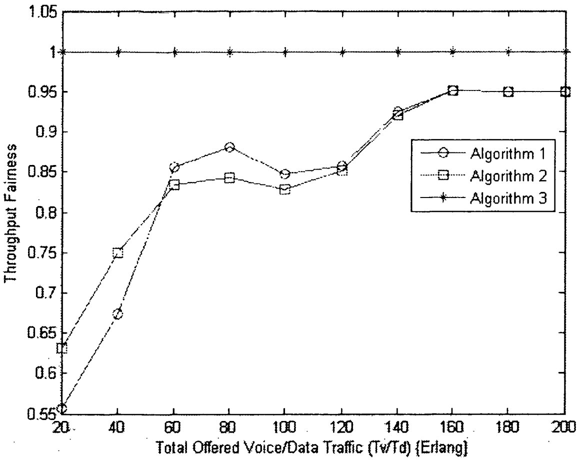


Figure 4.11: Throughput fairness

In summary, in terms of blocking probability, Algorithm 3 is the worst one when the traffic load is high. In terms of data throughput, Algorithm 1 performs better than the other two algorithms when the traffic load is low, however, Algorithm 3 outperforms the other two algorithms when the traffic load is high. In terms of throughput fairness, Algorithm 3 achieves the best performance.

4.3 Tradeoff between overall throughput and user satisfaction

The work in this section and the next section has been done in the early stage so that they focuses on a two co-located RAT scenario rather than a three co-located RAT scenario.

As discussed in Chapter 2, in a NCCB algorithm, the setting of path loss threshold is a key point. In this section, simulations are performed to find the relationship between overall downlink data throughput, user satisfaction rate, and path loss threshold.

In this simulation, the following assumptions are made: A GSM and a UMTS network are overlapped, UMTS/UTRAN model 2 is used and only data users are considered. Two simulation scenarios have been considered: cell size of $2km \times 2km$ and cell size of $4km \times 4km$. The path loss of every user is measured at every time interval during the simulation period. Both low network load (20 users served) and high network load (40 users served) cases are simulated. A user is assumed to be satisfied if the service bit rate is equal to or above a minimum accepted value R_s. The detailed parameters are summarized in Table 4.6.

Simulation results for the scenario where the cell size is $2km \times 2km$ are shown in Figs. 4.12 to 4.13.

Table 4.6: Simulation parameters

GSM parameters	
Number of carrier frequencies	3
Data user bit rate	12.2 kbps
UMTS parameters	
Activity factor	1
E_b/N_0	5dB
Downlink load factor threshold η_{DL}	0.75
WCDMA chip rate W	3.84 Mcps
Average orthogonality $\bar{\alpha}$	0.5
Other cells to own cell interference ratio f	0.65
Maximum base station transmission power	20W
Signalling channel power allocation	3W
Carrier frequency	1950MHZ
Thermal noise power	108dBm
Other parameters	
User speed	3km/hour
Base station antenna height	30m
Mobile antenna height	1.5m
R_s	32 kbps

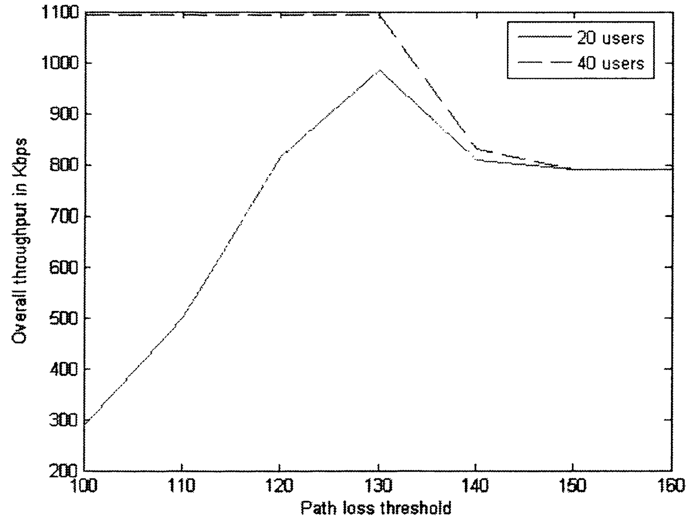


Figure 4.12: Throughput variation patterns when the cell size is $2km \times 2km$

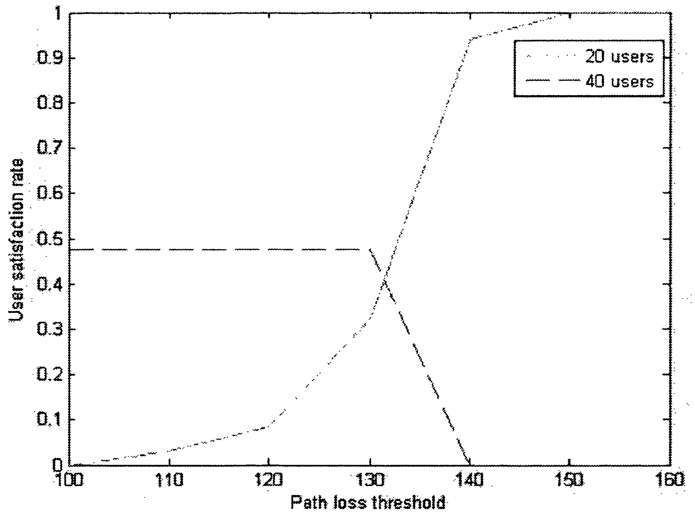


Figure 4.13: User satisfaction rate variation patterns when the cell size is $2km \times 2km$

From Figs. 4.12 to 4.13, it can be seen when the network load is low (20 users served), with the increase of path loss threshold, the overall throughput will increase to a maximum value (when the path loss threshold = 130dB) and then decrease, however, the user satisfaction rate keeps increasing. The reasons are as follows. When the path loss threshold is low, most or even all users are allocated to GSM, where only a relatively low throughput is available and the UMTS capacity is not utilized. This causes both low overall throughput and low user satisfaction rate. However, with the increase of path loss threshold from 100 dB to 130 dB, more users will be served by UMTS, where a relatively high data throughput is available and more UMTS capacity will be utilized so that both overall throughput and user satisfaction rate will increase. If the path loss threshold is above 130 dB, most of or even all users will be allocated to UMTS. The UMTS capacity will reach to its maximum value and the throughput per user in UMTS will be decreased. In this case, only a small number of users are assigned to GSM, most GSM channels will not be used, which will cause a lower overall throughput. However, because the network load is low, even though all users are allocated to UMTS; they still can get a throughput higher than R_s so that the user satisfaction rate remains increasing.

In the high network load case (40 users served), when the path loss threshold is between 100 dB and 130 dB, the overall throughput levels out. The reasons are as follows. For the GSM network, because the path loss threshold is relatively low, most users are allocated to GSM first so that all GSM channels are occupied. For the UMTS network, because of the small cell size, the UMTS throughput is determined by the load factor. So the overall throughput is a fixed value whatever the average user path loss is. If we keep increasing the path loss threshold, both overall throughput and user satisfaction rate will decrease. The reason for overall throughput decline is the same as the case where 20 users are served. The reason for user satisfaction rate decrease is that the increase of users served in UMTS causes the per user throughput less than R_s.

Figs. 4.14 to 4.15 show the simulation results when the cell size is enlarged to $4km \times 4km$. It can be seen that for the low network load case, the overall throughout and user satisfaction rate variation patterns are similar to the small cell size situation. However, the maximum throughput value occurs when the path loss threshold is 140 dB rather than 130 dB. For the high network load case, the overall throughput and user satisfaction rate variation patterns have some differences from the small cell size situation. This is because when the cell size becomes larger, the average path loss of users will in turn increase and the bottleneck of throughput will be BS transmission power rather than load factor. With the increase of path loss threshold from 100 dB to 140 dB, more high path loss users will be allocated to GSM so that the interference level in UMTS is reduced and higher throughput can then be achieved. When the path loss threshold is above 140 dB, the overall throughput will start to decrease. The reason is same as those for the small cell size situation. The user satisfaction rate will increase with the increase of path loss threshold from 120 dB to 140 dB, because more users are allocated to UMTS, where they can get a throughput higher than R_s. However, if the path loss threshold is above 140 dB, the user satisfaction rate will decrease, because there are too many users served in UMTS so that the throughput per user will be reduced.

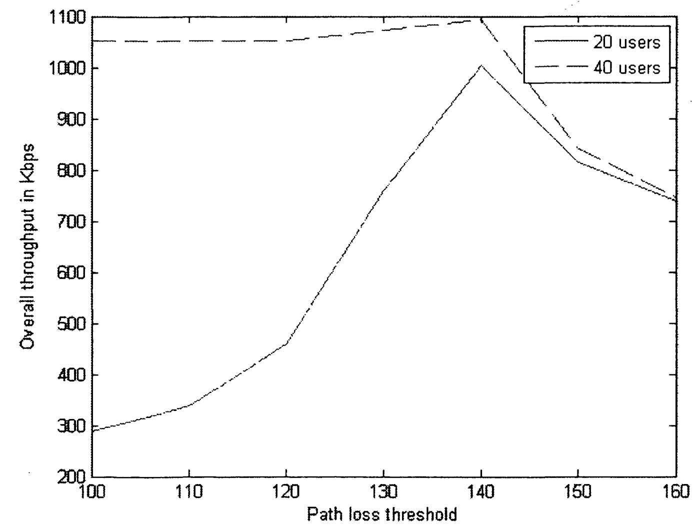


Figure 4.14: Throughput variation patterns when the cell size is $4km \times 4km$

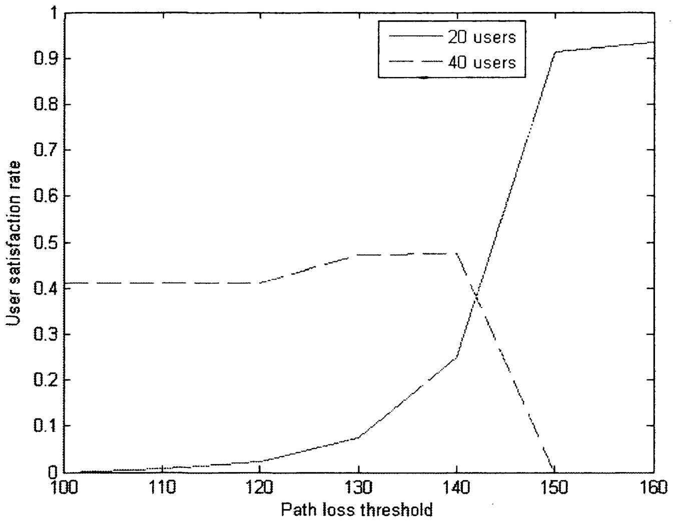


Figure 4.15: User satisfaction rate variation patterns when the cell size is $4km \times 4km$

In summary, the overall throughput will start to decrease if the path loss threshold is above a certain value, referred to as PL. The larger the cell size, the higher the value of PL (130 dB for a cell size of $2km \times 2km$ and 140 dB for a cell size of $4km \times 4km$). When the network load is low, the user satisfaction rate will keep increasing. When the network load becomes higher, the user satisfaction rate will start to decrease when the path loss threshold is above PL. When the network load is high, an optimum path loss threshold PL can be found in terms of both overall throughput and user satisfaction rate. However, when the network load is low, a tradeoff is required to balance the overall throughput and user satisfaction rate when the path loss threshold is above PL. The higher the path loss threshold is set, the lower the overall throughput but the higher the user satisfaction rate.

4.4 Performance evaluation of proposed policy based algorithms

In this section, simulations are performed to compare the performance of the two proposed policy based algorithms and the VG*VU algorithm. It is assumed that a UMTS cell and a GSM cell are overlapped in a square area of $1km \times 1km$. There are four types of users being defined: voice indoor, voice outdoor, data indoor, and data outdoor. During the simulation time, it is assumed that every voice user makes one call. The data users are continuously downloading during the whole simulation period. It is assumed that the locations of indoor users are fixed while the outdoor users are moving within the simulation area randomly. It is also assumed that the indoor users have an additional penetration loss compared to the outdoor users and voice calls have higher priority over data calls.

It is assumed that the total number of users is 200. Five scenarios are defined in this simulation (summarized in Table 4.7). Three algorithms: VG*VU, IN*VG*Load, and the Proposed Policy Based Algorithm 2 are compared in terms of average downlink data throughput. The detailed simulation parameters are shown in Table 4.8 and simulation results are shown in Figs. 4.16 to 4.20.

Table 4.7: Simulation scenarios

Scenario numbers	Voice users (%)	Data users (%)	Indoor users (%)	Outdoor users (%)
1	50	50	50	50
2	75	25	50	50
3	25	75	50	50
4	50	50	75	25
5	50	50	25	75

Table 4.8: Network and traffic parameters

GSM		
Parameters	Voice	Data
User bit rate	12.2 kbps	14.4 kbps
Number of carrier frequency	3	
UMTS		
Parameters	Voice	Data
Activity factor v_j	Uplink: 0.67, Downlink: 0.58	1
E_b/N_0	7dB	5dB
Downlink throughput	Up to 384 kbps	
Block error rate (BLER)	1%	10%
Bit rate of user R_j	12.2 kbps	Uplink: up to 64kbps, Downlink: up to 128kbps
Uplink load factor threshold	0.75	
Downlink load factor threshold	0.75	
WCDMA chip rate W	3.84 Mcps	
Average orthogonality $\bar{\alpha}$	0.5	
Other cells to own cell interference ratio f	0.65	
Maximum base station transmission power	20W	
Signalling channel power allocation	3W	
Base station antenna height	30m	
Mobile antenna height	1.5m	
Carrier frequency	1950MHZ	
Thermal noise power	-108dBm	
Other parameters		
User speed	3 km/hour	
Call duration	120s	
Penetration loss	20 dB	

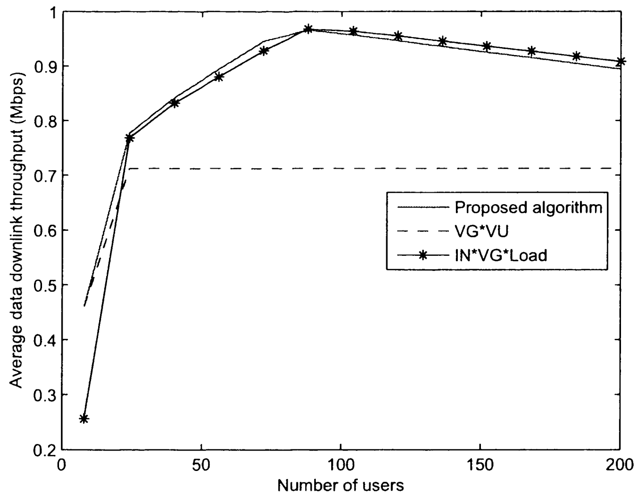


Figure 4.16: Simulation results of Scenario 1.

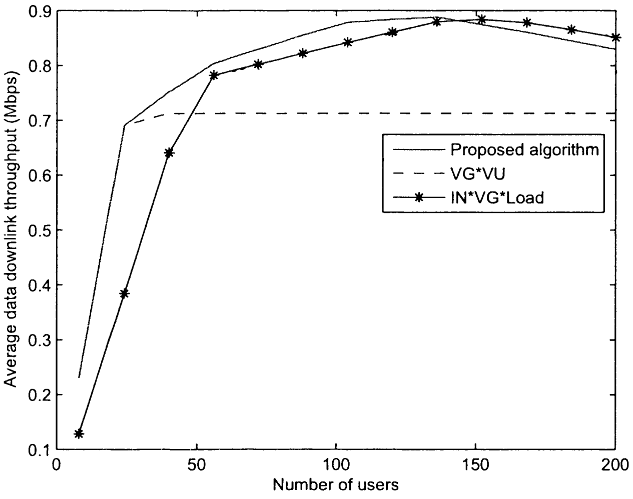


Figure 4.17: Simulation results of Scenario 2.

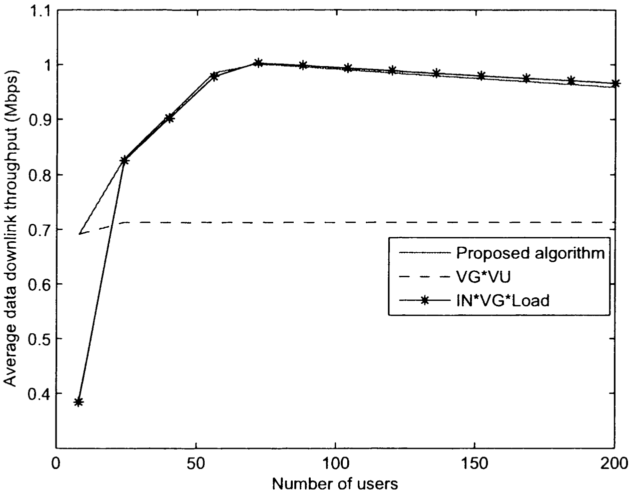


Figure 4.18: Simulation results of Scenario 3.

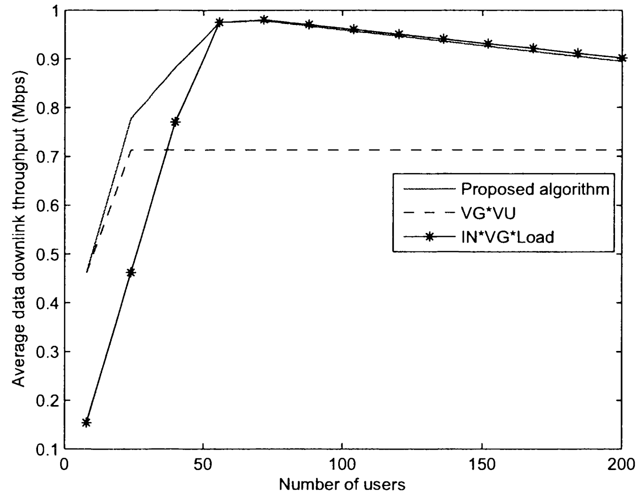


Figure 4.19: Simulation results of Scenario 4.

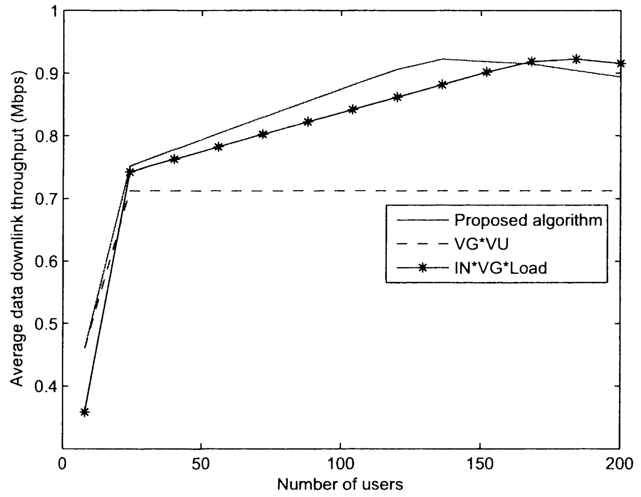


Figure 4.20: Simulation results of Scenario 5.

From Figs. 4.16 to 4.20, it can be seen that the IN*VG*Load algorithm performs worse than the other two algorithms when the system load is low in all of the five scenarios (for example, in Scenario 4, when the number of users is 8, the throughput is only about 11% of the other two algorithms). The main reasons are two-fold. Firstly, the IN*VG*Load algorithm allocates indoor data users into GSM, where only 14.4 kbps bit rate is available for data services. However, these users can get a higher bit rate in UMTS when the network load is low to medium. Another reason is that it allocates outdoor voice users to UMTS, which reduces the throughput of data users served by UMTS. However, the other two algorithms allocate all voice users to GSM and all data users to UMTS so that the data users can obtain a higher throughput.

When the system is highly loaded, the IN*VG*Load algorithm is slightly better than the other two algorithms. The main reasons are two-fold. Firstly, applying IN policy at the highest priority can minimize the number of indoor users in UMTS. Furthermore, in a highly loaded UMTS network, due to the increased number of users, the throughput per data user can be even lower than the one in GSM.

The VG*VU algorithm performs worse than the other two when the system load is medium to high. The reason is that it does not consider the IN policy at all so that a large number of indoor users are allocated to UMTS, which significantly degrades the system capacity due to higher interference levels. In VG*VU, all data users are allocated to UMTS. When the UMTS capacity reaches to its upper bound, the arrival of a new data user will decrease the throughput per user while the total throughput of UMTS will remain at the same or similar value as before. This is why the throughput performance of VG*VU levels out from medium to high system load.

The Proposed Policy Based Algorithm 2 outperforms the other two under a low to medium system load situation, because the throughput comparison function of this algorithm maximizes the overall data user throughput.

Network operators can select the most suitable solution according to system load estimation. For example, during busy hours, the IN*VG*Load algorithm can be used while the Proposed Policy Based Algorithm 2 can be used at other times.

4.5 Summary

In this chapter, a number of key performance metrics are evaluated for both existing and proposed RAT selection algorithms. Simulation results validate that the two proposed policy based RAT selection algorithms can work better than the VG*VU algorithm in terms of downlink data throughput in some cases. The results also show that there is no single algorithm that is suitable for all circumstances. Tradeoffs are always required.

Chapter 5

User level Markov models

Markov models have been applied in the RAT selection algorithms because it can provide flexible, powerful, and efficient means for a theoretical analysis of system performance metrics, such as blocking probability, traffic distribution, RAT load, and RAT throughput. In this chapter, basic knowledge of Markov models is introduced, a number of existing user level Markov models for RAT selection algorithms are discussed, and new user level Markov models are proposed.

5.1 Background knowledge of Markov models

A Markov process, named for Andrey Markov, is a family of random variables $\{X_n, n = 0, 1, 2,...\}$ with the Markov property, namely that, the next state depends only on the current state but not on the past. If $X_n = i$, the process is said to be in State i at time n. A Markov model is a particular type of Markov process, in which the process can only be in a finite or countable number of states. A one step transition probability of a Markov process from State i to State j is:

$$P_{ij} = P\{X_{(n+1)} = s_{(n+1)} | X_n = s_{(n)}, ...X_0 = s_{(0)}\}, \tag{5.1}$$

where $i = s_{(n)}$ and $j = s_{(n+1)}$. The one step transition probability has the following properties:

1) It must be a non-negative value:

$$P_{ij} \geq 0, \ i, j \geq 0, \tag{5.2}$$

2) The sum of all transition probabilities from State i to any State j is 1:

$$\sum_{j=0}^{\infty} P_{ij} = 1, \ i = 0, 1, ..., \tag{5.3}$$

3) The future state $X_{(n+1)}$ depends only on the present state X_n but not any previous states:

$$P_{ij} = P\{X_{(n+1)} = j | X_n = i\}. \tag{5.4}$$

The one step transition probability is usually summarized in a non-negative, stochastic transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} & . & . \\ P_{10} & P_{11} & P_{12} & . & . \\ P_{20} & P_{21} & P_{22} & . & . \\ . & . & . & . & \\ . & . & . & . & . \end{bmatrix}. \tag{5.5}$$

Fig. 5.1 shows a simple example, in which there are two states: State 0 and State

1. The transition probability matrix is:

$$\mathbf{P} = \begin{bmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \end{bmatrix}. \tag{5.6}$$

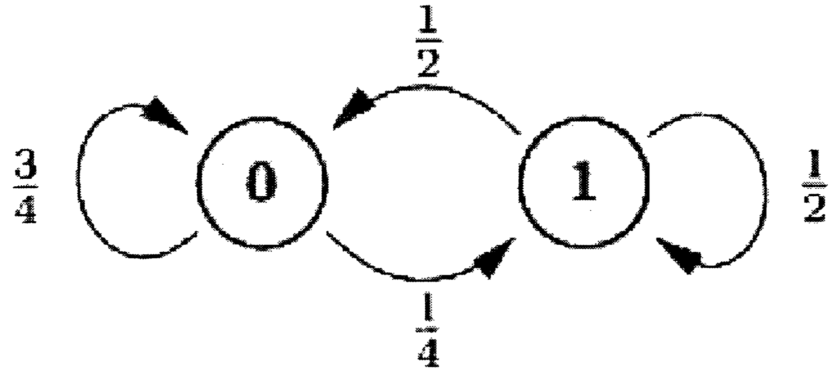


Figure 5.1: An example of Markov model

A special Markov process called birth and death process can be used to build mathematical models for RAT selection algorithms. The birth and death process assumes that transitions are only allowed between neighboring states. Fig. 5.2 shows an example of the birth and death process.

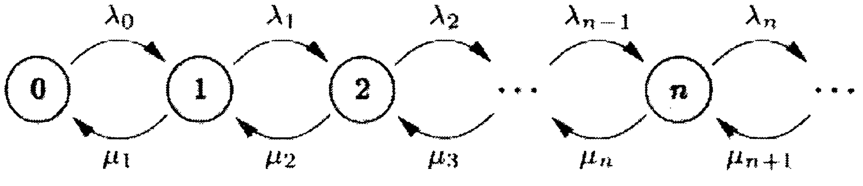


Figure 5.2: An example of the birth and death process

Transition rates λ_n, $n \geq 0$, are state-dependent arrival rates and transition rates

$\mu_{n+1},\ n \geq 0$, are state-dependent departure rates. A steady state is defined as a state, in which the sum of arrival rates is equal to the sum of departure rates. Steady state probabilities, $(\pi_n,\ n \geq 0)$ of the above Markov model can be worked out as follows.

$$\pi_0 \cdot \lambda_0 = \pi_1 \cdot \mu_1, \tag{5.7}$$

$$\pi_1 = \frac{\lambda_0}{\mu_1}\pi_0, \tag{5.8}$$

$$\pi_1(\lambda_1 + \mu_1) = \pi_0 \cdot \lambda_0 + \pi_2 \cdot \lambda_2, \tag{5.9}$$

$$\pi_2 = \frac{\lambda_0 \cdot \lambda_1}{\mu_1 \cdot \mu_2}\pi_0, \tag{5.10}$$

$$......,$$

$$\pi_k(\lambda_k + \mu_k) = \pi_{k-1} \cdot \lambda_{k-1} + \pi_{k+1} \cdot \lambda_{k+1}, \tag{5.11}$$

$$\pi_k = \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}},\ k \geq 1. \tag{5.12}$$

If $\lambda_0 = \lambda_1 = ... = \lambda_n,\ n \geq 0$, and $\mu_1 = \mu_2 = ...\mu_{n+1},\ n \geq 0$, then,

$$\pi_k = \pi_0(\frac{\lambda}{\mu})^k,\ k \geq 0. \tag{5.13}$$

From the law of total probability:

$$\sum_{k=0}^{\infty} \pi_k = 1, \tag{5.14}$$

$$\pi_0 + \pi_0 \cdot \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = 1. \tag{5.15}$$

So,

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}. \tag{5.16}$$

After working out π_0, the rest steady state probabilities can be solved using (5.12).

The above method to solve the steady state probabilities is called direct method. However, when the Markov model is complex, the steady state probabilities may be insolvable using the direct method. An iterative power method can be used to solve a complex Markov model. The iterative power method works as follows. A concept of generator matrix of a Markov model is defined as:

$$Q = [q_{ij}], \ \forall i, j \in S \tag{5.17}$$

where S is a countable infinite state space, q_{ij} is a instantaneous transition rate from State i to State j, which can be defined as [68]:

$$q_{ij}(t) = \lim_{\triangle t \to 0} \frac{P_{ij}(t, t + \triangle t)}{\triangle t}, \ i \neq j, \tag{5.18}$$

$$q_{ii}(t) = \lim_{\triangle t \to 0} \frac{P_{ii}(t, t + \triangle t) - 1}{\triangle t}. \tag{5.19}$$

At any instant time of t:

$$\sum q_{ij}(t) = 0. \tag{5.20}$$

After working out the generator matrix Q, the one step probability transition matrix can be solved by [68]:

$$\mathbf{P} = Q/\mathbf{q} + I, \tag{5.21}$$

where $\mathbf{q}$ is a uniformization factor, which is set to be a little bit larger than the maximum absolute element in Q and I is an identity matrix, which has the same size as Q. For a probability transition matrix, when n in the following equation is large enough, no matter what assumptions are made about the initial probability distribution, after a large number of steps have been taken, the probability distribution will converge to:

$$P^{(n)} = P^{(n-1)}. \tag{5.22}$$

More details of the Markov model theory can be found in [68].

5.2 User level Markov models

A user level Markov model is used to study a single user's behavior in a heterogeneous network. Through a well designed user level Markov model, we can know the probabilities of a user being in every RAT in the heterogeneous network, given the RAT selection algorithm being used, the call type, and other related information. In a user level Markov model, it is assumed that the network capacity is always sufficient to serve a user and hence blocking and dropping will not occur.

In [69], a simple user level Markov model is presented by Falowo et al. This model is designed for a co-located UMTS/WLAN network, in which a UMTS network has global coverage but a higher service cost and a WLAN network has limited coverage but a lower service cost. Three states are defined in Falowo's model:

- State 1: User is not connected to any RAT.

- State 2: User is located outside hotspot and is connected to UMTS.
- State 3: User is located inside hotspot and is connected to WLAN.

The state transition diagram of this model is shown in Fig. 5.3.

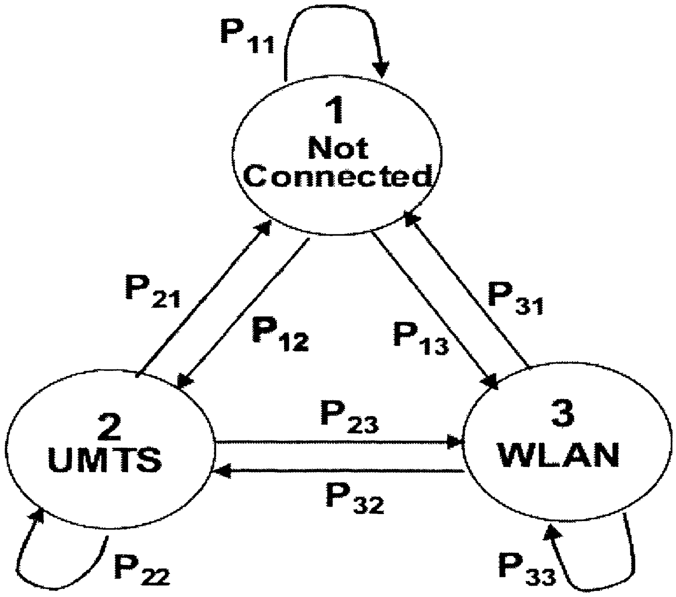


Figure 5.3: Falowo's user state transition diagram [69]

From the above state transition diagram, the state transition probability matrix can then be established:

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}. \tag{5.23}$$

In the above matrix, P_{ij} refers to the transition rate from State i to State j. The steady state probabilities can then be solved using this matrix and the details are given in [69].

Falowo's Markov model is relatively simple. In Falowo's model, if a user is in the hotspot area, where both UMTS and WLAN can be accessed, the call will be allocated to WLAN only. In [70], a more complex model is proposed by Hasib et al. In Hasib's model, the network topology is almost the same as Falowo's. The only difference is that the UMTS network is replaced by a CDMA2000 network. Five states are defined in Hasib's model:

- State 0: The user is not connected to any RAT.
- State 1: The user is located outside hotspot and is connected to CDMA2000.
- State 2: The user is located inside hotspot and is connected to CDMA2000.
- State 3: The user is located inside hotspot and is connected to both CDMA2000 and WLAN.
- State 4: The user is located inside hotspot and is connected to WLAN.

The state transition diagram is shown in Fig. 5.4.

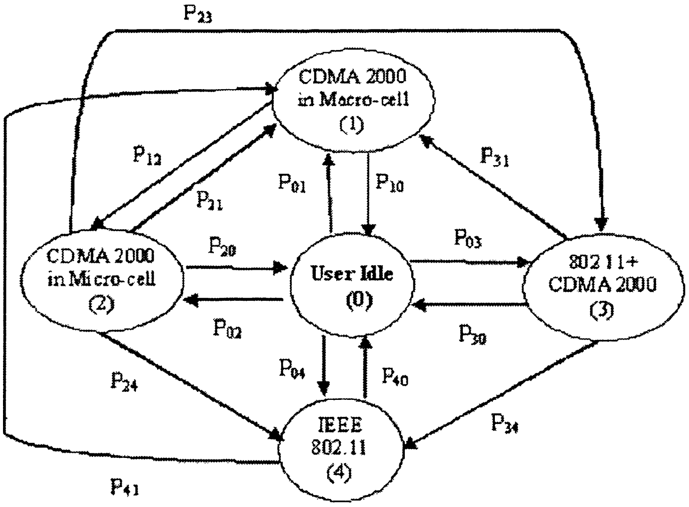


Figure 5.4: Hasib's user state transition diagram [70]

5.3 Proposed user level Markov models

In this section, new user level Markov models are proposed for a three co-located RATs network based on an extension of the two co-located RATs Markov models discussed above. The network topology is shown in Fig. 5.5. User level Markov models for two scenarios are proposed and two basic RAT selection algorithms, LB based and service based algorithms are analyzed using the proposed user level Markov models.

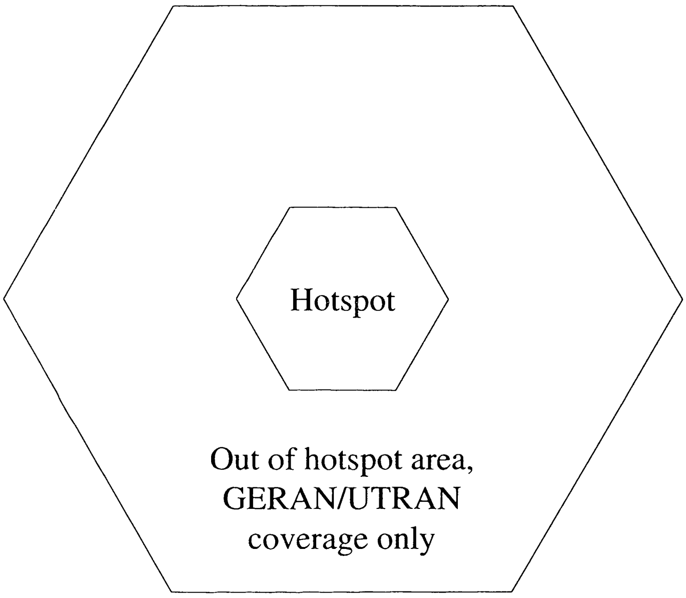


Figure 5.5: Network topology for proposed Markov models

5.3.1 Scenario 1

Scenario 1 only considers a very simple model, it is assumed that users only arrive and move within the hotspot area. Four states are defined in this model:

- State 0: The user is not connected to any RAT.

- State 1: The user is connected to GERAN
- State 2: The user is connected to UTRAN
- State 3: The user is connected to WLAN

Fig. 5.6 presents the user state transition diagram. Let P_0, P_1, P_2, and P_3 be the probabilities of a user being in States 0, 1, 2, and 3 respectively. The state transition matrix is:

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} & P_{03} \\ P_{10} & P_{11} & P_{12} & P_{13} \\ P_{20} & P_{21} & P_{22} & P_{23} \\ P_{30} & P_{31} & P_{32} & P_{33} \end{bmatrix}, \tag{5.24}$$

where P_{00} is the probability of the user staying in State 0; P_{01}, P_{02}, and P_{03} are the probabilities of a new call being allocated to GERAN, UTRAN, and WLAN respectively; P_{12}, P_{13}, P_{21}, P_{23}, P_{31}, and P_{32} are VHO probabilities; P_{11}, P_{22}, and P_{33} are probabilities that an ongoing call will stay in the RAT that is currently serving it. $0 \leq P_{ij} \leq 1$ for i,j = 0,1,2,3. According to the law of total probability:

$$\sum_{j=0}^{3} P_{ij} = 1, \; i = 0, 1, 2, 3. \tag{5.25}$$

The steady state probabilities can be solved using [68]:

$$\pi = \pi \cdot P. \tag{5.26}$$

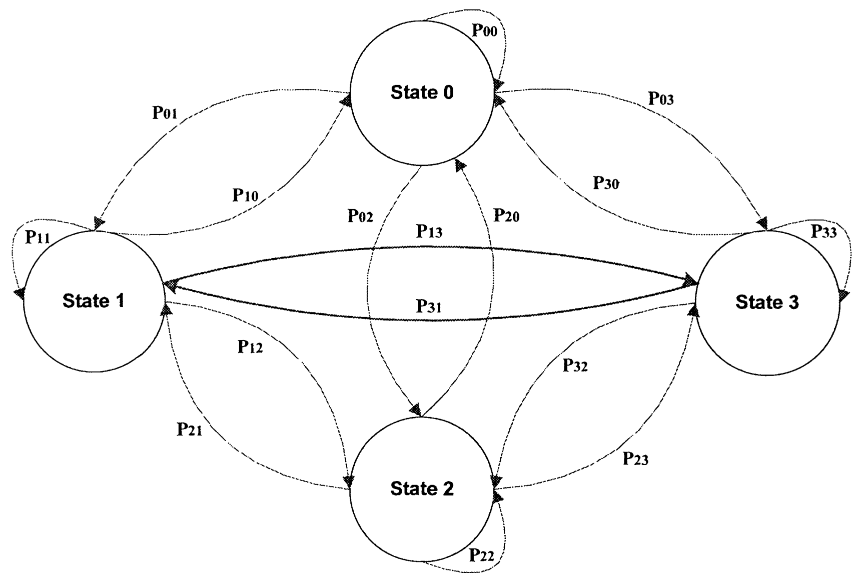


Figure 5.6: User state transition diagram of the proposed Markov model

where π is the state probability vector given by $\pi = [P_0, P_1, P_2, P_3]$. Since a user can only be in one of the four states at any point of time:

$$P_0 + P_1 + P_2 + P_3 = 1. \tag{5.27}$$

User level analysis for LB based RAT selection algorithm

In order to work out the steady state probabilities, firstly, we need to determine all state transition rates in (5.24). In a LB based RAT selection algorithm, users are allocated to the least loaded RAT. The following weighting parameter is introduced:

$$f_i = \begin{cases} 1 & \text{if } L_i = min(L_G, L_U, L_W), \\ 0 & \text{if } L_i \neq min(L_G, L_U, L_W), \end{cases} \quad i = 1, 2, 3. \tag{5.28}$$

where L_G, L_U, and L_W are the loads of GERAN, UTRAN, and WLAN respectively.

It is assumed that new calls arrive according to a Poisson process with a mean arrival rate of λ. Call duration h is exponentially distributed with a mean of $1/\mu$. The call completion probability is:

$$P_{term} = \mu. \tag{5.29}$$

The new call arriving probability is:

$$P_{new} = \lambda. \tag{5.30}$$

It is assumed that call completion probabilities are independent of the RAT in which the call is allocated:

$$P_{10} = P_{20} = P_{30} = P_{term}. \tag{5.31}$$

The probability of a user staying in the idle state is:

$$P_{00} = 1 - P_{new}. \tag{5.32}$$

The probabilities of an ongoing call to stay in it's current RAT or VHO to GERAN, UTRAN, and WLAN can be solved using:

$$P_{ij} = (1 - P_{term})f_j, \ where \ i, j = 1, 2, 3. \tag{5.33}$$

From (5.26), we can get:

$$\begin{cases} P_0 = P_{00}P_0 + P_{10}P_1 + P_{20}P_2 + P_{30}P_3 \\ P_1 = P_{01}P_0 + P_{11}P_1 + P_{21}P_2 + P_{31}P_3 \\ P_2 = P_{02}P_0 + P_{12}P_1 + P_{22}P_2 + P_{32}P_3 \\ P_3 = P_{03}P_0 + P_{13}P_1 + P_{23}P_2 + P_{33}P_3 \end{cases} \tag{5.34}$$

By using (5.27) and (5.34) , the steady state probabilities can then be worked out:

$$P_0 = P_{term}/(P_{term} + P_{new}). \tag{5.35}$$

$$P_i = (1 - P_{term} - P_{new}) \cdot f_i \cdot P_{new}/(P_{term} + P_{new}) + P_{new}f_i, \ i = 1, 2, 3. \tag{5.36}$$

User level analysis for service based RAT selection algorithm

Two types of users, voice and data users are considered. In a service based RAT selection algorithm, voice users are allocated to GERAN, UTRAN, and WLAN in order and data users are allocated in the inverse order [33]. It is assumed that a new user is randomly determined as a voice or data user and the probabilities of the user to be a voice user and a data user are defined as P_{voice} and P_{data} respectively, where

$$P_{voice} + P_{data} = 1. \tag{5.37}$$

The entire third row of the state transition matrix (shown in (5.24)) is zero because a user will not be allocated to UTRAN when GERAN and WLAN have sufficient capacities. The probabilities of a new user to be allocated to GERAN and WLAN are:

$$P_{01} = P_{new}P_{voice} = 1, \tag{5.38}$$

$$P_{03} = P_{new}P_{data} = 1. \tag{5.39}$$

In the service based algorithm, a voice user allocated to GERAN and a data user allocated to WLAN will be served by that RAT until call termination. Therefore,

$$P_{11} = P_{33} = (1 - P_{term}), \tag{5.40}$$

$$P_{22} = 0. \tag{5.41}$$

The calculations of P_{00}, P_{10}, and P_{30} are the same as that of the LB based algorithm case. According to the assumptions, the other state transition probabilities are all 0. By using (5.26) and (5.27), the steady state probabilities can then be worked out:

$$P_0 = P_{term}/(P_{term} + P_{new}), \tag{5.42}$$

$$P_1 = P_{new}P_{voice}/(P_{term} + P_{new}), \tag{5.43}$$

$$P_2 = 0, \tag{5.44}$$

$$P_3 = P_{new}P_{data}/(P_{term} + P_{new}). \tag{5.45}$$

5.3.2 Scenario 2

In Scenario 2, it is assumed that users can move out of the hotspot area. In this case, more user states need to be defined:

- State 0: The user is not connected to any RAT
- State 1: The user is within the hotspot area and connected to GERAN
- State 2: The user is within the hotspot area and connected to UTRAN

- State 3: The user is within the hotspot area and connected to WLAN
- State 4: The user is out of the hotspot area and connected to GERAN
- State 5: The user is out of the hotspot area and connected to UTRAN

Let P_0, P_1, P_2, P_3, P_4, and P_5 be the probabilities of a user being in States 0, 1, 2, 3, 4, and 5 respectively. According to the law of total probability,

$$P_0 + P_1 + P_2 + P_3 + P_4 + P_5 = 1. \tag{5.46}$$

The state transition matrix is:

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} & P_{03} & P_{04} & P_{05} \\ P_{10} & P_{11} & P_{12} & P_{13} & P_{14} & P_{15} \\ P_{20} & P_{21} & P_{22} & P_{23} & P_{24} & P_{25} \\ P_{30} & P_{31} & P_{32} & P_{33} & P_{34} & P_{35} \\ P_{40} & P_{41} & P_{42} & P_{43} & P_{44} & P_{45} \\ P_{50} & P_{51} & P_{52} & P_{53} & P_{54} & P_{55} \end{bmatrix}, \tag{5.47}$$

The following definitions are made. The probability of a user residing in the hotspot area is P_{in}. The probability of a user residing outside the hotspot area $P_{out} = 1 - P_{in}$ and the probability of a new user arriving in the hotspot area $P_{new\_h} = P_{new}P_{in}$. The probability of a new user arriving outside the hotspot area is then to be $P_{new}P_{out}$. The probability of a user exiting the hotspot area during a session is P_{ex} and the probability of a user entering the hotspot area during a session is P_{en}. By assuming that the numbers of users moving into and out of the hotspot area is the same on average, the following equilibrium equation applies:

$$P_{ex}P_{in} = P_{en}P_{out},\ where\ 0 < P_{in} < 1. \tag{5.48}$$

User level analysis for LB based RAT selection algorithm

One more parameter is introduced except for the weighting parameter f_i used in Scenario 1:

$$g_i = \begin{cases} 1 & \text{if } L_i = min(L_G, L_W), \\ 0 & \text{if } L_i \neq min(L_G, L_W), \end{cases} \quad i = 1, 2. \tag{5.49}$$

where L_G and L_W are the loads of GERAN and WLAN respectively. As in Scenario 1, the elements in the probability matrix are computed first. The call completion probabilities are:

$$P_{10} = P_{20} = P_{30} = P_{40} = P_{50} = P_{term}. \tag{5.50}$$

The state transition probabilities of a new call is:

$$P_{00} = 1 - P_{new}, \tag{5.51}$$

$$P_{0i} = P_{new\_h}f_i,\ where\ i = 1, 2, 3, \tag{5.52}$$

$$P_{04} = P_{new}P_{out}g_1, \tag{5.53}$$

$$P_{05} = P_{new}P_{out}g_2. \tag{5.54}$$

By assuming that an ongoing call will initially connect to the RAT currently serving it when the user crosses the hotspot boundary, then,

$$P_{42} = P_{43} = P_{51} = P_{53} = P_{15} = P_{24} = 0, \tag{5.55}$$

The other VHO probabilities are:

$$P_{41} = P_{52} = P_{en}(1 - P_{term}), \tag{5.56}$$

$$P_{14} = P_{ex}(1 - P_{term}), \tag{5.57}$$

$$P_{34} = P_{ex}(1 - P_{term})g_1, \tag{5.58}$$

$$P_{25} = P_{ex}(1 - P_{term}), \tag{5.59}$$

$$P_{35} = P_{ex}(1 - P_{term})g_2. \tag{5.60}$$

The probabilities of an active call staying in the hotspot area can be calculated using:

$$P_{ij} = (1 - P_{term})(1 - P_{ex})f_i, \ \ where \ i, j = 1, 2, 3. \tag{5.61}$$

The probabilities of an active call staying outside the hotspot area are:

$$P_{44} = P_{54} = (1 - P_{term})(1 - P_{en})g_1, \tag{5.62}$$

$$P_{45} = P_{55} = (1 - P_{term})(1 - P_{en})g_2, \tag{5.63}$$

By using (5.26) and (5.46), the steady state probabilities can then be worked out:

$$P_0 = P_{term}/(P_{term} + P_{new}), \tag{5.64}$$

$$P_3 = P_{new\_h}f_3P_0 + (1 - P_{term})(1 - P_{ex})f_3P_{in}, \tag{5.65}$$

$$P_4 = \frac{c_1g_1P_0 + c_2(P_{new\_h}f_1P_0 + (1 - P_{term})(1 - P_{ex})f_1P_{in}) + c_2g_1P_3 + c_3g_1P_{out}}{1 - c_2P_{en}(1 - P_{term})}, \tag{5.66}$$

$$P_5 = P_{out} - P_4, \tag{5.67}$$

$$P_1 = P_{new\_h} f_1 P_0 + (1 - P_{term})(1 - P_{ex}) f_1 P_{in} + P_{en}(1 - P_{term}) P_4, \tag{5.68}$$

$$P_2 = P_{new\_h} f_2 P_0 + (1 - P_{term})(1 - P_{ex}) f_2 P_{in} + P_{en}(1 - P_{term}) P_5, \tag{5.69}$$

$$\text{where } c_1 = 1 - P_{new\_h},$$
$$c_2 = P_{ex}(1 - P_{term}),$$
$$c_3 = (1 - P_{term})(1 - P_{en}).$$

User level analysis for service based RAT selection algorithm

For a service based RAT selection algorithm, similar to Scenario 1, the state transition probabilities can be calculated as follows:

$$P_{01} = P_{new\_h} P_{voice}, \tag{5.70}$$

$$P_{02} = 0, \tag{5.71}$$

$$P_{03} = P_{new\_h} P_{data}, \tag{5.72}$$

$$P_{04} = P_{new} P_{out} P_{voice}, \tag{5.73}$$

$$P_{05} = P_{new} P_{out} P_{data}, \tag{5.74}$$

$$P_{35} = P_{ex}(1 - P_{term}), \tag{5.75}$$

$$P_{12} = P_{13} = P_{21} = P_{22} = P_{31} = P_{32} = 0, \tag{5.76}$$

$$P_{23} = (1 - P_{term})(1 - P_{ex}), \tag{5.77}$$

$$P_{44} = P_{55} = (1 - P_{term})(1 - P_{en}), \tag{5.78}$$

$$P_{45} = P_{54} = 0. \tag{5.79}$$

The calculations of other state transition probabilities are same as the LB based algorithm case. Again, by using (5.26) and (5.46), the steady state probabilities can be solved:

$$P_0 = P_{term}/(P_{term} + P_{new}), \tag{5.80}$$

$$P_4 = [c_1 \times P_{voice}P_0 + c_2 d_1 P_0/(1 - d_2)]/[1 - c_2 d_3/(1 - d_2) - c_3], \tag{5.81}$$

$$P_5 = P_{out} - P_4, \tag{5.82}$$

$$P_2 = d_3 \times P_5, \tag{5.83}$$

$$P_3 = (P_{new\_h} \times P_{data}P_0 + d_2 P_2)/(1 - d_2), \tag{5.84}$$

$$P_1 = P_{in} - P_2 - P_3, \tag{5.85}$$

$$\begin{aligned}
\text{where } c_1 &= P_{new}P_{out}, \\
c_2 &= P_{ex}(1 - P_{term}), \\
c_3 &= (1 - P_{term})(1 - P_{en}), \\
d_1 &= P_{new\_h}P_{voice}, \\
d_2 &= (1 - P_{term})(1 - P_{ex}), \\
d_3 &= P_{en}(1 - P_{term}), \\
P_{out} &= [c_1 P_0 + c_2(1 - P_0)]/(1 + c_2 - c_3), \\
P_{in} &= 1 - P_0 - P_{out}.
\end{aligned}$$

5.4 Summary

This chapter reviews a number of user level Markov models and proposes new user level Markov models for a three overlapped RATs network (GERAN/UTRAN/WLAN). LB based and service based RAT selection algorithms have been analyzed using the proposed Markov models. The proposed Markov models are designed for two scenarios, users only staying in the hotspot area and users moving within the whole network coverage area. By using the proposed user level Markov models, the probabilities of a user being in different states can be derived mathematically so that it can be known, at which situation, a user will be allocated to which RAT, given the RAT selection algorithm being used, the call type, and other related information. Because the user level Markov models are relatively simple, simulations results are not included for validation purpose.

Chapter 6

Network level Markov models

In the last chapter, user level Markov models have been discussed. In this chapter, network level Markov models, which are used to analyze the overall system performance, are studied. In this chapter, a number of well known network level Markov models are reviewed and a new network level Markov model is proposed.

6.1 Review of network level Markov models

In [30, 70], a two-dimensional network level Markov model is proposed by Hasib et al. for a co-located CDMA2000/WLAN network (shown in Fig. 6.1). In this model, the capacities of CDMA2000 and WLAN are represented by an integer number of channels: C_1 and C_2 represent the maximum numbers of channels in CDMA2000 and WLAN respectively. A portion of CDMA2000 channels are reserved for HO calls. A predefined threshold C_{th} is set and if the number of occupied CDMA2000 channels reach C_{th}, only HO calls will be accepted. A transition from State (i,j) to State (i+1,j) or (i,j+1) represents a new call arrival; a transition from State (i,j) to State (i-1,j) or (i,j-1) rcprcsents a call complction; and a transition from State (i,j) to State (i+1,j-1) or (i-1,j+1) represents a VHO. In Hasib's model, it is assumed that new calls arrive at WLAN and CDMA2000 according to a Poisson distribution with mcan arrival rates of λ_{n_μ} and λ_{n_M} respectively. λ_h is the horizontal HO rate of CDMA2000, λ_{h_μ} and λ_{h_M} arc the VHO call arrival rates of WLAN and CDMA2000 respectively; μ is the avcrage call completion rate, v is the mean macrocell boundary crossing rate, and α is the sum of μ and v. User mobility in this model is reflected by call boundary cross rates, which are dependent on call duration and call residential time in the hotspot. Morc details of this mobility model can be found in [71].

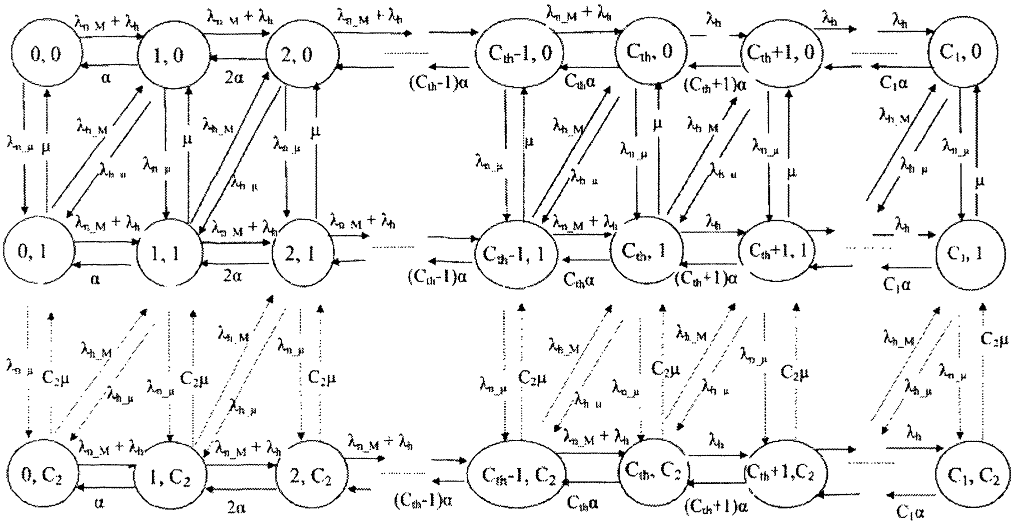


Figure 6.1: Hasib's network state transition diagram [70]

By solving the steady state probabilities of the Markov model using the iterative powcr method, a set of useful performance metrics can be expressed as follows [70]: New call blocking probability in CDMA2000, P_{nb}:

$$P_{nb} = \sum_{i=C_{th}}^{C_1} \sum_{j=C_{th}}^{C_2} \Pi_{ij}. \tag{6.1}$$

Handover call blocking probability in CDMA2000, P_{hb}:

$$P_{hb} = \sum_{j=0}^{C_2} \Pi_{C_1,j}. \tag{6.2}$$

New and handover call blocking probabilities in WLAN, P_{nb_wlan} and P_{hb_wlan}:

$$P_{nb\_wlan} = P_{hb\_wlan} = \sum_{i=0}^{C_1} \Pi_{i,C_2}. \tag{6.3}$$

Overall throughput T:

$$T = \sum_{i,j} \Pi_{ij}(iR_W + jR_L), \tag{6.4}$$

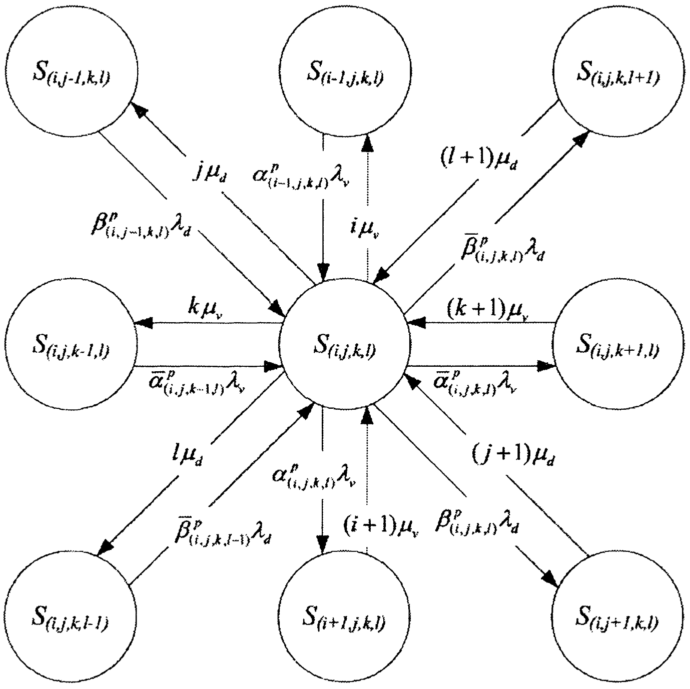
where R_W and R_L are the basic channel data rates of CDMA2000 and WLAN respectively. Detailed derivations of the above performance parameters can be found in [70].

Hasib's Markov model provides a good starting point for analytically evaluating RAT selection algorithm performance. Both VHO and user mobility aspects are considered in his model. However, service differentiation is not considered in Hasib's model. In [72], a two-dimensional network level Markov model is presented. In this model, voice and data users are differentiated, however, it is for a single RAT only. In [73, 74], a four-dimensional Markov model designed for a co-located GERAN/UTRAN network is proposed by Gelabert et al. The state transition diagram of this model is shown in Fig. 6.2. In Gelabert's model, two types of users, voice and data are considered. $S_{(i,j,k,l)}$ represents a state in which i voice users and j data users are currently served in GERAN; k voice users and l data users are currently served in UTRAN. Voice and data users consume different amounts of resources and have different priorities. In this model, GERAN load is represented by an integer number of channels.

One voice user is allocated to one GERAN channel and up to three data users can share one GERAN channel dependent on the traffic load. The UTRAN load L_U is calculated by load factors of voice and data calls [74]:

$$L_U = k[\frac{W/R_{b,v}}{(E_b/N_0)_v} + 1]^{-1} + l[\frac{W/R_{b,d}}{(E_b/N_0)_d} + 1]^{-1}, \tag{6.5}$$

where W is the WCDMA chip rate, $R_{b,v}$ is the bit rate for voice users, $R_{b,d}$ is the bit rate for data users, $(E_b/N_0)_v$ is the signal energy per bit to noise spectral density ratio for voice users, $(E_b/N_0)_d$ is the signal energy per bit to noise spectral density ratio for data users.



$S_{(i,j-1,k,l)}$
$S_{(i-1,j,k,l)}$
$S_{(i,j,k,l+1)}$
$j\mu_d$
$\alpha^p_{(i-1,j,k,l)}\lambda_v$
$(l+1)\mu_d$
$\beta^p_{(i,j-1,k,l)}\lambda_d$
$i\mu_v$
$\bar{\beta}^p_{(i,j,k,l)}\lambda_d$
$k\mu_v$
$(k+1)\mu_v$
$S_{(i,j,k-1,l)}$
$S_{(i,j,k,l)}$
$S_{(i,j,k+1,l)}$
$\bar{\alpha}^p_{(i,j,k-1,l)}\lambda_v$
$\bar{\alpha}^p_{(i,j,k,l)}\lambda_v$
$l\mu_d$
$(j+1)\mu_d$
$\alpha^p_{(i,j,k,l)}\lambda_v$
$\bar{\beta}^p_{(i,j,k,l-1)}\lambda_d$
$(i+1)\mu_v$
$\beta^p_{(i,j,k,l)}\lambda_d$
$S_{(i,j,k,l-1)}$
$S_{(i+1,j,k,l)}$
$S_{(i,j+1,k,l)}$

Figure 6.2: Gelabert's network state transition diagram [74]

As shown in Fig. 6.2, an arbitrary state $S_{(i,j,k,l)}$ can transit to one of eight other states. However, these states must be feasible states, which should meet the following requirements [74]:

$$0 \leq i/C + j/(n_c C) \leq 1, \tag{6.6}$$

where C is the number of GERAN channels and n_c is the maximum number of data users allowed to share one channel, and

$$0 \leq L_U \leq \eta_{max}, \tag{6.7}$$

where η_{max} is the load factor threshold. In Fig. 6.2, λ_v and λ_d represent the arrival rates of voice and data calls respectively; μ_v and μ_d represent the completion rates of voice and data calls respectively; and p represents the fraction of user terminals with multimode capabilities. α and β are state transition rate parameters, which are dependent on the RAT selection algorithm.

Compared to Hasib's model, in Gelabert's model, service types can be differentiated. Another advantage of Gelabert's model is that it is more generic so that it can be applied to different RAT selection algorithms by changing parameters α and β. However, Gelabert's model has some limitations too. Firstly, user mobility is not considered in Gelabert's model. Another limitation of Gelabert's model is that it only can be used for initial RAT selection but not VHO.

6.2 Proposed network level Markov models

The network level Markov models discussed above only considers a two co-located RATs scenario. In this section, a three co-located RATs scenario will be considered and the performance of two basic RAT selection algorithms, LB based and service based algorithms, are evaluated in terms of call blocking probability. The proposed three dimensional Markov model is not only an extension of the existing two co-located RATs models but is a more complex model with different state transitions. Compared to the existing network level Markov model, the proposed model considers both service differentiation and user mobility aspects.

The topology of thc heterogeneous wireless network is shown in Fig. 6.3. Three RATs are included in this model: GERAN, UTRAN, and WLAN. A GERAN cell and a UTRAN cell overlap with each other while a WLAN cell with smaller coverage area is located in the center of GERAN/UTRAN cells.

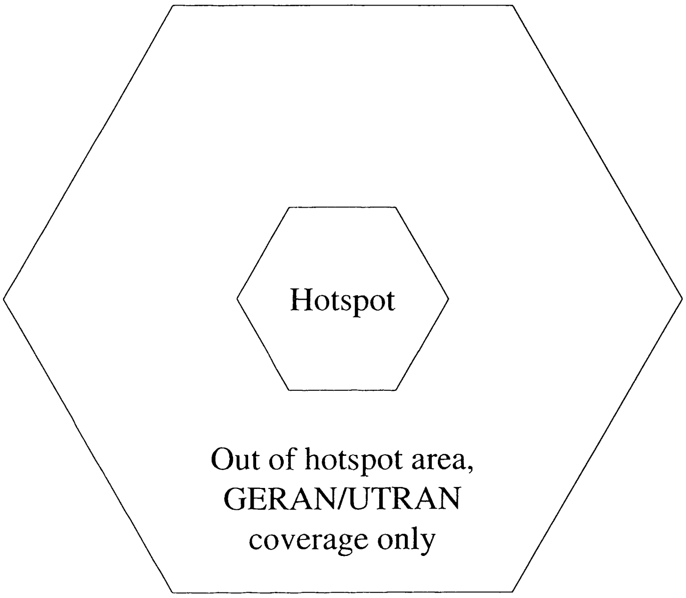


Figure 6.3: Network topology

The following assumptions are made:

1) In order to simplify the model, the RAT capacity is represented by an integer number of channels. GERAN, UTRAN, and WLAN are allocated with C_1, C_2, and C_3 basic channels, respectively.

2) Every call (i.e., real time or non real time) will be allocated to a channel in one of the three RATs.

3) New calls arrive according to a Poisson process with a mean arrival rate of λ_n. The call duration is exponentially distributed with a mean of $1/\mu$. So, the call completion rate is μ. Hence, the average offered traffic load N is λ_n/μ according to the Little's law.

4) The probability of a new call originating within the hotspot area is P_{userh}, which is independent of the type of service. The probability of a user exiting the hotspot area during a session is μ_{ex}. The probability of a user entering the hotspot area during a session is μ_{en}. By assuming that the number of users moving into and the number of users moving out of the hotspot area are the same on average, the relationship between μ_{ex} and μ_{en} is given by

$$\mu_{ex} P_{userh} = \mu_{en}(1 - P_{userh}), \quad 0 < P_{userh} < 1. \tag{6.8}$$

5) It is assumed that the VHO algorithm is coverage driven. For every new call, an initial RAT selection is carried out. VHOs will be considered when a user crosses the boundary of the hotspot area. When an active user moves into the hotspot area, a VHO from GERAN or UTRAN to WLAN may occur because the user has more RAT options. When an active user moves out of the hotspot area, a VHO from WLAN to GERAN or UTRAN may occur because WLAN does not have coverage in the out of hotspot area. VHOs between GERAN and UTRAN will not happen.

In general, one can characterize the above scenario using a three-dimensional Markov model as shown in Fig. 6.4. Let $S_{(i,j,k)}$ represents a state in which i, j, and k users are served in GERAN, UTRAN, and WLAN, respectively. The feasibility of a state depends on the maximum number of calls that can be simultaneously served in each RAT. Let S denote a set of feasible states; a state $S_{(i,j,k)} \in S$ if $0 \leq i \leq C_1$, $0 \leq j \leq C_2$, and $0 \leq k \leq C_3$.

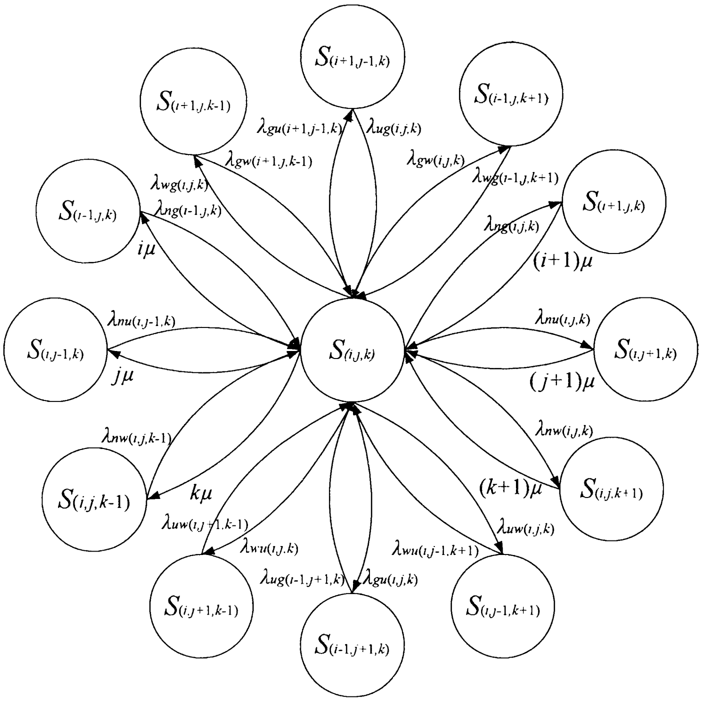


Figure 6.4: State transition diagram for the proposed three dimensional Markov model

For any feasible state $S_{(i,j,k)}$, theoretically, it can transit to one of 12 other states if they are feasible (as shown in Fig. 6.4). Transitions from state $S_{(i,j,k)}$ to states $S_{(i+1,j,k)}$, $S_{(i,j+1,k)}$, and $S_{(i,j,k+1)}$ represent a new call arrival in GERAN, UTRAN, and WLAN respectively. Transitions from state $S_{(i,j,k)}$ to states $S_{(i-1,j,k)}$, $S_{(i,j-1,k)}$ and $S_{(i,j,k-1)}$ represent a call completion in GERAN, UTRAN, and WLAN respectively. Transitions from state $S_{(i,j,k)}$ to the other six states means VHOs. For every new call, an initial RAT selection is performed. λ_{ng}, λ_{nu}, and λ_{nw} represent new call arrival rates to GERAN, UTRAN, and WLAN, respectively; Let define λ_{ug} and λ_{wg} as VHO call arrival rates from UTRAN and WLAN to GERAN, respectively; λ_{gu} and λ_{wu} as VHO call arrival rates from GERAN and WLAN to UTRAN, respectively; and finally, λ_{gw} and λ_{uw} as VHO call arrival rates from GERAN and UTRAN to WLAN, respectively.

Let define $P_{(i,j,k)}$ as the steady state probability of the system being in state $S_{(i,j,k)}$. The following steady state balance equation for any state $S_{(i,j,k)} \in S$ can be written,

$$\sum Rates\ In = \sum Rates\ Out. \tag{6.9}$$

An indicator function $\delta_{(i,j,k)}$ is defined to exclude non-feasible states in the above balance equation:

$$\delta_{(i,j,k)} = \begin{cases} 1 & \text{if } S_{(i,j,k)} \in S, \\ 0 & \text{otherwise.} \end{cases} \tag{6.10}$$

From the state transition diagram shown in Fig. 6.4, the balance equation for a state

$S_{(i,j,k)} \in S$ can be derived as

$$\begin{aligned}
&P_{(i,j,k)}[\lambda_{ng(i,j,k)}\delta_{(i+1,j,k)} + \lambda_{nu(i,j,k)}\delta_{(i,j+1,k)} \\
&+ \lambda_{nw(i,j,k)}\delta_{(i,j,k+1)} + i\mu\delta_{(i-1,j,k)} + j\mu\delta_{(i,j-1,k)} \\
&+ k\mu\delta_{(i,j,k-1)} + \lambda_{gu(i,j,k)}\delta_{(i-1,j+1,k)} + \lambda_{gw(i,j,k)}\delta_{(i-1,j,k+1)} \\
&+ \lambda_{ug(i,j,k)}\delta_{(i+1,j-1,k)} + \lambda_{uw(i,j,k)}\delta_{(i,j-1,k+1)} \\
&+ \lambda_{wg(i,j,k)}\delta_{(i+1,j,k-1)} + \lambda_{wu(i,j,k)}\delta_{(i,j+1,k-1)}] \\
&= \mu(i+1)P_{(i+1,j,k)}\delta_{(i+1,j,k)} + \mu(j+1)P_{(i,j+1,k)}\delta_{(i,j+1,k)} \\
&+ \mu(k+1)P_{(i,j,k+1)}\delta_{(i,j,k+1)} \\
&+ \lambda_{ng(i-1,j,k)}P_{(i-1,j,k)}\delta_{(i-1,j,k)} \\
&+ \lambda_{nu(i,j-1,k)}P_{(i,j-1,k)}\delta_{(i,j-1,k)} \\
&+ \lambda_{nw(i,j,k-1)}P_{(i,j,k-1)}\delta_{(i,j,k-1)} \\
&+ \lambda_{ug(i-1,j+1,k)}P_{(i-1,j+1,k)}\delta_{(i-1,j+1,k)} \\
&+ \lambda_{wg(i-1,j,k+1)}P_{(i-1,j,k+1)}\delta_{(i-1,j,k+1)} \\
&+ \lambda_{gu(i+1,j-1,k)}P_{(i+1,j-1,k)}\delta_{(i+1,j-1,k)} \\
&+ \lambda_{wu(i,j-1,k+1)}P_{(i,j-1,k+1)}\delta_{(i,j-1,k+1)} \\
&+ \lambda_{gw(i+1,j,k-1)}P_{(i+1,j,k-1)}\delta_{(i+1,j,k-1)} \\
&+ \lambda_{uw(i,j+1,k-1)}P_{(i,j+1,k-1)}\delta_{(i,j+1,k-1)}.
\end{aligned} \tag{6.11}$$

The steady state probabilities can then be solved using (6.8) and the following constraint equation:

$$\sum_{i=0}^{C_1}\sum_{j=0}^{C_2}\sum_{k=0}^{C_3} P_{(i,j,k)} = 1. \tag{6.12}$$

By knowing the steady state probabilities, some useful performance metrics can be achieved, which will be discussed later in Section 6.2.3.

6.2.1 LB based RAT selection algorithm

Let L_G, L_U, and L_W be defined as the loads of GERAN, UTRAN, and WLAN after accepting a new call when the system is currently in state $S_{(i,j,k)} \in S$ respectively,

$$L_G = \frac{i+1}{C_1}, \tag{6.13}$$

$$L_U = \frac{j+1}{C_2}, \tag{6.14}$$

$$L_W = \frac{k+1}{C_3}. \tag{6.15}$$

A user will not be accepted by a RAT if its load exceeds 1. In order to determine which RAT should be selected for a user, the following load weighting parameters is introduced:

$$a_{1(i,j,k)} = \begin{cases} 1 & \text{if } L_G < min(L_U, L_W) \\ 1/2 & \text{if } L_G = min(L_U, L_W) \ \& \ L_U \neq L_W \\ 1/3 & \text{if } L_G = L_U = L_W \\ 0 & \text{if } L_G \neq min(L_G, L_U, L_W), \end{cases} \tag{6.16}$$

$$a_{2(i,j,k)} = \begin{cases} 1 & \text{if } L_G < L_U \\ 1/2 & \text{if } L_G = L_U \\ 0 & \text{if } L_G > L_U, \end{cases} \tag{6.17}$$

$$b_{1(i,j,k)} = \begin{cases} 1 & \text{if } L_U < min(L_G, L_W) \\ 1/2 & \text{if } L_U = min(L_G, L_W) \ \& \ L_G \neq L_W \\ 1/3 & \text{if } L_G = L_U = L_W \\ 0 & \text{if } L_U \neq min(L_G, L_U, L_W), \end{cases} \tag{6.18}$$

$$b_{2(i,j,k)} = \begin{cases} 1 & \text{if } L_U < L_G \\ 1/2 & \text{if } L_G = L_U \\ 0 & \text{if } L_U > L_G, \end{cases} \tag{6.19}$$

$$c_{(i,j,k)} = \begin{cases} 1 & \text{if } L_W < min(L_G, L_U) \\ 1/2 & \text{if } L_W = min(L_G, L_U) \ \& \ L_G \neq L_U \\ 1/3 & \text{if } L_G = L_U = L_W \\ 0 & \text{if } L_W \neq min(L_G, L_U, L_W), \end{cases} \tag{6.20}$$

where $a_{1(i,j,k)}$, $b_{1(i,j,k)}$, and $c_{(i,j,k)}$ are parameters used to determine the probabilities of a call to be allocated to GERAN, UTRAN, and WLAN respectively, if it is located inside the hotspot area. $a_{1(i,j,k)} + b_{1(i,j,k)} + c_{(i,j,k)} = 1$. Parameters $a_{2(i,j,k)}$ and $b_{2(i,j,k)}$ are used to determine the probabilities of a call to be allocated to GERAN and UTRAN respectively, when it is located outside the hotspot area. $a_{2(i,j,k)} + b_{2(i,j,k)} = 1$. The new call arrival rates can then be calculated as follows:

$$\lambda_{ng(i,j,k)} = \lambda_n P_{userh} a_{1(i,j,k)} + \lambda_n (1 - P_{userh}) a_{2(i,j,k)}, \tag{6.21}$$

$$\lambda_{nu(i,j,k)} = \lambda_n P_{userh} b_{1(i,j,k)} + \lambda_n (1 - P_{userh}) b_{2(i,j,k)}, \tag{6.22}$$

$$\lambda_{nw(i,j,k)} = \lambda_n P_{userh} c_{(i,j,k)}, \tag{6.23}$$

It should be noticed that:

$$\lambda_n = \lambda_{ng} + \lambda_{nu} + \lambda_{nw}. \tag{6.24}$$

The VHO call arrival rates are:

$$\lambda_{wg(i,j,k)} = kP_{ex} a_{2(i,j,k)}, \tag{6.25}$$

$$\lambda_{wu(i,j,k)} = kP_{ex}b_{2(i,j,k)}, \tag{6.26}$$

$$\lambda_{gw(i,j,k)} = i(1 - P_{userh})P_{en}c_{(i,j,k)}, \tag{6.27}$$

$$\lambda_{uw(i,j,k)} = j(1 - P_{userh})P_{en}c_{(i,j,k)}, \tag{6.28}$$

$$\lambda_{gu(i,j,k)} = \lambda_{ug(i,j,k)} = 0. \tag{6.29}$$

6.2.2 Service based RAT selection algorithm

A service based RAT selection algorithm is proposed in [33]. In this algorithm, RT users are allocated in the following order: 1) GERAN, 2) UTRAN, 3) WLAN and NRT users are allocated in the inverse order: 1) WLAN, 2) UTRAN, 3) GERAN. Let us define P_{rt} as the probability of a RT call arrival and P_{nrt} as the probability of a NRT call arrival. $P_{rt} + P_{nrt} = 1$.

In order to determine the RAT that will be selected for a user, the following weighting parameters are used:

$$d_{1(i,j,k)} = \begin{cases} 1 & \text{if } L_G \leq 1 \\ 0 & \text{if else,} \end{cases} \tag{6.30}$$

$$d_{2(i,j,k)} = \begin{cases} 1 & \text{if } L_W > 1 \;\&\; L_U > 1 \;\&\; L_G \leq 1 \\ 0 & \text{if else,} \end{cases} \tag{6.31}$$

$$d_{3(i,j,k)} = \begin{cases} 1 & \text{if } L_U > 1 \;\&\; L_G \leq 1 \\ 0 & \text{if else,} \end{cases} \tag{6.32}$$

$$e_{1(i,j,k)} = \begin{cases} 1 & \text{if } L_G > 1 \;\&\; L_U \leq 1 \\ 0 & \text{if else,} \end{cases} \tag{6.33}$$

$$e_{2(i,j,k)} = \begin{cases} 1 & \text{if } L_W > 1 \text{ \& } L_U \leq 1 \\ 0 & \text{if else,} \end{cases} \tag{6.34}$$

$$e_{3(i,j,k)} = \begin{cases} 1 & \text{if } L_U \leq 1 \\ 0 & \text{if else,} \end{cases} \tag{6.35}$$

$$f_{1(i,j,k)} = \begin{cases} 1 & \text{if } L_G > 1 \text{ \& } L_U > 1 \text{ \& } L_W \leq 1 \\ 0 & \text{if else,} \end{cases} \tag{6.36}$$

$$f_{2(i,j,k)} = \begin{cases} 1 & \text{if } L_W \leq 1 \\ 0 & \text{if else,} \end{cases} \tag{6.37}$$

where $d_{1(i,j,k)}$ and $e_{1(i,j,k)}$ are parameters used to determine the probabilities of a RT call to be allocated to GERAN and UTRAN respectively. Parameters $d_{2(i,j,k)}$ and $e_{2(i,j,k)}$ are used to determine the probabilities of a NRT call to be allocated to GERAN and UTRAN respectively when it is located inside the hotspot area. Parameters $d_{3(i,j,k)}$ and $e_{3(i,j,k)}$ are used to determine the probabilities of a NRT call to be allocated to GERAN and UTRAN respectively when it is located outside the hotspot area. Parameters $f_{1(i,j,k)}$ and $f_{2(i,j,k)}$ are used to determine the probabilities of a RT call and a NRT call to be allocated to WLAN respectively. New call arrival rates can then be calculated as follows:

$$\begin{aligned} \lambda_{ng(i,j,k)} &= \lambda_n(P_{rt}d_{1(i,j,k)} + P_{userh}P_{nrt}d_{2(i,j,k)} \\ &+ (1 - P_{userh})P_{nrt}d_{3(i,j,k)}), \end{aligned} \tag{6.38}$$

$$\begin{aligned} \lambda_{nu(i,j,k)} &= \lambda_n(P_{rt}e_{1(i,j,k)} + P_{userh}P_{nrt}e_{2(i,j,k)} \\ &+ (1 - P_{userh})P_{nrt}e_{3(i,j,k)}), \end{aligned} \tag{6.39}$$

$$\lambda_{nw(i,j,k)} = \lambda_n P_{userh}(P_{rt} f_{1(i,j,k)} + P_{nrt} f_{2(i,j,k)}), \tag{6.40}$$

The VHO call arrival rates are:

$$\lambda_{wg(i,j,k)} = k P_{ex}(P_{rt} d_{1(i,j,k)} + P_{nrt} d_{3(i,j,k)}), \tag{6.41}$$

$$\lambda_{wu(i,j,k)} = k P_{ex}(P_{rt} e_{1(i,j,k)} + P_{nrt} e_{3(i,j,k)}), \tag{6.42}$$

$$\lambda_{gw(i,j,k)} = i(1 - P_{userh}) P_{en} f_{2(i,j,k)}, \tag{6.43}$$

$$\lambda_{uw(i,j,k)} = j(1 - P_{userh}) P_{en} f_{2(i,j,k)}, \tag{6.44}$$

$$\lambda_{gu(i,j,k)} = \lambda_{ug(i,j,k)} = 0. \tag{6.45}$$

6.2.3 Numerical Results

In this section, Markov chain validation and performance comparison between LB based and service based RAT selection algorithms have been conducted. An iterative power method is used to solve the steady state probabilities. After working out the steady state probabilities, the average carried traffic (average number of users) for every RAT can be derived as follows:

$$N_G = \sum_{\forall (i,j,k) | S_{(i,j,k)} \in S} i P_{(i,j,k)}, \tag{6.46}$$

$$N_U = \sum_{\forall (i,j,k) | S_{(i,j,k)} \in S} j P_{(i,j,k)}, \tag{6.47}$$

$$N_W = \sum_{\forall (i,j,k) | S_{(i,j,k)} \in S} k P_{(i,j,k)}, \tag{6.48}$$

where N_G, N_U, and N_W are the average traffic carried in GERAN, UTRAN, and WLAN respectively. The call blocking probability, including new call and VHO call blocking probabilities, can be computed by:

$$B = 1 - \frac{N_G + N_U + N_W}{N}, \tag{6.49}$$

Table 6.1 summarizes network and traffic parameters. A parameter will be set to its default value if it is not a variable.

Table 6.1: Network and traffic parameters

Number of channels	3 for GERAN, 3 for UTRAN and 9 for WLAN
Default average traffic load N	10 Erlangs
Default value of P_{userh}	0.5
Default value of P_{ex}	0.5
Default value of P_{rt}	0.5

Markov model validation

In order to validate the proposed Markov chain model, numerical and simulation results are compared. Simulation results are obtained by using the LB and service based algorithms discussed before. The simulation tool is MATLAB. Figs. 6.5 and 6.6 present the call blocking probability patterns with the increase of average traffic load for LB based and service based algorithms respectively. It can be seen that the analytical and simulated results match very closely, which validates the proposed Markov model.

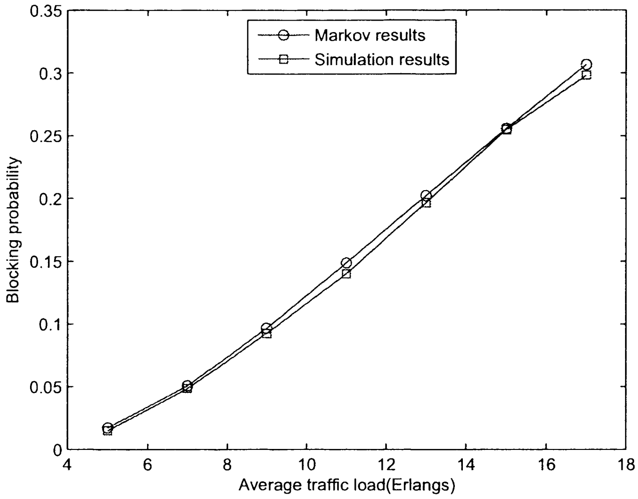


Figure 6.5: Call blocking probability for LB based RAT selection algorithm

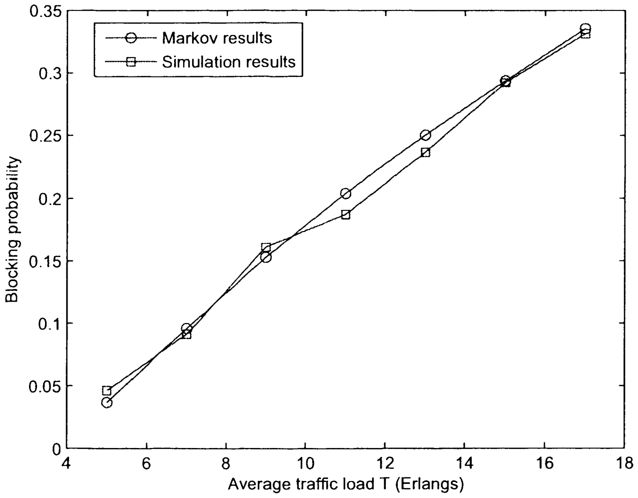


Figure 6.6: Call blocking probability for service based RAT selection algorithm

Performance comparison

Figs. 6.7 to 6.9 compare the performance of LB based and service based RAT selection algorithms in terms of call blocking probability using the proposed Markov model.

These figures present the call blocking probability patterns under varying traffic loads, probabilities of users arriving within the hotspot area, and probabilities of a call being real time respectively. The results present that in terms of call blocking probability, LB based algorithm performs better than service based algorithm, especially in the case of high P_{rt}. Fig. 6.9 proves that the variation of P_{rt} has no influence on LB based algorithm, because it does not consider user service types when making RAT selection decisions. However, for the service based RAT selection algorithm, with the increase of P_{rt}, the call blocking probability will increase. This is because RT users within the hotspot are allocated to GERAN and UTRAN first, which reduces the capacity available for users out of the hotspot area.

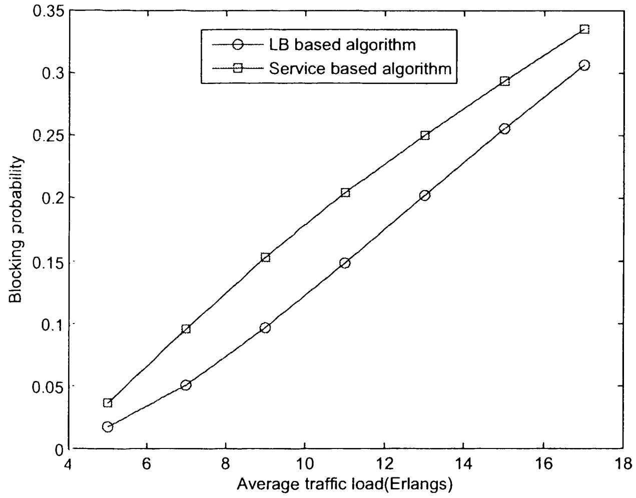


Figure 6.7: Call blocking probability comparison between LB based and service based algorithms under varying traffic loads

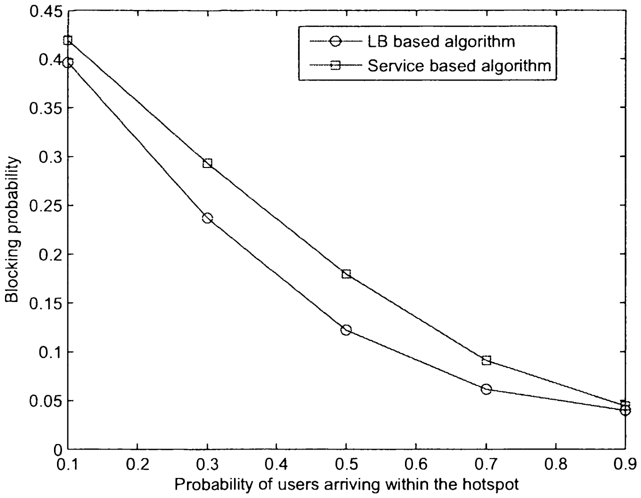


Figure 6.8: Call blocking probability comparison between LB based and service based algorithms under varying probabilities of a user arriving within the hotspot area

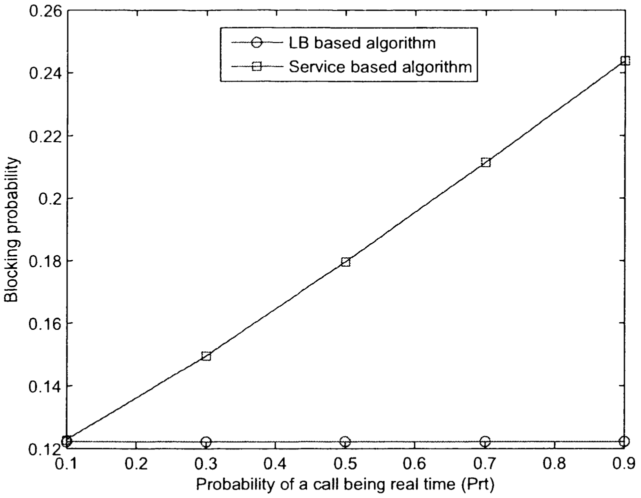


Figure 6.9: Call blocking probability comparison between LB based and service based algorithms under varying probabilities of a call being real time

In this work, the proposed Markov model is used to evaluate the load balancing based and service based algorithms. However, this model can be used to evaluate any algorithms proposed in the future by changing the transfer rates between different states. Compared to numerical simulations, the Markov model approach has its advantage. The accuracy of a numerical simulation is depended on the number of samples. More samples in the simulation, more accurate results can be obtained. However, longer simulation time will be caused. If the simulation model is simple, it may not be a serious problem; however, if the simulation model is very complex, the simulation time may be too long to wait. A proper theoretical model can solve this problem by obtain accurate results in an acceptable time constraint.

6.3 Summary

In this Chapter, existing network level Markov models for RAT selection algorithms are reviewed and a new three-dimensional network level Markov model for the performance evaluation of RAT selection algorithms in a co-located GERAN/UTRAN/WLAN network is proposed. Compared to the existing network level Markov model, the proposed model considers both service differentiation and mobility issues. Numerical results obtained from this model are validated by simulation results. The LB and service based RAT selection algorithms have been compared using the proposed Markov model. Numerical results show that the LB based algorithm outperforms the service based algorithm in terms of call blocking probability. In general, a more realistic Markov model requires more dimensions. However, it may be computationally insolvable. A tradeoff between reality and computational solvability has to be made before a Markov model is designed.

Chapter 7

Conclusions and future work

This chapter summaries the thesis contributions and discusses potential directions for future research in the RAT selection area.

7.1 Summary of contributions

The research work in this thesis focuses on the RAT selection part of the CRRM area. This thesis contains the following contributions.

1. This thesis evaluates the effects of load threshold setting on the performance of LB based RAT selection algorithm for real time traffic. It is found that setting a proper load threshold may achieve a more balanced load distribution among overlapped RATs. However, it also may cause higher DR/VHO probability and in turn higher signaling overhead and blocking/dropping probability. Tradeoffs need to be conducted before making decisions.

2. This thesis evaluates the performance of three RAT selection algorithms, LB based using maximum resource consumption, LB based using minimum resource con-

sumption, and service based algorithms, for a co-located GERAN/UTRAN/WLAN network. Simulation results show that in terms of blocking probability, the service based algorithm is the worst algorithm when the traffic load is high. In terms of data throughput, the LB based using maximum resource consumption algorithm performs better than the other two algorithms when the traffic load is low, however, the service based algorithm outperforms the other two algorithms when the traffic load is high. In terms of throughput fairness, the service based algorithm achieves the best performance.

3. This thesis studies the relationship between overall downlink data throughput, user satisfaction rate, and path loss threshold in the NCCB algorithm. It is found that the overall throughput will start to decrease if the path loss threshold above a certain value, referred to as PL. The larger the cell size, the higher the value of PL (130 dB for a cell size of $2km \times 2km$ and 140 dB for a cell size of $4km \times 4km$). When the network load is low, the user satisfaction rate will keep increasing. When the network load becomes higher, the user satisfaction rate will start to decrease when the path loss threshold is above PL. When the network load is high, an optimum path loss threshold PL can be found in terms of both overall throughput and user satisfaction rate. However, when the network load is low, a tradeoff is required to balance the overall throughput and user satisfaction rate when the path loss threshold is above PL. The higher the path loss threshold is set, the lower the overall throughput but the higher the user satisfaction rate.

4. This thesis proposes two improved policy based RAT selection algorithms and compares them with the VG*VU algorithm proposed in the literature. It was found that the Proposed Policy Based Algorithm 2 outperforms the other two algorithms in a low to medium system load case while the proposed IN*VG*Load algorithm is the best choice for highly loaded networks. Network operators can select the most

suitable solution according to system load estimation. For example, during busy hours, the proposed IN*VG*Load algorithm can be used while the Proposed Policy Based Algorithm 2 can be used at other times.

5. This thesis proposes new user level Markov models for a three co-located RATs network (GERAN/UTRAN/WLAN). LB and service based RAT selection algorithms have been mathematically analyzed using the proposed Markov model. The proposed Markov models are designed for two scenarios, users only staying in the hotspot area and users moving within the whole network coverage area. By using the proposed user level Markov model, it can be known which RAT a user will be allocated to, given related information, such as the environment, RAT selection algorithm, call type, etc.

6. This thesis proposes a new three-dimensional network level Markov model for performance evaluation of RAT selection algorithms in a co-located GERAN/UTRA-N/WLAN network. Compared to the existing network level Markov models, the proposed model considers both service differentiation and mobility issues. Numerical results obtained from the proposed model are validated by simulation results. The LB and service based RAT selection algorithms have been compared using the proposed Markov model. Numerical results show that the LB based algorithm outperforms the service based algorithm in terms of call blocking probability.

7.2 Future work

In this thesis, the performance of a number of RAT selection algorithms has been evaluated in terms of load threshold, fairness, throughput, blocking and dropping probabilities. In the future work, more performance metrics, such as packet delay, cost, operator's revenue should be evaluated and the performance of more RAT selection algorithms should be compared.

In this thesis, two improved policy based RAT selection algorithms have been proposed. However, the proposed algorithms are only for a co-located UMTS/GSM network. In the future work, a more generic algorithm should be proposed. The generic algorithm is expected to be applied to more types of RATs and is expected to make RAT selection decisions by considering both technical factors, such as user, service, and network properties and non-technical factors, such as cost and revenue.

In this thesis, simulation models are developed for specific RATs, such as GERAN, UTRAN, and WLAN. In the future work, a more generic simulation model should be developed. The generic simulation model is expected to be able to work for any type of RAT and be more realistic.

In this thesis, only voice and data users are simulated. In the future work, more types of users, such as video users, should be simulated.

In this thesis, improved user level and network level Markov models for a three co-located RAT scenario are proposed. However, in the proposed Markov models, the RAT capacity is represented by an integer number of channels, which is not realistic. In the future work, more realistic Markov models should be developed.

In this thesis, only blocking probability is evaluated using the proposed network level Markov model. In the future work, more performance metrics should be evaluated.

In this thesis, only two and three co-located RAT scenarios have been considered. In the future work, four or even more co-located RAT scenario will be considered.

Bibliography

[1] S. Hasan, N. Siddique, S. Chakraborty, "Femtocell versus WiFi - A survey and comparison of architecture and performance," *The 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009 (Wireless VITAE 2009)*, Aalborg, May 2009, pp. 916-920.

[2] J. Schiller, "Mobile Communications," *Boston: Addison-Wesley.*, 2nd. Ed, 2003.

[3] P. Nicopolitidis, M. Obaidat, G. Papadimitriou, and A. Pomportsis, "Wireless Networks," *Chichester: John Wiley & Sons, Ltd.*, 2003.

[4] A. Tolli, P. Hakalin, and H. Holma, "Performance Evaluation of Common Radio Resource Management (CRRM)," *IEEE International Conference on Communications 2002*, New York, USA, May 2002, pp. 3429-3433.

[5] N. Passas, S. Paskalis, A. Kaloxylos, F. Bader, R. Narcisi, E. Tsontsis, A. S. Jahan, H. Aghvami, M. O'Droma, and I. Ganchev, "Enabling technologies for the 'always best connected' concept," *Wireless Communications & Mobile Computing*, Vol. 6, pp. 523-540, Apr. 2006.

[6] J. Pérez-Romero, O. Sallent, R. Agustí, P. Karlssont, A. Barbaresit, L. Wang, F. Casadevall, M. Dohler, H. Gonzfilezt, and F. Cabral-Pintot, "Common Radio

Resource Management: Functional Models and Implementation Requirements," *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, Germany Sep. 2005, pp. 2067-2071.

[7] J. Pérez-Romero, O. Sallent, R. Agustí, and M. A. Diaz-Guerra, "Radio Resource Management Strategies in UMTS," *John Wiley & Sons.*, 2nd. Ed, 2005.

[8] J. Pérez-Romero, O. Sallent, and R. Agustí, "On Evaluating Beyond 3G Radio Access Networks: Architectures, Approaches and Tools," *IEEE 61st Vehicular Technology Conference*, Stockholm, Sweden, May-Jun., 2005, pp. 2964-2968.

[9] 3GPP TR v5.0.0, "Improvement of RRM across RNS and RNS/BSS (Release 5)," 2001.

[10] F. Casadevall, P. Karlsson, O. Sallent, H. Gonzalez, A. Barbaresi, and M. Dohler, "Overview of the EVEREST Project," *13th IST Mobile & Wireless Communications Summit*, Lion, France, 2004.

[11] 3GPP TR v0.3.0, "Improvement of RRM across RNS and RNS/BSS (Post Rel-5) (Release 6)," 2003.

[12] Y. Cui, Y. Xue, H. Shang, X. Sha, and Z. Ding, "A Novel Scheme and Access Architecture for Joint Radio Resource Management in Heterogeneous Networks," *International Forum on Information Technology and Applications, 2009 (IFITA'09)*, Chengdu, China, May 2009, pp. 24-27.

[13] P. Magnusson, J. Lundsj, J. Sachs, and P. Wallentin, "Radio Resource Management Distribution in a Beyond 3G Multi-Radio Access Architecture," *Global Telecommunication Conference*, 2004, pp. 3472-3477.

[14] O. Sallent, "A Perspective on Radio Resource Management in B3G," *3rd International Symposium on Wireless Communication Systems (CCECE'06)*, Valencia, Sept. 2006, pp. 30-34.

[15] L. Wu and K. Sandrasegaran, "A Survey on Common Radio Resource Management," *The Second Australia Conference on Wireless Broadband and Ultra Wideband Communications (Auswireless'07)*, Sydney, Australia, Aug. 2007 pp. 66.

[16] O. E. Falowo and H. A. Chan, "Joint call admission control algorithms: Requirements, approaches, and design considerations," *Computer Communications*, Vol. 31, pp. 1200-1217, Nov. 2007.

[17] A. Serrador and L. Correia, "A Cost Function for Heterogeneous Networks Performance Evaluation Based on Different Perspectives," *The 16th ISTMobile and Wireless Communications Summit*, Budapest, 2007, pp. 1-5.

[18] A. Serrador and L. Correia, "Policies for a Cost Function for Heterogeneous Networks Performance Evaluation," *The IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2007)*, Athens, Greece, Sep. 2007, pp. 1-5.

[19] G. Cybenko, "Dynamic load balancing for distributed memory multiprocessors," *Journal of Parallel and Distributed Computing*, Vol. 7, pp. 279-301, Oct. 1989.

[20] T. Chu and S. Rappaport, "Overlapping coverage with reuse partitioning in cellular communication systems," *IEEE Transactions on Vehicular Technology, 2006 (CCECE '06)*, Vol. 46, pp. 41-54, Feb. 2006.

[21] B. Eklundh, "Channel Utilization and Blocking Probability in a Cellular Mobile Telephone System with Directed Retry," *IEEE Transactions on Communications*, Vol. 34, No. 4, Apr. 1986.

[22] K. Suleiman, H. Chan, M. Dlodlo, "Load balancing in the call admission control of heterogeneous wireless networks," *International Conference On Communications And Mobile Computing, 2006 (IWCMC'06)*, Vancouver, Canada, Jul. 2006, pp. 245-250.

[23] A. Tolli and P. Hakalin, "Overlapping coverage with reuse partitioning in cellular communication systems," *IEEE 56th Vehicular Technology Conference*, 2002, pp. 1691-1695.

[24] Y. Zhang, K. Zhang, Y. Ji, and P. Zhang, "Adaptive Threshold Joint Load Control in an End-to-end Reconfigurable System," *15th IST Mobile & Wireless Communications Summit*, Mykonos, 2006.

[25] R. Piqueras, J. Pérez-Romero, O. Sallent, and R. Agusti, "Dynamic Pricing for Decentralised RAT Selection in Heterogeneous Scenarios," *IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications*, Helsinki, Sep. 2006, pp. 1-5.

[26] Z. Ding, Y. Xu, X. Sha, Y. Cui, "A Novel JRRM Approach Working Together with DSA in the Heterogeneous Networks," *International Forum on Information Technology and Applications, 2009. (IFITA'09)*, Chengdu, May 2009, pp. 347-350.

[27] J. Luo, R. Mukerjee, M. Dillinger, E. Mohyeldin, and E. Schulz, "Investigation of radio resource scheduling in WLANs coupled with 3G cellular network," *IEEE Communications Magazine*, Vol. 41, Jun. 2003, pp. 108-115.

[28] R. B. Ali, S. Pierre, "An efficient predictive admission control policy for heterogenous wireless bandwidth allocation in next generation mobile networks," *International conference on Wireless communications and mobile computing (IWCMC'06)*, Vancouver, Canada Jul. 2006, pp. 635-640.

[29] O. Yilmaz, A. Furuskar, J. Pettersson, and A. Simonsson, "Access selection in WCDMA and WLAN multi-access networks," *The IEEE 61stVehicular Technology Conference (VTC 2005-Spring)*, May-Jun. 2005, pp. 2220-2224.

[30] A. Hasib and A. Fapojuwo, "Performance Analysis of Common Radio Resource Management Scheme in Multi-service Heterogeneous Wireless Networks," *IEEE Wireless Communications and Networking Conference (WCNC) 2007*, Hong Kong, Mar. 2007, pp. 3296-3300.

[31] I. Koo, A. Furuskar, J. Zander, K. Kim, "Erlang capacity of multiaccess systems with service- based access selection," *IEEE Communications Letters*, Vol. 8, No. 11, pp. 662-664, 2004.

[32] W. Song, H. Jiang, W. Zhuang, X. Shen, "Resource management for QoS support in WLAN/cellular interworking," *IEEE Network Magazine*, Vol. 19, No. 5, pp. 12-18, 2005.

[33] A. Baraev, L. Jorguseski, and R. Litjens, "Performance Evaluation of Radio Access Selection Procedures in Multi-Radio Access Systems," *WPMC'05*, Aalborg, Denmark, 2005.

[34] B. Abuhaija and K. Al-Begain, "Enhanced Common Radio Resources Managements Algorithm in Heterogeneous Cellular Networks," *The 3rd International Conference on Next Generation Mobile Applications, Services and Technologies, 2009 (NGMAST'09)*, Cardiff, Wales, Sep. 2009, pp. 335-342.

[35] J. Pérez-Romero, O. Sallent, R. Agustí, N. Garcia, L. Wang, and H. Aghvami, "Network-controlled cell-breathing for capacity improvement in heterogeneous CDMA/TDMA scenarios," *Wireless Communications and Networking Conference 2006*, Las Vegas, Apr. 2006, pp. 36-41.

[36] J. Pérez-Romero, O. Sallent, R. Agustí, "A Novel Algorithm for Radio Access Technology, Selection in Heterogeneous B3G networks," *The IEEE 63rd Vehicular Technology Conference (VTC 2006-Spring)*, Melbourne, May 2006, pp. 471-475.

[37] J. Pérez-Romero, R. Ferrus, O. Salient, and J. Olmos, "RAT Selection in 3GPP-Based Cellular Heterogeneous Networks: From Theory to Practical Implementation," *The IEEE Wireless Communications and Networking Conference 2007*, Kowloon, Mar. 2007, pp. 2115-2120.

[38] L. Wang, H. Aghvami, N. Nafisi, O. Sallent, and J. Pérez-Romero, "Voice Capacity with Coverage-based CRRM in a Heterogeneous UMTS/GSM Environment," *The Second International Conference on Communications and Networking in China (CHINACOM '07)*, Shanghai, China, Aug. 2007, pp. 1085-1089

[39] J. Delicado and J. Gozalvez, "User Satisfaction Based CRRM policy for heterogeneous wireless networks," *The 6th International Symposium on Wireless Communication Systems, 2009 (ISWCS 2009)*, Tuscany, Sep. 2009, pp. 176-180.

[40] J. Gozalvez and J. Delicado, "CRRM strategies for improving user QoS in multimedia heterogeneous wireless networks," *The IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, 2009* , Tokyo, Sep. 2009, pp. 2250-2254.

[41] J. Pérez-Romero, O. Sallent, and R. Agustí, "Policy-based Initial RAT Selection algorithms in Heterogeneous Networks," *7th Mobile Wireless Communication Networks (MWCN)*, Marrakech, Morocco, Mar. 2005, pp. 1-5.

[42] W. Zhang, "Performance of real-time and data traffic in heterogeneous overlay wireless networks," *The 19th International Teletraffic Congress (ITC 19)*, Beijing, Aug.-Sep. 2005.

[43] J. Pérez-Romero, O. Sallent, and R. Agustí, "On the Capacity Degradation in W-CDMA Uplink/Downlink Due to Indoor Traffic," *Vehicular Technology Conference*, Los Angeles, USA, Sep. 2004.

[44] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "On the suitability of Load Balancing Principles in Heterogeneous Wireless Access Networks," *Wireless Personal Multimedia Communications Symposium (WPMC'05)*, Denmark, Sep. 2005.

[45] J. Pérez-Romero, O. Sallent, and R. Agustí, "Network Controlled Cell Breathing in Multi-Service Heterogeneous CDMA/TDMA Scenarios," *IEEE 64th Vehicular Technology Conference (VTC-2006)*, Canada, Sep. 2006, pp. 1-5.

[46] J. Pérez-Romero, O. Sallent, and R. Agustí, "A Novel Metric for Context-Aware RAT Selection in Wireless Multi-Access Systems," *The IEEE International Conference on Communications, 2007(ICC'07)*, Glasgow, Jun. 2007, pp. 5622-5627.

[47] O. Sallent, J. Pérez-Romero, R. Ljung, P. Karlsson, and A. Barbaresi, "Operator's RAT Selection Policies Based on the Fittingness Factor Concept," *The 16th IST Mobile and Wireless Communications Summit*, Budapest, Jul. 2007, pp. 1-5.

[48] J. Pérez-Romero, O. Sallent, A. Umbert, A. Barbaresi, R. Ljung, and R. Azevedo, "RAT Selection in Wireless Multi-Access Systems," *The First Ambient Networks Workshop on Mobility, Multiaccess, and Network Management (M2NM-2007)*, Sydney, Australia, 2007.

[49] J. Pérez-Romero, O. Salient, and R. Agustí, "A Generalized Framework for Multi-RAT Scenarios Characterisation," *The IEEE 65thVehicular Technology Conference (VTC2007-Spring)*, Dublin, Apr. 2007, pp. 980-984.

[50] X. Liu, V. Li, and P. Zhan, "Joint Radio Resource Management through Vertical Handoffs in 4G Networks," *IEEE Global Telecommunications Conference (GLOBECOM '06)*, Carlifornia, USA, Nov.-Dec. 2006, pp. 1-5.

[51] A. Umbert, L. Budzisz, N. Vucevic, and F. Bernardo, "An all-IP heterogeneous wireless testbed for RAT selection and e2e QoS evaluation," *The 2007 International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST '07)*, Cardiff, UK, Sep. 2007, pp. 310-315.

[52] A. Pillekeit, F. Derakhshan, E. Jugl, and A. Mitschele-Thiel, "Force-based load balancing in co-located UMTS/GSM networks," *IEEE 60th Vehicular Technology Conference*, Sep. 2004, pp. 4402-4406.

[53] F. Yu and V. Krishnamurthy, "Optimal joint session admission control in integrated CDMA WLAN cellular networks with vertical handoff," *IEEE Transactions on Mobile Computing*, Vol. 6, No. 1, pp. 42-51, Dec. 2007.

[54] A. Hasib and A. O. Fapojuwo, "A QoS Negotiation Framework for Heterogeneous Wireless Networks," *Canadian Conference on Electrical and Computer Engineering (CCECE 2007)*, Vancouver, BC, Apr. 2007, pp. 769-772.

[55] R. Agustí, O. Salient, J. Pérez-Romero, and L. Giupponi, "A fuzzy-neural based approach for joint radio resource management in a beyond 3G framework," *First International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE)*, 2004, pp. 216-224.

[56] L. Giupponi, R. Agustí, J. Pérez-Romero, and O. Sallent, "A novel joint radio resource management approach with reinforcement learning mechanisms," *24th IEEE International Performance, Computing, and Communications Conference (IPCCC)*, Apr. 2005, pp. 621-626.

[57] C. Lin and C. Lee, "Neural-Network-Based Fuzzy Logic Control and Decision System," *IEEE Transactions on Computers*, Vol. 40, No. 12, pp. 1320-1336, Dec. 1991.

[58] K. Lo, C. Shung, "A Neural Fuzzy Resource Manager for Hierarchical Cellular Systems Supporting Multimedia Services," *IEEE Transactions on Vehicular Technology*, Vol. 52, No. 5, pp. 1196-1206, Sep. 2003.

[59] P. Chan, R. Sheriff, Y. Hu, P. Conforto, and C. Tocci, "Mobility management incorporating fuzzy logic for a heterogeneous IP environment," *IEEE Communications Magazine*, Vol. 39, No. 12, pp. 42-51, Dec. 2001.

[60] Q. Guo, X. Xu, J. Zhu, H. Zhang, "A QoS-guaranteed cell selection strategy for heterogeneous cellular systems," *ETRI Journal*, Vol. 28, No. 1, pp. 77-83, Feb. 2001.

[61] W. Zhang, "Handover decision using fuzzy MADM in heterogeneous networks," *IEEE Wireless Communications and Networking Conference, 2004 (WCNC 2004)*, Atlanta, GA, Mar. 2004.

[62] M. Alkhawlani and A. Hussein, "Intelligent radio network selection for next generation networks," *The 7th International Conference on Informatics and Systems (INFOS 2010)*, Cairo, Egypt, 28-30 March 2010.

[63] M. Alkhawlani and A. Hussein, "Radio network selection for tight-coupled wireless networks," *The 7th International Conference on Informatics and Systems (INFOS 2010)*, Cairo, Egypt, 28-30 March 2010.

[64] J. Laiho, A. Wacker, and T. Novosad, "Radio Network Planning and Optimisation for UMTS," *John Wiley & Sons.*, 2001.

[65] H. Holma and A. Toskala, "WCDMA for UMTS:radio access for third generation mobile communications," *John Wiley & Sons.*, 3rd. Ed, 2004.

[66] K. Sipila, Z. Honkasalo, J. Laiho-Steffens, and A. Wacker, "Estimation of Capacity and Required Transmission Power of WCDMA Downlink Based on a Downlink Pole Equation," *IEEE 51st Vehicular Technology Conference*, Tokyo, Japa, 2000.

[67] R. Jain, A. Durresi, and G. Babic, "Throughput fairness index: an explanation," *ATM Forum/99-0045*, Feb. 1999.

[68] G. Bolch, S. Greiner, H. Meer, and K. S. Trivedi, Queueing networks and Markov chains: modeling and performance evaluation with computer science applications, 2nd ed., John Wiley and Sons, Inc., Hoboken, New Jersey, 2006.

[69] O. E. Falowo and H. A. Chan, "Joint Call Admission Control for Next Generation Wireless Network," *Canadian Conference on Electrical and Computer Engineering, 2006 (CCECE '06)*, Ottawa, Ont, 2006, pp. 1151-1154.

[70] A. Hasib and A. O. Fapojuwo, "Analysis of Common Radio Resource Management Scheme for End-to-End QoS Support in Multiservice Heterogeneous Wireless

Networks," *IEEE Transactions on Vehicular Technology*, Vol. 57, pp. 2426-2439, Jul. 2008.

[71] A. Hasib and A. O. Fapojuwo, "A mobility model for Heterogeneous wireless networks," *IEEE Radio and Wireless Symposium 2008*, Orlando, FL, Jan. 2008 pp. 815-818.

[72] S. J. Lincke, "Balancing CRRM Performance Goals with Load Shared Packet Services," *IEEE Wireless Communications and Networking Conference, 2008 (WCNC 2008)*, Las Vegas, NV, Mar.-Apr. 2008, pp. 2969-2974

[73] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A 4-Dimensional Markov Model for the Evaluation of Radio Access Technology Selection Strategies in Multi-service Scenarios," *IEEE 64th Vehicular Technology Conference, 2006 (VTC-2006 Fall)*, Montreal, Que, Sep. 2006, pp. 1-5.

[74] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A Markovian Approach to Radio Access Technology Selection in Heterogeneous Multiaccess Multiservice Wireless Networks," *IEEE Transactions on Mobile Computing*, Vol. 7, No. 10, pp. 1257-1270, Oct. 2008.