An intelligent radio access network selection and optimisation system in heterogeneous communication environments

Luo, Weizhi

An Intelligent Radio Access Network Selection and Optimisation System in Heterogeneous Communication Environments

Weizhi Luo

Submitted for the degree of Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary, University of London

Abstract

The overlapping of the different wireless network technologies creates heterogeneous communication environments. Future mobile communication system considers the technological and operational services of heterogeneous communication environments. Based on its packet switched core, the access to future mobile communication system will not be restricted to the mobile cellular networks but may be via other wireless or even wired technologies. Such universal access can enable service convergence, joint resource management, and adaptive quality of service. However, in order to realise the universal access, there are still many pending challenges to solve. One of them is the selection of the most appropriate radio access network.

Previous work on the network selection has concentrated on serving the requesting user, but the existing users and the consumption of the network resources were not the main focus. Such network selection decision might only be able to benefit a limited number of users while the satisfaction levels of some users are compromised, and the network resources might be consumed in an ineffective way. Solutions are needed to handle the radio access network selection in a manner that both of the satisfaction levels of all users and the network resource consumption are considered.

This thesis proposes an intelligent radio access network selection and optimisation system. The work in this thesis includes the proposal of an architecture for the radio access network selection and optimisation system and the creation of novel adaptive algorithms that are employed by the network selection system. The proposed algorithms solve the limitations of previous work and adaptively optimise network resource consumption and implement different policies to cope with different scenarios, network conditions, and aims of operators. Furthermore, this thesis also presents novel network resource availability evaluation models. The proposed models study the physical principles of the considered radio access network and avoid employing assumptions which are too stringent abstractions of real network scenarios. They enable the implementation of call level simulations for the comparison and evaluation of the performance of the network selection and optimisation algorithms.

Acknowledgement

I would like to express my appreciation to Doctor Eliane L. Bodanese for her supervision, knowledge, constant and invaluable support and encouragement she gave me during my PhD at Queen Mary, University of London. I also would like to thank Chris Philips, Yue Chen, John Schormans, and Raul Mondragon for their help and valuable advice during my study.

Many thanks go to all the staff of the School of Electronic Engineering and Computer Science who create the welcoming environment and give me so much help during my stay.

I would like to thank the colleagues and friends I met at the school, specially, Jiayi, Oliver, Hengzhi, Peng, Ling, Vindya, Kejing, Frank and Dapeng for the good times, friendship and help.

Finally, with my love and gratitude, this thesis is dedicated to my family, as well as my girlfriend Jin. Without their persistent support and encouragement, I cannot imagine I am able to complete my study.

To my family and to Jin

Contents

List of Figures

List of Tables

Glossary

AAA	Authentication, Authorisation, and Accounting
AHP	Analytic Hierarchy Process
ANG	Access Network Gateway
ANR	Access Network Router
AP	Access Point
API	Application Programming Interface
BAN	Basic Access Network
BAS	Basic Access Signalling
BS	Base Station
CBR	Constant Bit Rate
CCN	Common Core Network
CC/PP	Composite Capabilities/Preference Profile
CoopRRM	Cooperative Radio Resource Management entity
CRE	Composite Radio Environment
CS	Circuit-Switched
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
DCF	Distributed Coordination Function
eNodeB	evolved Node B
ePDG	evolved Packet Data Gateway
GAN	Generic Access Network
GANC	Generic Access Network Controller
GERAN	GSM/EDGE Radio Access Network
GGSN	Gateway GPRS Service Node
GPS	Global Positioning System
GR	Gateway Routers
GRA	Grey Relational Analysis
GRC	Grey Relational Coefficient
GUI	Graphical User Interface
HSPA	High Speed Packet Access
IMS	IP Multimedia Subsystem
LTE	Long Term Evolution

MADM	Multiple Attribute Decision Making
MICS	Media Independent Command Service
MIES	Media Independent Event Service
MIH	Media Independent Handover
MIHF	MIH Function
MIRAI	Multimedia Integrated network by Radio Access Innovation
MIS	Multiple Interface Selection
MISSA	Multiple Interface Selection and Service Adaptation
MME	Mobility Management Entity
MS-CRE	Management System for Composite Radio Environment
MUSE-VDA	Multiservice Vertical Handoff Decision Algorithm
NGMN	Next Generation Mobile Networks
OFDMA	Orthogonal Frequency Division Multiple Access
OUSI	Overall User Satisfaction Improvement
P-RASO	Policy-based Radio Access Selection and Optimisation
PCF	Point Coordination Function
PLMN	Public Land Mobile Network
PS	Packet-Switched
RAN	Radio Access Network
RAO	Radio Access Optimisation
RAT	Radio Access Technology
RPS	Resource Prediction System
RRM	Radio Resource Management
SAE	System Architecture Evolution
SAW	Simple Additive Weighting
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDP	Session Description Protocol
SGML	Standard Generalized Markup Language
SGSN	Serving GPRS Service Node
SIP	Session Initiation Protocol
SM	Session Manager
SN	Signalling Node
SNMP	Simple Network Management Protocol
SRRM	Specific Radio Resource Management entity

TCP	Transmission Control Protocol
TMS	Terminal Management System
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
UAC	User Agent Client
UAS	User Agent Server
UE	User Equipment
UMA	Universal Mobile Access
UNI	User Number Increase
UPE	User Plane Entity
URI	Uniform Resource Indicator
VoIP	Voice over Internet Protocol
WINNER	Wireless World Initiative New Radio
WLAN	Wireless Local Area Network
WRAN	WINNER Radio Access Network
XML	Extensible Markup Language

Chapter 1 Introduction

1.1 Introduction

Wireless communications have attracted many efforts in research and this trend is continuing. Nowadays, multiple wireless network technologies are available as commercial wireless systems. They can be classified into four main categories. The first category is the mobile cellular network technologies, including the Second-Generation (2G) cellular networks (e.g. GSM and cdmaOne), the 2G evolution (e.g. GPRS, EDGE, and CDMA2000 1xRTT), and the Third-Generation (3G) cellular networks technologies (e.g. UMTS, CDMA2000 1xEV-DO). The second category comprises the Wireless Local Area Network (WLAN) technologies, including HiperLAN and IEEE 802.11 based WiFi. The third category includes the Wireless Metropolitan Area Network (WMAN) technologies, such as Digital Video Broadcasting (DVB) and IEEE 802.16-2004 based WiMAX. The last category is the Wireless Personal Area Network (WPAN), such as Bluetooth. Among these technologies, the mobile cellular networks initially focused on offering telephony applications. The 2G mobile cellular networks only provide packet services with very low data rates. The 2G evolution and 3G mobile cellular networks offer packet services and provide wider application sets. Also, the WLAN, WMAN and WPAN technologies concentrate on packet services and support a variety of applications.

The overlapping of different wireless network technologies creates heterogeneous communication environments. Future mobile communication systems, such as 3GPP *Long Term Evolution* (LTE) [LTE08], consider the technological and operational services of heterogeneous communication environments. Future mobile communication systems will have a packet switched core. The access to the mobile communication systems will not be restricted to the mobile cellular networks but may be via other wireless or even wired technologies. This foreseen universal access facilitates service convergence, joint resource management and adaptive quality of service [LTE08]. Network operators or service providers would not have to reject the user requests, but may redirect them to an appropriate network. The redirection might be through a different radio access network (RAN). The network operators or service providers can benefit from the redirection for appropriate network utilisation and optimal resource management. The users also can obtain optimal radio access which will consider their requirements and status. One of the challenges in realising the advantages of future mobile communication systems is the development of novel and efficient RAN selection and optimisation systems.

In homogeneous communication environments, the Radio Access Technology (RAT) is the same and the criterion for RAN selection is primarily based on the received signal strength. However, in heterogeneous communication environments, the measurement of the signal strength can only be done by thoroughly scanning the available access points of different RATs. This scanning can be inefficient and power consuming. Furthermore, considering the variety of the service types and the heterogeneity of the RATs, the RAN selection should be context aware and take into account other influencing factors, such as the knowledge of service type, user/terminal status, network condition, etc. Based on the above context information and a predefined criterion, the performance of the candidate RANs would be investigated and compared, and the most appropriate RAN would be selected. The above context information not only contributes to the RAN selection, but also assists reconfigurable devices (e.g. software-defined radio) to appropriately adapt to user requirements and network conditions [WSG06][Tro06]. In order to exchange the service request and transmit the context information, the RAN selection and optimisation systems need to implement a specific signalling mechanism, which should be flexible and network independent [CRZO04].

The use of the context information not only facilitates an intelligent and optimal RAN selection but also increases complexity in the selection procedure by introducing more influencing factors. Furthermore, due to user handovers and the initiation and termination of service sessions, network resource availability may vary frequently, even during the lifetime of an active service session. Such variability may affect QoS guarantees and user admission, and make the selection procedure more dynamic and complex. Therefore, in the RAN selection procedure, adaptations of service sessions and network resource utilisation would be necessary. The RAN selection and optimisation systems should be efficient and adaptive and also able to cope with different scenarios and network conditions by implementing different policies in a dynamic way.

1.2 Contribution

This work presents an in-depth study of the integration schemes for network interworking and cooperation in heterogeneous communication environments, and the challenges in the implementation of context information for the development of context-aware RAN selection algorithms. This work also presents a comprehensive review of the RAN selection algorithms which have been proposed previously in the literature.

The contribution of this work includes:

- The proposal of an architecture for an intelligent RAN selection and optimisation system in heterogeneous communication environments.
- The creation of novel network resource availability evaluation models that facilitate the effective use of RAN selection and optimisation algorithms.
- The creation and evaluation of novel adaptive RAN selection and optimisation algorithms, which can optimise the usage of network resources and cope with different scenarios, network conditions, and aims of operators.

The author’s publications are listed in Appendix A.

1.3 Organisation of the Thesis

Chapter 2 firstly presents several integration schemes for interworking and cooperation of different wireless systems in heterogeneous communication environments. It also studies the challenges in the implementation of context information for developing context-aware RAN selection and optimisation algorithms.

Chapter 3 presents a literature review of algorithms used for developing RAN selection schemes.

Chapter 4 introduces the architecture of the proposed RAN selection and optimisation system. This chapter also describes the system’s functional configuration and presents the signalling and protocol message exchanges between different functional entities.

Chapter 5 describes the network resource availability evaluation models for the UTRAN and IEEE 802.11a/b based WLAN. The evaluation models, based on appropriate assumptions, are the abstractions of real networks and able to effectively obtain the context information of network conditions which helps the development of dynamic and adaptive RAN selection and optimisation algorithms.

Chapter 6 validates the network resource availability evaluation models. It demonstrates the implementation of the evaluation models and compares the evaluation results with call level and packet level simulation results for validation.

Chapter 7 presents the development of efficient and adaptive RAN selection and optimisation algorithms. Simulation results and analysis are presented in chapter 6.

Chapter 8 gives a general discussion and evaluation of the work. A conclusion is presented. Finally, this chapter also discusses the future work.

Chapter 2 Integration Aspects of Heterogeneous Wireless Networks

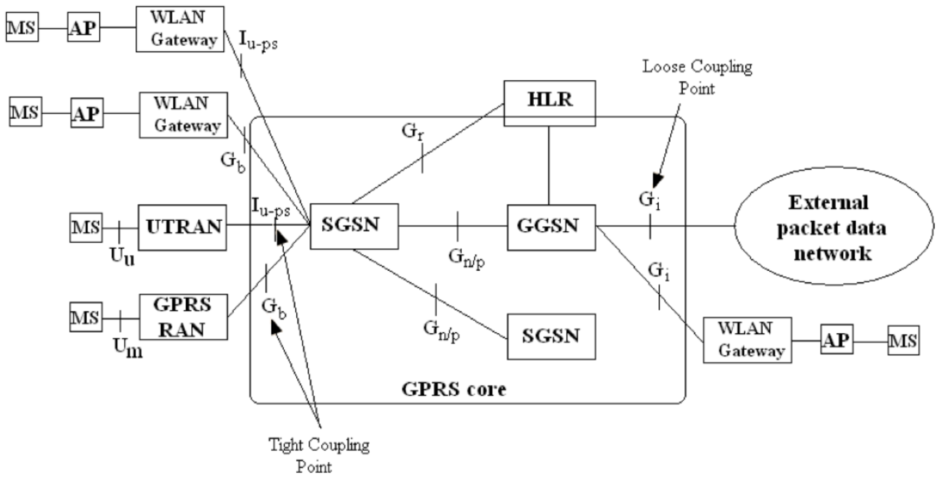
The overlapping of different wireless network technologies creates heterogeneous communication environments. Future mobile communication systems not only concentrate on developing novel radio access technologies and system architectures, but also consider the cooperation and interworking with legacy communication systems. Future mobile communication systems facilitate the provisioning of diversified and flexible services which can fulfil different user requirements [AIPN05]. Furthermore, with the cooperation and interworking with other communication systems, these services can be ubiquitous. The access to the mobile services will not be restricted to the mobile cellular networks, but may be via other wireless or even wired technologies [LTE08]. Such universal access facilitates service convergence, joint resource management, and adaptive quality of service [LTE08]. Network operators or service providers can perform dynamic and adaptive radio resource management policies for the increasing traffic load, such as redirecting some of the traffic to other networks, and benefit for effective and optimal resource utilisation. The users also can obtain optimal radio access which will consider their requirements and status.

A RAN selection and optimisation system can facilitate the realisation of the above advantages and functionalities. The development of a RAN selection and optimisation system also relies on two factors: the system interworking architecture and the adequate context information.

2.1 System Interworking Architecture

Currently, most of the RANs are connected to their own core networks and it is a challenge to make the RANs cooperate with each other. In order to achieve seamless communications and joint resource management in a heterogeneous communication environment, interworking among different wireless systems is necessary and the development of an interworking architecture should be taken into account.

The European Telecommunications Standards Institute has specified two models for interworking between cellular networks and wireless local area data networks [ETSI01]. The models are the loosely coupling model and the tightly coupling model. The architecture of these two models is shown in Figure 2-1.



AP: Access Point   MS: Mobile Station   HLR: Home Location Register
SGSN: Serving GPRS Service Node   GGSN: Gateway GPRS Service Node

Figure 2-1 Interworking Models for Cellular network and WLAN [ETSI01]

The loosely coupling model has a WLAN gateway, which connects to the cellular core network through a *Gateway GPRS Service Node* (GGSN). This connection is via a packet switched network, such as Internet. The loosely coupling model separates the data path in the WLAN and the cellular networks, and the WLAN data packets would not be injected into the cellular core network. However, interoperation must be achieved in order to provide seamless services for users. The WLAN gateway is required to support Mobile-IP, Authentication, Authorisation, and Accounting (AAA) functionalities as the cellular network, so as to handle mobility and security management. Furthermore, interoperation also enables cellular network operators to collect WLAN accounting records and produce a unified billing statement for usage and various price schemes for both networks. The loosely coupling model presents several advantages for network integration. Firstly, it is possible to deploy the WLAN and cellular networks independently. Secondly, after the roaming mechanism agreement is achieved among service providers covering a widespread area, users can benefit from one service provider for all network access. The users will not be required to establish separate accounts for different service providers, areas, or technologies. Finally, the loosely coupling model also provides the possibilities to integrating private WLANs with public WLAN and the cellular networks. However, the mobility support that normally uses Mobile IP may introduce inefficiency and high latency for handovers [SFP02].

The basic idea of the tightly coupling model is to make the WLAN act as another cellular access network to the cellular core network. The WLAN is connected to the core network via a standardised communication interface (e.g. Iu-ps). The WLAN also needs to emulate the functions that belong to cellular access network. In Figure 2-1, the attachment point in the tightly coupling internetworking is presented. In this architecture, a WLAN gateway appears to the cellular core network as a *Serving GPRS Service Node* (SGSN). This gateway is responsible for hiding the details of the WLAN to the cellular core, mapping and implementing all the cellular protocols, such as mobility management, accounting and authentication. In this situation, user terminals should implement the cellular network protocol stack on the top of the standard WLAN interface, and they should be able to switch from one mode to another. Furthermore, all the traffic generated by user terminals in the WLAN would be injected to the cellular core network. The two different wireless networks will share identical authentication, signalling, transport and billing systems. The main advantages of the tightly coupling model are: a more efficient mobility management; the reuse of the cellular standard protocols; and the reuse of network resources, such as subscriber databases and billing systems. However, this model may be more suitable when the WLANs are owned by cellular network operators. Also the cellular core network may have to expand its capacity to support the increased high traffic load injected from WLANs. One example of an implementation of the tightly coupling model is the *Universal Media Access*, which will be described later in this chapter.

In [Wu02], three typical models providing interworking for different wireless systems are classified, which are shown in Figure 2-2. The models include the tunnelled network model, the heterogeneous network model, and the hybrid network model. The difference between these models lies on the layer where the wireless systems are connected with each other.

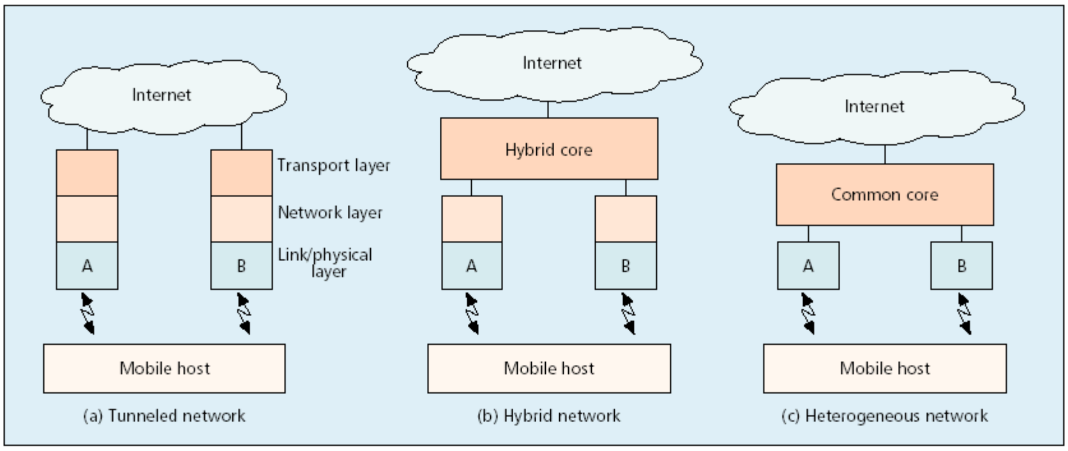


Figure 2-2 Three Interworking Models [Wu02]

The tunnelled network model allows different wireless systems, supporting layers from physical to transport, to be connected with each other. The connectivity is made through a packet-switched network, such as the Internet. The tunnelled network model requires service agreements between the operators of several independent wireless systems. The advantage of this model lies in few modifications to the existing wireless systems. It is easy to deploy even if the existing wireless systems belong to different administration domains. However, this model also may introduce inefficiency in exchanging data between different wireless systems.

The hybrid network model employs a hybrid core network for the communications of different wireless systems. The hybrid core network situates above the network layer and provides a means for data exchange between different systems. The connected communication systems implement the functionalities of the network layers and the layers below. Therefore, fewer duplicate functions at the transport layer are included.

The heterogeneous network model employs a common core network. As a single network, the common core network deals with core network functionalities and operations of different wireless systems. Therefore, the connected wireless systems just handle the tasks related to the physical layer and link layer. This model can effectively reduce the overhead and improve the network performance and efficiency. However, the challenge to implement this model is the introduction of new functionalities or amendments to the protocol stacks of existing wireless systems, and the necessary commitments between service providers.

2.1.1 Research Proposition

In the subsection, three research propositions for enabling the interworking between different wireless systems are presented. They are the *Wireless Network Management Evolution*, the *Multimedia Integrated network by Radio Access Innovation* project, and the *Wireless World Initiative New Radio* project.

2.1.1.1 Wireless Network Management Evolution

Demestichas et al. [DKKTOSPTVE-K03][DVE-KT04] propose an evolution in wireless network management: *Composite Radio Environment* (CRE). The architecture of the CRE is shown in Figure 2-3. The architecture is based on the tunnelled network model and the wireless systems are integrated together through an IP-based backbone network. Each wireless system maintains the functionalities in the physical, data link, network and transport layers. Furthermore, the Mobile-IP technology is used for maintaining IP level connectivity regardless the underlying changes in lower layers.

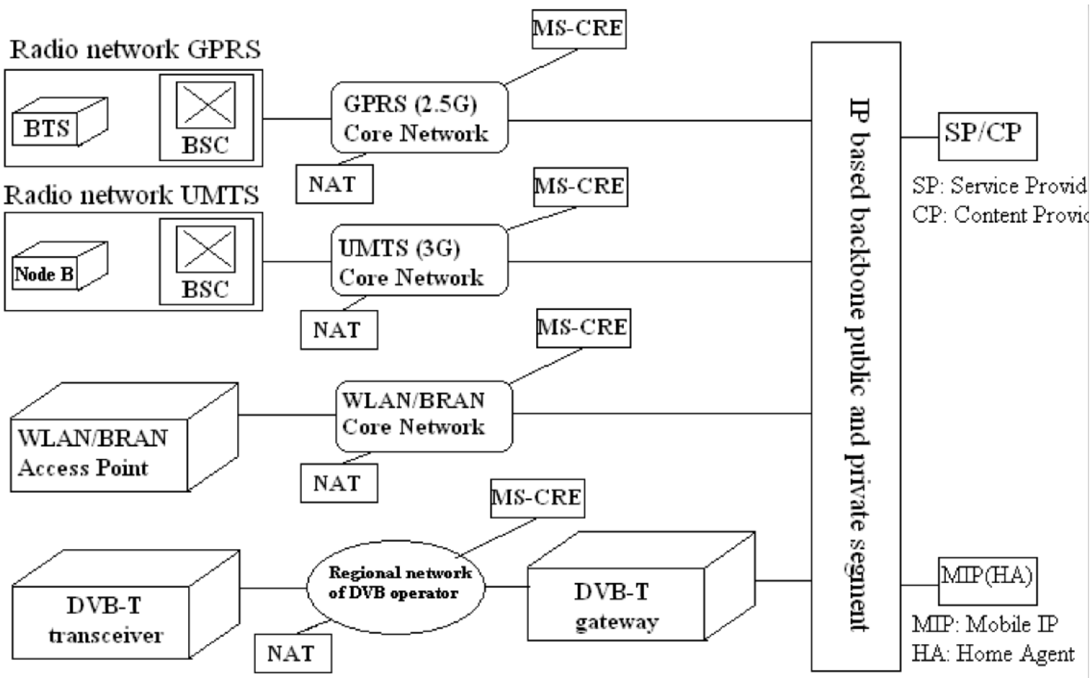


Figure 2-3 Architecture of a Composite Radio Environment [DVE-KT04]

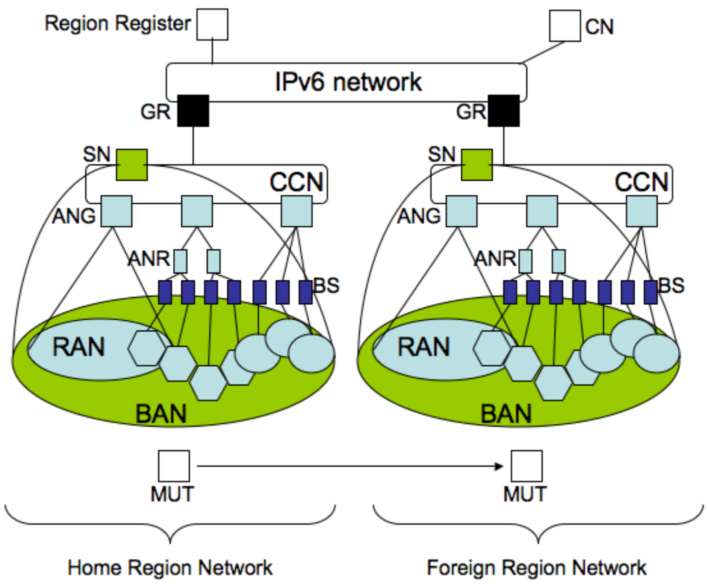
In each wireless system's core, an intelligent entity, *Management System for Composite Radio Environment* (MS-CRE), is employed for management and cooperation purposes. The MS-CRE is responsible for exploiting network capabilities by exchanging information with the MS-CRE peers in other wireless systems. Based on the network information, the MS-CRE directs users to the appropriate networks according to service costs, provided QoS levels, and user and terminal profiles. It also provides the brokerage functionality and facilitates the cooperation between different wireless systems. In addition, a *Terminal Management System* (TMS) is also introduced to the user terminal. The TMS assists and supports network discovery and selection, vertical handover control, QoS monitoring, and user and terminal profile control.

In the CRE, a user terminal may support more than one RAT. There is no common signalling network. Therefore, the user terminal is pre-configured offline to work with one or more candidate wireless systems. When the user initialises a service request, the user terminal interacts with the *Session Manager* (SM) within the MS-CRE situated in each candidate wireless system. The SM manages the interaction between the user and the wireless system. It also supports network selection by providing commands and recommendations which can direct the user to a suitable network.

2.1.1.2 Multimedia Integrated network by Radio Access Innovation Project

The *Multimedia Integrated network by Radio Access Innovation* (MIRAI) project [SDA99][BI03][MWIMM02][IMMHM04] is a part of the e-Japan Plan prompted by the Japanese Government. This project aims to develop new technologies which will enable seamless integration of a variety of wireless access systems. A conceptual view of the MIRAI system is illustrated in Figure 2-4.

In MIRAI, a *Common Core Network* (CCN) is proposed as a management core for different wireless systems and is shown in Figure 2-4. The CCN has a common database for managing user profiles, such as authentication, location, preferred access system, billing, policy, and the capabilities of the user terminal system. The base stations (BSs) belonging to different wireless systems act as the access points (APs) and interfaces to the CCN. A BS is connected to the CCN via an *Access Network Gateway* (ANG). For some RANs with small coverage area, such as WLAN, several BSs are firstly connected to an *Access Network Router* (ANR) then to the ANG.



GR: Gateway Router; SN: Signalling Node; CCN: Common Core Network
ANG: Access Network Gateway; ANR: Access Network Router; BS: Base Station
BAN: Basic Access Network; MUT: Multi-service User Terminal

Figure 2-4 Conceptual view of the MIRAI system [IMMH03]

Within a CCN, a *Signalling Node* (SN) is used to transmit, receive and route signalling information. The CCNs are connected to the Internet through *Gateway Routers* (GRs). Mobile IPv6 is used for connecting CCNs and supporting global mobility management. Within a given area managed by one CCN, local mobility management is provided by fast handover.

In MIRAI, two concepts are also proposed: *Basic Access Signalling* and *Basic Access Network*. The *Basic Access Signalling* (BAS) provides a set of functionalities specific to heterogeneous communication environments, including user registration, information transfer and reception, heterogeneous paging support, location information provision, RAN discovery, network selection decision transfer, vertical handover support, and mobility management. The BAS messages are transmitted between the users and the network management entities over a *Basic Access Network* (BAN). The BAN is supposed to have a broad coverage area and a reliable communication link. The BAN also can act as a RAN and provide access service for applications which require low bandwidth. Inoue et al. present three approaches for establishing the BAN [IMMH03][IMMHM04].

The first approach is to implement a specially designed, out-of-band and dedicated signalling network [IMMH03]. In [IMMH03], a wireless network with frequency around 394 MHz and a specific MAC layer frame is presented. In both experimental and simulation situations, this network shows good properties as a dedicated signalling network for paging call initiation and temporary address release. The second approach does not use any specific wireless network for signalling. It lists all available RANs and allows a user to make a choice [IMMHM04]. The last approach is to make use of the existing Public Land Mobile Network (PLMN) as a signalling network. The PLMN has the following characteristics: wide area coverage, reliable link, adequate transmission rate, relatively low power consumption, and high popularity [IMMHM04]. Implementing a signalling network by using existing PLMN means that the signalling and the context information should be delivered by using the existing technologies as well.

Based on the BAS and BAN functionalities, an example of the signalling procedure of RAN discovery and selection is presented in Figure 2-5.

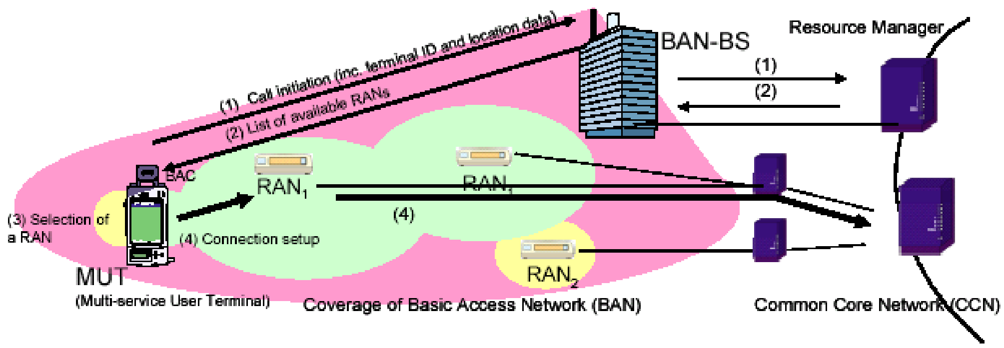


Figure 2-5 Example of signalling procedure of RAN discovery and selection using BAN and BAS [IMMH03]

In order to initialise a call, the user terminal firstly sends a BAS message containing the ID and location of the user terminal to the *Resource Manager* via BAN. The *Resource Manager* makes use of the location information and other information such as the network status and user preferences and then sends a list of available RANs back to the user terminal. The user can select the appropriate RAN from the list and finally set up the connection to it.

2.1.1.3 The Wireless World Initiative New Radio Project

In 2003, ITU-R approved recommendation M.1645, *Framework and Overall Objectives of the Future Development of IMT-2000 and Systems beyond IMT-2000*. This recommendation considers that the development of systems beyond 3G comprises two major paths: first, the legacy and evolved access networks will be integrated on a packet-based platform to enable cooperation and interworking of heterogeneous systems in the sense of "*optimally connected anywhere, anytime*" [M.1645]; second, a new radio access system for mobile and nomadic/local area wireless access will be developed to provide services with significantly improved performance. The *Wireless World Initiative New Radio* (WINNER) project aims to fulfil the recommendations of M.1645. The WINNER project [WINNER] concentrates on evaluating promising technologies for future mobile communication systems and understanding what structures and functionalities future system entities might have. It develops a new radio access system and an interworking architecture for cooperative radio resource management. The outcomes of the WINNER project are not products, but a clear concept of what future mobile communication systems should look like [B3G08].

The radio access system developed by WINNER supports multiple operating modes and is able to provide wireless access for local, metropolitan and wide areas. The access system aims to provide seamless services to users in different environments and support a full range of mobility scenarios [TMM-DK07]. The *WINNER Radio Access Network* (WRAN) offers ubiquitous access services and is also able to interwork with other access technologies. Furthermore, the WRAN not only handles tasks related to radio access, but also implements functionalities more related to the core network. The RAN of future mobile systems will be more powerful and intelligent in a flatter network architecture [BSMPR07]. The logical node architecture of the WRAN is presented in Figure 2-6. The logical node architecture includes six types of entities: an IP anchor GW_IPA_{LN}, a control node GW_C_{LN}, a base station BS_{LN}, a relay node RN_{LN}, a user terminal UT_{LN}, a spectrum server $SpectrumServer_{LN}$, and a radio resource management (RRM) server $RRMserver_{LN}$.

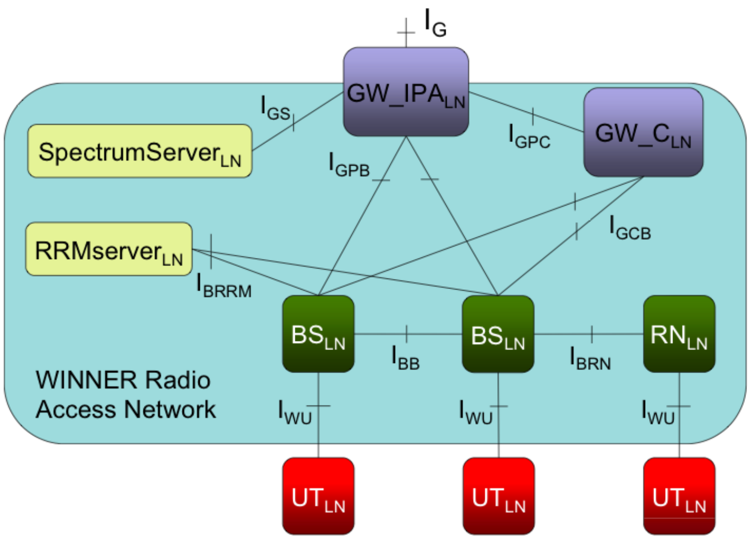


Figure 2-6 WINNER Radio Access Network Logical Node Architecture [WINNER 6.13.14]

The WRAN supports two types of radio access points: the BS_{LN} and the RN_{LN}. A RN_{LN} serves as a relay node between a base station BS_{LN} and a user terminal UT_{LN}. A RN_{LN} can be wirelessly connected to a BS_{LN} or another RN_{LN}. The radio transmission and reception between a UT_{LN} and a RN_{LN} are governed by a BS_{LN}, which manages the resources used by the RN_{LN}. The BS_{LN} also handles all radio related tasks for active user terminals. Therefore, a UT_{LN} can be connected to the WRAN either directly via the BS_{LN} or indirectly via the RN_{LN}.

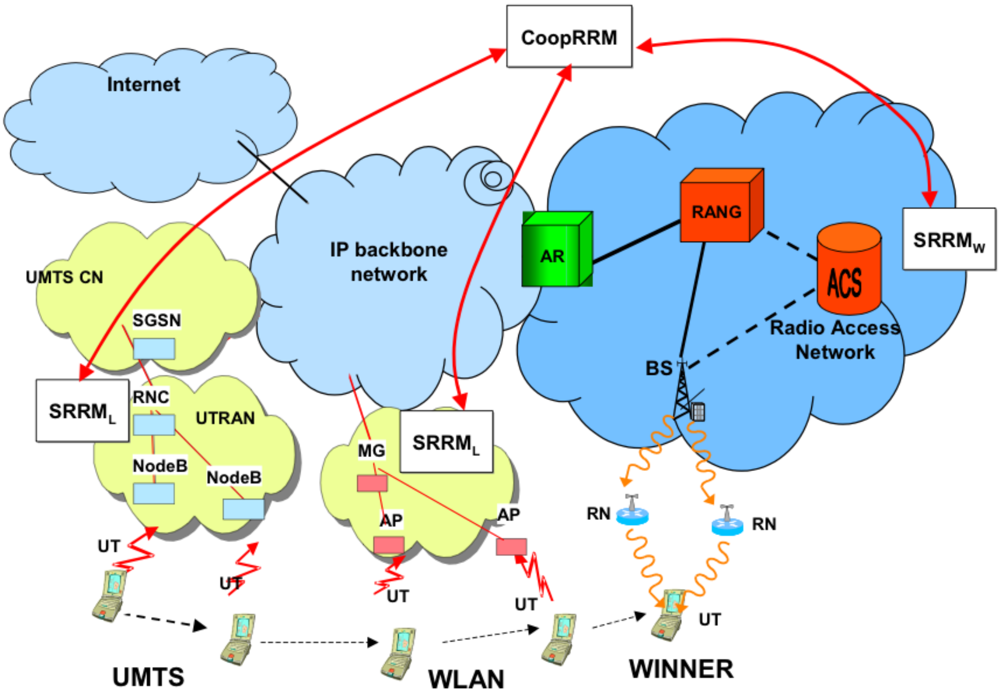
A BS_{LN} is connected to the GW_IPA_{LN} and GW_C_{LN} pair via the I_{GPB} and I_{GPC} interfaces, and these connections can be multipoint-to-multipoint. That means, a BS_{LN} is able to connect to multiple GW_IPA_{LN} and GW_C_{LN} pairs and a GW_IPA_{LN} and GW_C_{LN} pair also can connect to multiple UT_{LN}. GW_IPA_{LN} operates on user data traffic and facilitates users to access external data networks and operator services. GW_IPA_{LN} also terminates the data flows from external data networks and acts as an anchor point for external routing. Each GW_IPA_{LN} is accompanied by a GW_C_{LN}, which controls the inactive user terminals and also controls and configures the associated GW_IPA_{LN}. When a GW_IPA_{LN} and GW_C_{LN} pair serves a UT_{LN}, the UT_{LN} is said to be associated with the GW_IPA_{LN} and GW_C_{LN} pair. Considering the interconnection between the BS_{LN} and the GW_IPA_{LN} and GW_C_{LN} pair, two UT_{LN} connected to the same BS_{LN} may be associated with different GW_IPA_{LN} and

GW_C_{LN} pairs, while two UT_{LN} associated with the same GW_IPA_{LN} and GW_C_{LN} pair may be connected to different BS_{LN} entities.

The $SpectrumServer_{LN}$ controls spectrum sharing and assignment in a WRAN and between WRANs. It also enables co-existence with other radio access technologies. The $SpectrumServer_{LN}$ receives information from a GW_C_{LN} about the associated UT_{LN} entities and their characteristics and transfers spectrum sharing and assignment decisions to the BS_{LN} entities. For spectrum negotiation, the access to the $SpectrumServer_{LN}$ can be made via GW_IPA_{LN}. The $RRMserver_{LN}$ is responsible for centralised radio resource management.

One requirement for deploying the WINNER system is to interwork with existing wireless communication systems and coordinate the RRM activities in each RAN. Such interworking can facilitate seamless introduction of WINNER's new radio interface while exploiting the installed base of legacy systems [WINNER4.4]. The functionalities supported by the interworking and cooperation architecture include admission control, handover, scheduling, and QoS based management. Other functionalities, such as AAA which is outside of the scope of WINNER, also can be supported. Figure 2-7 presents the interworking and cooperation architecture and the relationship between different RRM entities.

As presented in Figure 2-7, different communication systems are integrated via an IP-based backbone network. A *Specific Radio Resource Management* entity (SRRM) is situated in each RAN and it cooperates with other SRRMs via a *Cooperative Radio Resource Management* entity (CoopRRM). The SRRM is responsible for adapting the RAN to cooperate with the CoopRRM. The SRRMs situated in the legacy RANs have two functionalities: traffic monitoring and CoopRRM command translation. When an SRRM is monitoring the traffic of the RAN, it compares the measurement with a threshold. If the measurement surpasses a certain threshold and the serving RAN cannot solve the situation, a trigger will be activated and a request for cooperation will be transmitted to the CoopRRM. The SRRM in legacy RANs will have two interfaces: one is to the CoopRRM and the other one is to the legacy RAN RRM entity. The interfaces enable the SRRM not only to receive and translate commands from the CoopRRM, but also to transfer the commands and actuate the RAN.



AR: Access Router   RANG: Radio Access Gateway   ACS: Access Control Server

Figure 2-7 WINNER Cooperation Architecture [WINNER4.4]

The CoopRRM is situated at a neutral position among the RANs and coordinates the RRM entities in each RAN. The CoopRRM comprises two parts: a common part and a specific part. The common part of CoopRRM includes functionalities which are common to all RANs and also provides a common interface to the upper layers of the protocol stack. The specific part includes functionalities which are devoted to cope with specific tasks of each RAN. When the CoopRRM receives a request from a SRRM, the CoopRRM may reply with a message containing the resource status of other RANs. This would allow an inter-RAN handover. The CRRM may send a command to the target SRRM to actuate its RAN. Several scenarios can activate the cooperation between RANs, including unsupported QoS in serving RAN, handover requests caused by congestion, or coverage loss [WINNER4.4].

Currently, the WINNER project has been extended into a subsequent project, the WINNER+. The WINNER+ project intends to develop, optimise, and evaluate the IMT-Advanced compliant technologies. The IMT-Advanced system goes beyond the objectives of M.1645 and provides universal access to a wider range of telecommunication services. The IMT-Advanced system utilises advanced radio access technologies and is supported by the packetised mobile and fixed networks.

Some of the key features of the IMT-Advanced system include: compatibility with the services provided by IMT systems and fixed networks; support for advanced services and applications with enhanced peak data rates, such as 100 Mbit/s for high mobility and 1 Gbit/s for low mobility [IMT-ADV08].

2.1.2 Interworking Architecture Standards and Recommendations

In this subsection, efforts for deploying and realising the interworking between different wireless systems are presented. They are the *Universal Media Access*, the 3GPP *Long Term Evolution*, and the *Next Generation Mobile Networks*.

2.1.2.1 Universal Media Access and Generic Access Network

Universal Mobile Access (UMA) implements the tightly coupling model and provides a standard, scalable and secure access to a mobile core network over IP [UMA]. Based on the UMA, mobile operators can extend circuit, packet or *IP Multimedia Subsystem* (IMS) based services via any IP-based access network. For example, by using a dual-mode cellular/WLAN mobile station, a user can access GSM/UMTS services through the WLAN on his/her premises.

To enable the secure transport of mobile signalling and user plane traffic over IP, the UMA deploys a *UMA Network Controller* at the network side and develops a set of associated protocols. The specifications of UMA are included into the 3GPP UMTS release 6 specifications [UMA], where UMA and *UMA Network Controller* are referred to as *Generic Access Network* (GAN) and *Generic Access Network Controller* (GANC), respectively. The GAN functional architecture is illustrated in Figure 2-8.

To the cellular system's core, the GANC serves as a *GSM/EDGE Radio Access Network* (GERAN) base station subsystem or a UTRAN radio network controller. It also includes a *Security Gateway* (SEGW) which terminates remote access from user terminal and provides security functionalities for signalling, voice and data traffic, e.g. mutual authentication, encryption, and data integrity [GAN09]. Between the GANC and the user terminal, an interface U_p is defined and a generic IP access

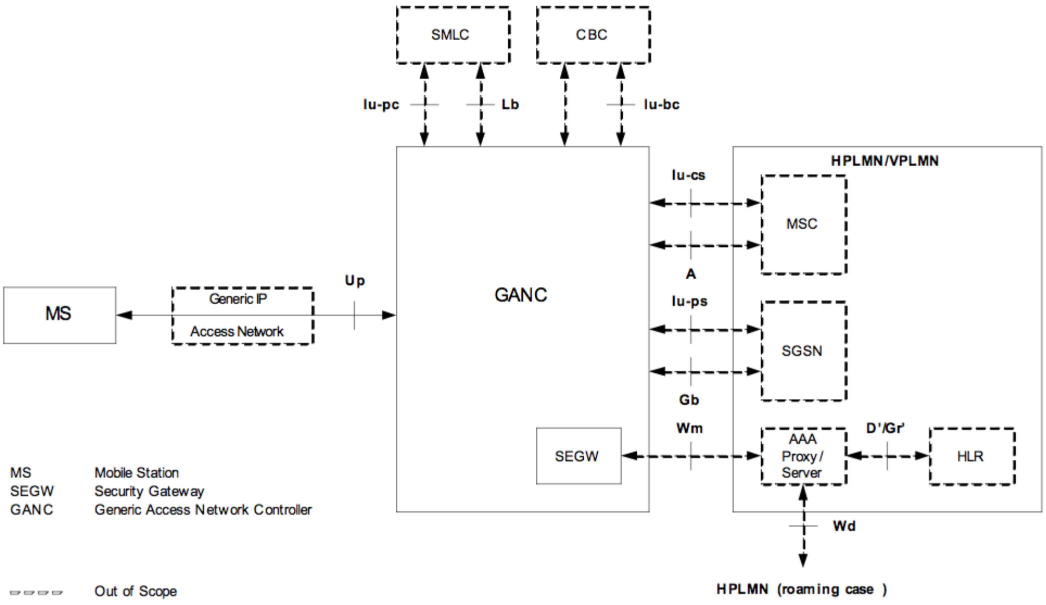


Figure 2-8 Generic Access Network Functional Architecture [GAN09]

network provides the connectivity. The GANC interworks between the generic IP access network and the cellular system's core for both control and user planes. In the control plane, the GANC supports a set of functionalities, including security functionalities (carried out by the SEGW), GAN access registration and system information provision, GAN *Circuit-Switched* (CS) and *Packet-Switched* (PS) services bearer paths management, paging and handover support, and layer 3 message transfer between mobile stations and the core network [GAN09]. In the user plane, when connected to the GSM/GPRS core, the GANC provides reframing and transcoding functionalities for circuit switched services and connects data transport channels over the U_p interface to packet flows over the G_b interface [GAN09]. When the GANC is connected to a UMTS core, it transports circuit switched user data between the U_p interface and the Iu-cs interface and transports packet switched user data between the U_p interface and the Iu-ps interface.

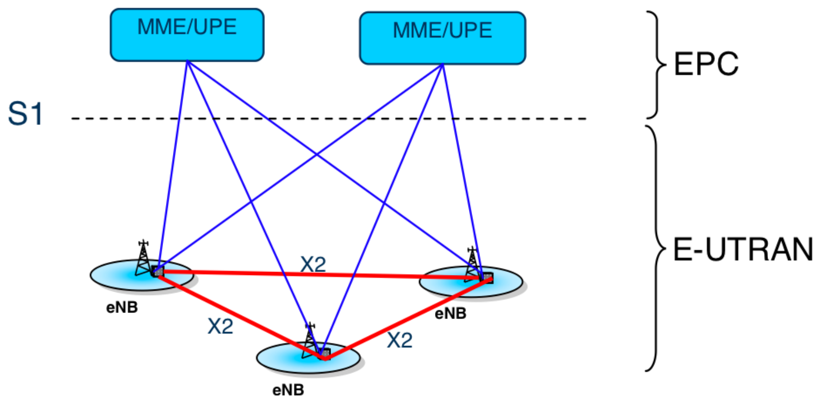
A dual-mode user terminal can activate any operation mode at anytime. Furthermore, preferences on the operation modes are also supported and can be configured by users or operators through different mechanisms [GAN09]. The preferences on the operation modes can possibly include GERAN/UTRAN-only, GERAN/UTRAN-preferred, GAN-preferred, and GAN-only [GAN09]. When GERAN/UTRAN-only is activated, the mobile station will operate only in the GERAN/UTRAN mode, but not switch to the GAN mode. When GERAN/UTRAN-preferred is configured, the mobile station will operate in the GERAN/UTRAN mode until the GERAN/UTRAN cell is not available. Then, if the mobile station has successfully registered with a GANC over the generic IP access network, it will switch to the GAN mode. Once a suitable GERAN/UTRAN cell is available, the mobile station will switch back to the GERAN/UTRAN mode. When GAN-preferred is selected, the mobile station which has registered with a GANC over the generic IP access network will remain in the GAN mode until the generic IP access network is not anymore available. Then, the mobile station will switch to the GERAN/UTRAN mode. When GAN-only is activated, the mobile station will firstly operate in the GERAN/UTRAN mode to obtain mobile network information, and then switch to the GAN mode.

Initially, the UMA/GAN services are provided through UMA/GAN-enable dual-mode mobile stations. The UMA/GAN also offers solutions for deploying femtocells, terminal adaptors, and softmobiles [UMA]. The UMA/GAN technology not only enables mobile operators to effectively and efficiently utilise network resources, but also it provides a means to integrate heterogeneous wireless systems.

2.1.2.2 3GPP Long Term Evolution and System Architecture Evolution

3GPP *Long Term Evolution* (LTE) focuses on improving universal terrestrial radio access towards to a high speech, low delay and packet optimised radio access technology [LTE08]. Some of the key requirements for LTE include: optimised packet switched domain, low server to *User Equipment* (UE) round-trip time (below 30ms) and access delay (below 300ms), enhanced peak uplink and downlink data rates (50 Mbps and 100 Mbps), ensured mobility and security, improved UE power efficiency, flexible spectrum allocation, and improved spectrum efficiency compared with *High Speed Packet Access* (HSPA) reference case (three to four times in the downlink and two to three times in the uplink) [HT07][LTE08].

For multiple access, LTE uses *Orthogonal Frequency Division Multiple Access* (OFDMA) in the downlink and *Single Carrier Frequency Division Multiple Access* (SC-FDMA) in the uplink. By using these two multiple access mechanisms, LTE can facilitate larger bandwidth and bandwidth flexibility in downlink direction and enable better power and amplifier efficiency in uplink direction. Furthermore, LTE adds more intelligence to base stations which are called *evolved Node Bs* (eNodeBs). An eNodeB not only handles radio related tasks, but also hosts radio resource management functionalities, including radio bearer control, radio admission control, connection mobility control, and dynamic resource allocation (scheduling) [LTE08]. The radio resource management functionalities improve radio access performance and enable a flat RAN architecture, which is shown in Figure 2-9. An eNodeB can connect to another eNodeB via reference point X2. An eNodeB is connected to the *Mobility Management Entity* (MME) and *User Plane Entity* (UPE) via a reference point S1 and this connection can be multipoint-to-multipoint. The MME is in charge of UE control plane context, such as identities, mobility states, security parameters, and it generates and allocates temporary identities to the UEs. It also authenticates users and authorises user's admission in the tracking area or PLMN [SAE08]. The UPE is in charge of UE user plane context, such as IP bearer service parameters or network internal routing information. It terminates downlink data path for UEs at idle state, triggers paging for the UE when its downlink data arrive and it performs user traffic replication in case of lawful interception [SAE08].

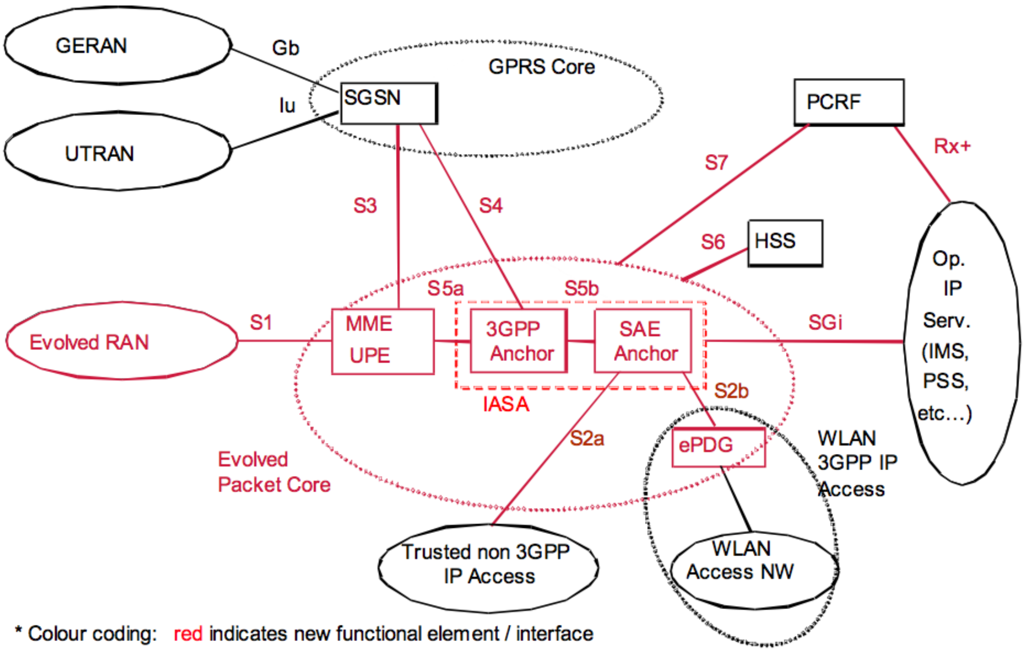


MME: Mobility Management Entity   UPE: User Plane Entity   EPC: Evolved Packet Core
E-UTRAN: Evolved UTRAN   eNB: Evolved Node B

Figure 2-9 Evolved UTRAN Architecture [EUTRAN08]

In the same time frame as the development of LTE, 3GPP also undertakes the *System Architecture Evolution* (SAE) and aims to develop an evolved core network to provide an architectural support to the LTE [LTE08]. The evolved core network is "*all-IP*" and optimised to provide various packet switched services and support IMS.

The evolved core network also supports multiple RATs and the access to the provided services is not restricted to 3GPP access systems, but it can be via WiFi, WiMAX or even wired technologies. The main objectives of the SAE include: development of a core network architecture which can support the high speech and low delay LTE access system, development of a core network which provides all services via PS domain, and development of a core network architecture which supports mobility between heterogeneous access systems [LTE08]. The logical high-level architecture for the evolved core network is presented in Figure 2-10.



IASA: Inter Access System Anchor HSS: Home Subscriber Server

PCRF: Policy Charging Rule Function ePDG: Evolved Packet Data Gateway

Figure 2-10 Logical High Level Architecture for the Evolved System [SAE08]

The *SAE Anchor* and *3GPP Anchor* are the functional entities which process user plane data. The *SAE Anchor* handles tasks relevant to mobility management between non 3GPP access systems and 3GPP access systems. The *3GPP Anchor* handles tasks related to mobility management between 2G/3G access system and LTE access system. It connects to the SGSN in the GSM/UMTS core and provides GGSN functionalities. The SGSN in the GSM/UMTS core is also connected to the MME, which handles control plane signalling and mobility management between 2G/3G access system and LTE access system. As shown in Figure 2-10, trusted non 3GPP access system can be directly connected to the *SAE Anchor* via reference point S2a. In other respects, the non 3GPP access system also can be connected to the *SAE Anchor* through an *evolved Packet Data Gateway* (ePDG). The functionalities of the ePDG may include routing, address translation and mapping, packet de-capsulation and encapsulation, management of remote IP address and local IP address, authentication, filtering of unauthorised and/or unsolicited traffic, charging information generation, etc.

2.1.2.3 Next Generation Mobile Networks

The *Next Generation Mobile Networks* (NGMN) is an alliance initiated by a group of operators and infrastructure manufacturers, e.g. Vodafone and Alcatel-Lucent [NGMN]. Facing competitions and challenges from other wireless and wired broadband technologies, the NGMN alliance presents a vision for technology evolution beyond 3G for providing mobile broadband services in a competitive way. The objective of the NGMN alliance includes establishment of clear performance targets, provision of fundamental recommendations and deployment scenarios for a future mobile broadband network, whose price/performance should be competitive with alternative technologies.

The NGMN alliance intends to introduce the seamless mobile broadband services by the provision of an integrated network. The NGMN alliance focuses on the core network and radio technology. The integrated network is proposed to be based on a PS core, to enable intelligence in the network edge, and to support various business models. With new radio access technologies, novel NGMN access network will be developed and supported by the NGMN integrated network. The integrated network also will coexist with other networks, provide a smooth migration of existing 2G and 3G networks to an IP network, and maximise the resource exploitation. The generic NGMN integrated network is illustrated in Figure 2-11. The NGMN integrated network will coexist with the CS segment of current mobile networks, whose PS segment can be supported by the NGMN system. However, in the long run, the CS segment will be phased out [NGMN06]. A full PS system will emerge and provide transparent services to the customers of the legacy mobile networks. The CS services will be replaced or emulated in the PS system. A Session Initiation Protocol based subsystem may be implemented for the control of access, service and network functions [NGMN06].

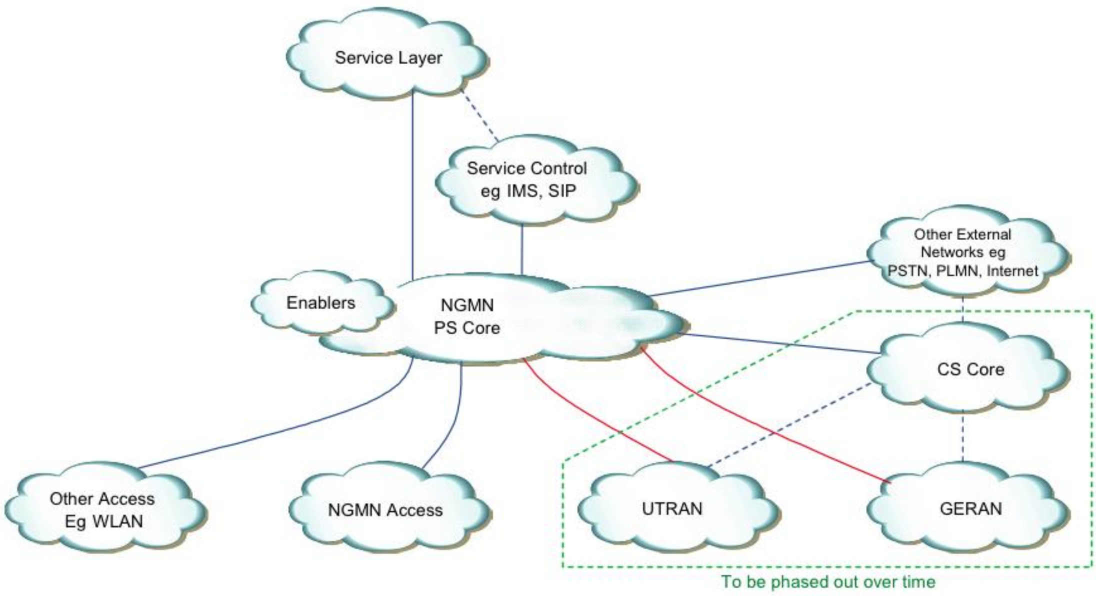


Figure 2-11 High-Level NGMN System Architecture [NGMN06]

In order to support the work of standardisation bodies and manufacturers towards a cost effective mobile communication system, the NGMN alliance provides a set of recommendations. There are three groups of recommendations. The first group are the functional recommendations which provide recommendations in several criteria enabling operators to offer flexible and attractive services. Some of the functional recommendations are related to RAN selection research. For example [NGMN06]:

1. End to end QoS in all segments and preferably optimum end to end QoS with service continuity
2. Seamless mobility management, preferably based on intelligent infrastructure
3. QoS based global roaming and interworking
4. Real time conversational and streaming in packet switched across all required bearers
5. Valued based charging for integrated network
6. Scalable core throughput to allow for deployment options that match specific operators and traffic requirements, and optimise radio resources

The second group of recommendations are related to cost efficiency. Several recommendations of this group also influence the way which RAN selection solutions should be considered [NGMN06]:

1. Fully integrated multi frequency sites IP backhaul and IP/MPLS backhaul

2. Maximum throughput without proportional incremental costs
3. One integrated network with convergence of fixed and mobile technologies
4. IMS like service management
5. Negotiated access between the terminal and the network (under the guidance of the network), preferably optimised access for the application and terminal with user preferences
6. Highly intelligent multipurpose handsets and devices

The third group are the overarching recommendations which provide guidance to evaluate deployment suitability. The NGMN also expects the integrated network to maximise resource exploitation, where terminals are required to support other RATs.

2.1.3 IEEE802.21 — A Standard Supporting System Interworking

IEEE 802.21 is a developing standard to realise *Media Independent Handover* (MIH) between interoperable heterogeneous networks and enhance user experience of user terminals [MIH09]. IEEE 802.21 standard comprises a set of elements:

- A framework which facilitates service continuity when a user terminal handovers between heterogeneous link-layer technologies;
- A logical entity, the *MIH Function* (MIHF), which consists of a group of media independent handover-enabling functions;
- The *Media Independent Handover Service Access Point*, MIH_SAP, and a set of associated primitives which enable the MIH users to access the services provided by the MIHF;
- The *Link-layer Service Access Point*, MIH_LINK_SAP, and associate primitives which assists the MIHF in collecting link-layer information and control link-layer behaviour during handovers.

The MIHF reference model is presented in Figure 2-12, which also shows the MIH services and their initiation.

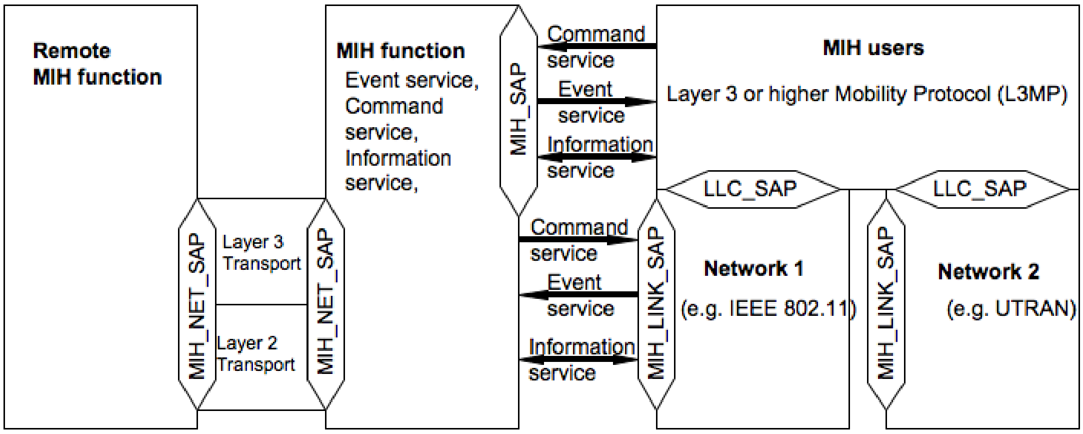


Figure 2-12 MIH Function Reference Model [MIH09]

The MIHF obtains information and event messages from lower layers and control the behaviour of lower layers through the abstracted interface MIH_LINK_SAP. Such operations are also applied between peer MIHF entities. The information and event message exchanges and command execution between MIHF peers are at layer 2 or layer 3 through the abstracted interface MIH_NET_SAP. MIH_NET_SAP supports the operations by providing transport services over the data plane on the local nodes. The MIHF provides MIH users at higher layers with *Media Independent Event Service*, *Media Independent Command Service*, and *Media Independent Information Service* and assists MIH users in performing handover operations, including service continuity maintenance, service adaptation to varying QoS, power consumption conservation, and network discovery and selection.

The *Media Independent Event Service* (MIES) takes into account the dynamic changes in link characteristics, status and quality (e.g. “Link Up”, “Link Down”, “degrading wireless link quality”), and provides event classification, filtering and reporting services. IEEE 802.21 standard categorises the event service into two types: Link Events and MIH Events. The Link Events originate from event source entities in lower layers and terminate at the MIHF. The MIH Events either originate from the MIHF, or they are Link Events, which are propagated by the MIHF to MIH users with or without additional processing. Both Link Events and MIH Events support five types of events: MAC and PHY State Change events, Link Parameter events, Predictive events, Link Handover events, and Link Transmission events. An event can be local or remote. A local event originates from and propagates across different layers of the local protocol stack of an MIH entity. A remote event traverses across different network medium from one MIH entity to another MIH entity.

For MIH users, the *Media Independent Command Service* (MICS) facilitates: the management and control of link behaviour related to handovers and mobility; the control of user terminals for optimal performance; and the implementation of optimal handover policies. Furthermore, the command service defines commands to retrieve dynamic information from lower layers, such as signal-to-noise ratio, bit error rate, etc. Also, when the MIHF entity receives certain command requests, it can generate event indications. These indications reveal and notify MIH users of future state changes in the lower layers of local protocol stack. The indications and dynamic information provided by the command service complement the media independent information service (MIIS), which typically provides less dynamic information about network operators and the availability of higher layer services.

IEEE 802.21 standard defines two categories of commands: Link Commands and MIH Commands. The MIH Commands are generated by MIH users in higher layers and transmitted to the MIHF. The MIH Commands can be local or remote. The local MIH Commands terminate at the MIHF within the local protocol stack. The destinations of a remote MIH Command are the MIHF peers in other networks or user terminals. A remote MIH Command delivered to the destination will be processed in one of three possible ways: executed by the lower layers below the MIHF peer as a Link Command; or executed by the MIHF peer as its MIH Command; or executed by the MIH user of the MIHF peer as an indication. The functionalities of the MIH Commands include: link status acquisition, link parameter threshold configuration, control of the behaviour of a group of links, handover initiation, notification of selected network, notification of handover commitment, and notification of handover completion [MIH09]. The functionalities of the Link Commands include: query and discovery of the list of supported events and commands at the link-layer, subscription/unsubscription of link event(s), acquisition of the parameter measurements of an active link, configuration of the thresholds for reporting link events, and request of an action on a link [MIH09].

The *Media Independent Information Service* (MIIS) comprises a framework and the corresponding mechanisms enabling the MIHF entities within a given geographical area to discover and collect information about the neighbouring networks. The information can be accessed via any network. For example, assuming a user is using a channel provided by a cellular network, the information about the neighbouring WLAN hot spot could be obtained via the serving cellular network.

Based on the generic mechanism provided by the information service, mobile users and network operators can exchange information about different handover candidate networks in heterogeneous communication environments. The collected network information enables a network selector or a mobility management entity, at higher layers, to assemble the knowledge of neighbouring networks and perform intelligent, efficient and optimal inter-media handover decisions. The information supported by the information service can be divided into four categories. The first category is the information about the availability of the networks within a given geographical area. The second category is the information about link-layer parameters of the networks, for example the knowledge of whether security and QoS are supported by a candidate network. The third category is the information about the capabilities of candidate networks, such as their supported channels. This type of information can facilitate optimal radio configuration and efficient handover decision-making without thorough scanning or beaconing. The fourth category is the information about the supported services in higher layers of the candidate networks. The information service also defines two ways to represent the information: XML or binary coding.

2.2 Context Information

The implementation of a RAN selection and optimisation system in heterogeneous communication environments helps to facilitate service convergence, joint resource management, adaptive quality of service, and the provision of ubiquitous and diverse radio access services.

Assuming the user terminal has inbuilt more than one RAT and different applications can be executed on it, there should be mechanisms to select transparently the most appropriate RAN for the user, and adjust network resource usage in the communication start up and handover stages. In heterogeneous communication environments, RANs differ in access technologies and some may be more suited to provide certain types of services. In the RAN selection and optimisation, not only the received signal strength, but also other metrics should be taken into account, including network performance and conditions, user terminal conditions, service types, user preference, etc [MZ04]. These metrics cover the knowledge and conditions of the networks, user terminals, and users. Based on these metrics, the RAN selection and optimisation system can dynamically and adaptively adjust the network resource usage and select an appropriate RAN, which can best serve the user's request. Therefore, in this thesis context-aware selection is defined as the selection based on the context of the users, user terminals, and networks involved.

Context is a general word and has a loose definition. In [DA99], context is defined as "*any information which can be used to characterise the situation of an entity, where an entity can be a person, place, or a physical or computational object*". In [CK00], the definition of context is enumerated into four categories, such as computing context, user context, physical context, and time context. User context comprises the user profile, location, people nearby, and even the current social situation. Physical context consists of the lighting, noise levels, traffic conditions, and temperature. Time context is about the time of the day, week, month, and season of the year. Some researchers also define context as the knowledge about the user's and IT device's state, which involves the surroundings, situation, and even location. Therefore, context can be considered to be the set of environmental states and settings which either determines an application's behaviour or causes an application event to take place. The above understanding of context can be visualised in the following scenario: a user with an intelligent mobile terminal gets appropriate services or information in specific areas. The use of context information can be applied to enhance network services, such as handover, paging, etc. Christian Prehofer et al. propose a framework for handover decision-making based on context information [PNW03]. The context information is regarded as coming from different network layers and device entities. Table 2.1 presents a snapshot of the context information classification for mobile networks and devices.

	Mobile Device	Network
Static	User settings and profile	User profile and history
	Application settings	Network location, capabilities, and services
	Willingness to pay	Charging models
Static in a cell	Reachable access points	Potential next access point
Dynamic	Type of application	Location information and location prediction
	Application requirements	Network status such as signal strength
	Device status (battery, interface status, etc)	Network traffic load

Table 2.1 Context Information Classification [PNW03]

With the emergence of the context-aware decision-making, the problems for managing the context information also come forth [PNW03]. Firstly, context information is distributed and not situated in one single entity. Secondly, some context information is dynamic and changes frequently. It also looses accuracy as time passes. Thirdly, as context information changes, the relevant methods to interpret the information may need to change as well. These problems introduce difficulties and challenges in processing and transferring the context information. Efforts for understanding what types of context information are required, how they are described, acquired, and delivered should be conducted in order to clarify the feasibility of developing a context-aware RAN selection and optimisation system.

2.2.1 Context Information Definition and Description

As presented in the previous section, several metrics should be considered while selecting the appropriate RAN and adjusting network resource utilisation in heterogeneous communication environments. These metrics also provide an insight about the types of context information that are required in the selection. They are specified as follows:

- **Network Performance and Conditions:** Various parameters, such as channel propagation characteristics, interchannel interference, signal-to-interference-noise ratio, bit error rate, latency, jitter, etc., reflect network performance. Network conditions include a set of parameters as traffic load, available bandwidth, and RAN location. In some ways, the network performance can be affected by the network conditions. For example, in a CDMA cellular system, heavy traffic load may introduce higher interference, which decreases the signal-to-interference-noise ratio. The network performance and conditions influence QoS, user satisfaction, and the utilisation of network resources. Therefore, they need to be taken into account in the RAN selection and optimisation.

- **User terminal Conditions:** Some attributes and dynamic factors of the user terminal, such as access interface types, battery status, running applications, and location information, also need to be considered. For example, by knowing the access interface types, the RAN selection and optimisation system just needs to consider certain types of candidate networks covering the user. This effort can simplify the computation complexity and reduce the computation time. By understanding the position information, including user's trajectory and moving history, together with the knowledge of network location and coverage area, the RAN selection and optimisation system can determine the RANs along the user's path, reserve sufficient resources, and coordinate the handover procedure in advance. The position information can help the RAN selection and optimisation system to improve user satisfaction.

- **Service Types:** Different types of services or applications have different requirements about network performance and necessary network resources to maintain their QoS requirements (e.g. latency, jitter, data rate, packet loss, bit error rate). In [HT07], according to their QoS requirements, the services are differentiated into four classes: conversational class, streaming class, interactive class, and background class. Their main parameters are presented in Table 2.2.

	Conversational class	**Streaming class**	**Interactive class**	**Background class**
Transfer latency	≤80ms	≤250ms	N/A	N/A
Guaranteed bit rate	Up to 2 Mbps	Up to 2 Mbps	N/A	N/A
Traffic handling priority	N/A	N/A	1, 2, 3	N/A
Allocation/retention priority	1, 2, 3	1, 2, 3	1, 2, 3	1, 2, 3

Table 2.2 UMTS QoS classes and their main parameters

The conversational class comprises the services which require low end-to-end delay and symmetric or nearly symmetric traffic between uplink and downlink. Streaming class services have the same requirements as conversational class services for bandwidth, but streaming services tolerate some delay variations. Interactive class services can be characterised by a request-response pattern: the end user expects the response within a certain period of time and can tolerate longer delays. The background class services are meant for delay tolerant services, such as email. In the RAN selection and optimisation, the characteristics of different service types should be thoroughly considered.

- **User Preference:** Users may have specific requirements for the candidate networks and have preference over certain metrics, such as available bandwidth, delay, monetary cost, etc. Users also may have preference over specific RATs.

The use of the above context information considerably enhances the RAN selection and optimisation. However, the variety of the context information and the heterogeneity of the networks and user terminals introduce challenges and difficulties for one to provide a universal approach for context description. The approach should not only be interoperable and device independent, but also be extensible and flexible, because the new types of context information may appear in future. *Extensible Markup Language* (XML) [XML04] is one possible solution to describe the context information. XML is a simple and flexible text format from *Standard Generalized Markup Language* (SGML). It is originally designed to overcome the challenges in large-scale electronic publishing and is currently playing an important role in the exchange and transfer of various data on the web and other places [XML04]. XML is widely accepted in industry and regarded by many as the de facto standard for document encoding [Li03].

Another context description proposition is the *Composite Capabilities/Preference Profile* (CC/PP) [CCPP07]. CC/PP provides an XML-based format for context information presentation. It provides a vocabulary for expressing context information relevant to device capabilities and user preferences. The original purpose of CC/PP was limited in allowing a server to deliver customised content to a device. However, Indulska et al. [IRRH03] introduced an extension to the vocabulary of CC/PP, which allows mobile devices to capture basic context information (such as user context information, device capabilities and application profiles) in a pervasive system environment.

2.2.2 Context Information Acquisition

Context information is distributed among networks and user terminals. It is a challenge to collect the required context information for RAN selection and optimisation. Context information about networks includes network performance and conditions. Context information about user terminals includes user preference and user terminal conditions.

2.2.2.1 Network Information

A RAN selection and optimisation system makes use of the context information of candidate networks for performance assessment, network selection, and service and resource utilisation adaptation. Information such as network congestion can be obtained by analysing implicit network feedbacks, for example, packet loss. However, this method requires the RAN selection and optimisation system to have the ability to analyse various network feedbacks, which may increase the complexity of system design. Moreover, the information provided by using this method may not be precise.

Remos system [LMSGSS98][MS00][DGKLMSS01] is one solution for collecting network context information. The Remos system provides resource information for distributed applications. It supports resource measurement in various environments with different network architectures, operating systems, and hardware. The Remos system also considers the resource requirements from different applications, such as real-time video. It supplies the applications with explicit resource information at an appropriate level of detail. The explicit information provided by the Remos system is via a special *Application Programming Interface* (API) [LMSGSS98]. For example, clients can specify the information (such as bandwidth or delay) and the nodes they are interested in by sending a Remos query through the API. The Remos response will be a logical topology, which is a set of switch nodes and links annotated with bandwidth or delay, as shown in Figure 2-13.

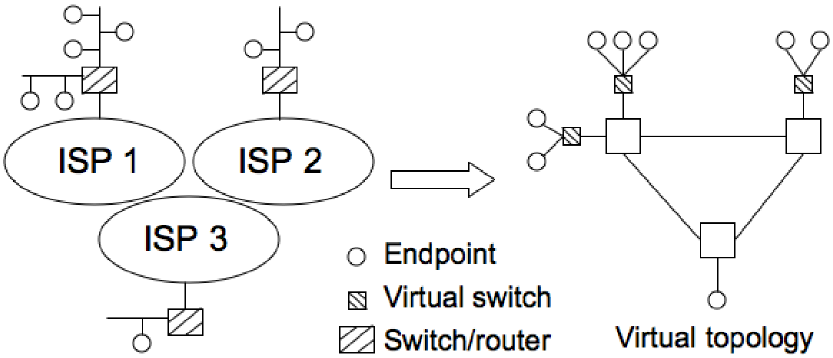


Figure 2-13 Example of a Logical Topology [MS00]

The services provided by the Remos system are basically based on two types of elements [DGKLMSS01]: collectors and modellers. Figure 2-14 presents the architecture of the Remos system.

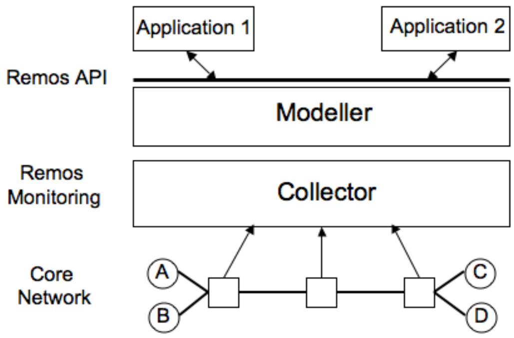


Figure 2-14 Remos Architecture [MS00]

The Remos modellers provide the APIs to the applications, and answer the queries received from the applications. The Remos modeller communicates with the Remos collector to obtain the required information. Instead of directly transmitting the information back to the clients, the modellers refine and convert the information into the form that the applications are interested in. Furthermore, if a network behaviour prediction is necessary, the modeller also can use a toolkit, *Resource Prediction System* (RPS) [DGKLMSS01][Din06], together with the measurement history, to provide prediction services. The Remos collectors are responsible for network information collection and integration. The Remos collectors can be organised into three levels: local collectors, global collectors, and master collectors. The local collectors are responsible for collecting information about specific local area networks. The global collectors are responsible for collecting information about the networks connecting to the local area networks. The local and global collectors obtain network information by using two methods: *Simple Network Management Protocol* (SNMP) or explicit benchmarking. The master collectors communicate with a Remos modeller. After receiving a request from the modeller, the master collector queries the appropriate local and global collectors, gathers and combines the information from the collectors into a response to the Remos modeller.

[MS00] presents results of some experiments using the Remos system, such as bandwidth measurements, timing measurements, and prediction accuracy measurements. The results show that the information provided by the Remos system is accurate. The Remos system has several advantages. Firstly, it is efficient because several clients can share the same information. Secondly, the information presented by Remos is network-independent so that clients can be less concerned to the network details. Finally, it moves the task of collecting network information to the network side so as to simplify the task of the service developers.

Remos is a network information collection and behaviour prediction system for distributed applications. Besides Remos, there are other projects focusing on this field, such as Prophet [WZ98] and Network Weather Service [WSP97][SW02]. The Prophet project and the Network Weather Service project concentrate on predicting the availability and performance of the network resources. In order to obtain accurate predictions, these two projects implement benchmark-based measurement by exchanging messages between different computing nodes in the network.

2.2.2.2 User Terminal Information

In this work, the context information about user terminals mainly includes user preference and user terminal location.

For the information about user preference, the *Graphical User Interface* (GUI) is one of most popular choices for information acquisition. In [E-KLGGB03], a HTML frame based GUI is presented. This GUI not only acts as a service advertisement, which describes the available services from the service providers, but it also can be used to configure the user preferences so that a suitable service is chosen for the user.

In [IMMHM04], a communication management software on the user terminal is presented and called the *seamless client application*. This software uses a GUI (shown in Figure 2-15) to interact with the users. The GUI can present the users with a list of available RANs and allow the users to input their preference. Based on the information about available RANs and user preference, the communication management software can select an appropriate RAN at the communication start up phase (in Figure 2-15, the IEEE 802.11b WLAN is selected).

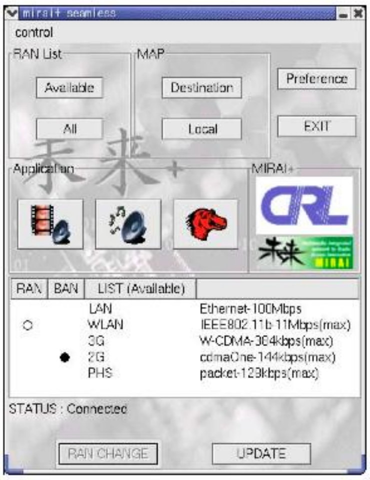


Figure 2-15 GUI of the communication management software [IMMHM04]

User terminal location is another important context information to the RAN selection and optimisation. By combining the terminal's location with the knowledge of the networks' coverage areas and the user's moving history, the RAN selection and optimisation system can determine the RANs along the user's path, reserve sufficient resources, and coordinate the handover operation in advance. In outdoor environments, two positioning methods are considered [HT07]: Cell Coverage based positioning and *Global Positioning System* (GPS) based positioning.

The Cell Coverage based positioning method is network-based and does not require new functionalities in the user terminal. The accuracy of the user location is of cell level and it can be obtained when the user terminal has a dedicated channel or is in *cell_FACH* or *cell_PCH* states [HT07]. When the user terminal is in the idle state, its location can be acquired by forcing it to *cell_FACH* state with a location update [HT07]. The accuracy of the location information depends on the cell size.

The GPS-based positioning method requires the user terminal to have an integrated GPS receiver, which can measure and provide more accurate location information. With a reference GPS receiver situated in every base station, the network also can assist the measurement by providing additional information, such as visible GPS satellites, reference time and Doppler [HT07]. Based on the assisted GPS measurement, the accuracy of the location information can reach 10 meters outdoor [HT07].

The Cellular Coverage based positioning and the GPS-based positioning are complementary. By combining them together, accurate location information can be obtained in the outdoor environment. However, in the indoor environments, these two technologies seem unsuitable. The obstacles inside the building, such as walls, doors, furniture, equipments, and even human bodies, may affect electromagnetic wave propagations and introduce multi-path effects which will worsen the performance of the Cellular Coverage based positioning. The building walls also block the line-of-sight transmission between the GPS receiver and the satellites and make the GPS service limited. As a result, it is necessary to implement an indoor positioning system. Nowadays, different technologies are developed and being researched to provide indoor positioning services, including *Infrared* (IR) signals, ultrasound waves, radio frequency, electromagnetic waves, vision-based analysis, and audio sound [GLN09]. Indoor location information can be obtained by carefully studying the specific indoor environment and selecting and combining appropriate positioning technologies.

2.2.3 Session Initiation Protocol

As described in the previous sections, the context information can be collected from entities in different network layers and user terminals. However, the efficient delivery of the context information is another challenge. Considering the heterogeneity of the networks and user terminals, context delivery protocols should be efficient, programmable, robust, and network independent. The *Session Initiation Protocol* is a possible protocol for context information delivery.

The *Session Initiation Protocol* (SIP) [SJ01] [Joh04], which is developed by the IETF, is a general-purpose application-layer control or signalling protocol for session establishment and termination. It is independent of the underlying transport layer protocol and without any dependencies on the type of the session. An SIP-enabled communication network basically consists of SIP endpoints and SIP servers. SIP endpoints are also called SIP User Agents, which can be located in user terminals or network gateways. SIP is working on a client-server transaction model similar to HTTP. The session management is performed basically by exchanging SIP requests and responses between the SIP User Agents. Each User Agent is composed of a *User Agent Client* (UAC), which initiates and transmits SIP requests, and a *User Agent Server* (UAS), which receives SIP requests and returns responses.

SIP servers assist SIP User Agents to establish communication sessions. Basically there are four types of SIP servers: SIP Proxy, Redirect Server, Register Server, and Location Server. The SIP Proxy performs a routing functionality and forwards an SIP request to another entity closer to the target SIP User Agent. The Redirect Server receives an SIP request and responds with a redirection indicating an alternative target User Agent. The Register Server receives registration request from a User Agent and updates the User Agent's information in a location server and other databases. The Location Server is a database that records user identities and location information. User Agents communicate with the Location Server indirectly through SIP Proxies or other servers.

In order to establish a communication session, an SIP client initiates an SIP request and transmits it to the network side and waits for the response from the target User Agent or other SIP servers. Each SIP request comprises two parts, a header and a body. The header part consists of parameters including the identity of the caller, the identity of the receiver, a unique call ID, a sequence number, subject, the hop traversed path to deliver the message, etc. So every SIP message has enough routing and session status information to be delivered to its destination. The body part of a SIP message typically makes use of the *Session Description Protocol* (SDP) in order to describe the session that is being negotiated. There are several key features of SIP and they are listed as follows [Wis01]:

- **Simplicity:** SIP is designed as a lightweight protocol which means that it can interoperate with just four headers and three request types. It can run on very thin devices such as pagers.

- **Generic session description:** SIP separates the signalling of a session from the description of the session. SDP is not compulsory and SIP can be utilised to establish and control a new type of session.

- **Modularity and extensibility:** SIP is extensible and allows implementations with different features to be compatible. 3GPP specified SIP as the means for supporting IP multimedia applications, such as video telephony, in 3GPP UMTS release 5 onwards.

- **Programmability:** A SIP server offers the possibility of running scripts or code that can change, re-direct or copy SIP messages. Therefore SIP servers can provide *Intelligent Network* (IN) services for voice oriented networks, but it also can be extended to provide intelligent control of advanced multimedia services.

- **Integration with other IP component technologies:** Like HTTP, SMTP, and *Boundary Gateway Protocol* (BGP), SIP can be integrated with IP protocols, such as *Real-Time Streaming Protocol* (RTSP), to provide voice mail service and even enable video streaming service during a video conference session.

In [LL02], Latvakoski and Laurila describe an application based access system selection in an all-IP environment. An SIP REGISTER message is adopted to deliver the information about the initial access system which is being used by the user terminal during the negotiation between the user terminal and the IMS entity.

BT also presents in a paper their experience in using SIP to provide virtual home environment services [Wis01]. Wei Li [Li03] presents an infrastructure for providing adaptive and context-aware wireless services, which relies on SIP and its sibling protocols (such as *SIP for Instant Messaging and Presence Leveraging Extensions* (SIMPLE)) to transfer XML encoded context information.

2.3 Concluding Remarks

This chapter presents the challenges related to context-aware and intelligent RAN selection and optimisation in heterogeneous communication environments.

The difficulties in achieving effective cooperation between RANs are discussed. Section 2.1 presents an insight of how cooperation can be realised between RANs by describing six network integration schemes and a mobile independent handover standard. IP-based network is commonly proposed as a means of interworking different RANs.

In order to perform a context-aware and optimised RAN selection, context information plays an important part. Section 2.2 studies the challenges in using context information to enhance the RAN selection and optimisation. These challenges include context information definition, description, acquisition, and delivery.

The next chapter presents decision-making mechanisms proposed for intelligent RAN selection algorithms.

Chapter 3 Radio Access Network Selection Algorithms

The use of metrics presented in Chapter 2 increases the complexity of RAN selection algorithms. RAN selection algorithms solely based on radio signal strength cannot be used in complex heterogeneous communication environments. Different RAN selection algorithms have been proposed in the literature and they can be classified into three categories: cost/utility function based algorithms, multiple criteria based algorithms, and computational intelligence based algorithms.

3.1 Cost/Utility Function based Algorithms

These algorithms use cost/utility functions to measure the benefits from connecting to each candidate RAN. The candidate RAN that produces the highest utility or lowest cost, is selected by the algorithms.

In [ZM06], Zhu and McNair propose a *Multiservice Vertical Handoff Decision Algorithm* (MUSE-VDA) based on a cost function. They use the cost function to evaluate the cost of receiving each of the requested services from every network covering the user. The network with the lowest cost is selected to handover to. Assuming network *n* is being evaluated, the cost introduced by network *n*, C^n, can be calculated as follows [ZM06]:

$$C^n = \sum_s C_s^n \tag{3.1}$$

s is the index representing the user's requested service and C_s^n represents the per-service cost for network *n*. C_s^n can be calculated as [ZM06]:

$$C_s^n = \sum_j W_{s,j}^n \times Q_{s,j}^n \tag{3.2}$$

$Q_{s,j}^n$ is the normalised QoS provided by network *n* for parameter *j* of service *s*. The parameter can be the data rate provided by the network, the network delay, the power consumption, and so on. In [ZM06], Zhu and McNair present an example to normalise the value of parameter *j* by using a logarithmic function. $W_{s,j}^{n}$ is the weight specifying the importance of parameter *j* to service *s*. For example, the user requests a service with a specified minimum delay and minimum power consumption requirements. For a user terminal that has a low battery life, the power consumption requirement is of greater importance than the delay requirement. A higher weight for power consumption may direct the user to a network where the user terminal consumes less power. By combining equations 3.1 and 3.2, the general form of the cost function is [ZM06]:

$$C_s^n = \sum_s \sum_j W_{s,j}^n \times Q_{s,j}^n \qquad \text{s.t. } E_{s,j}^n \neq 0, \qquad \forall s,j \tag{3.3}$$

$E_{s,j}^n$ is a *Boolean* value called the network elimination factor. It indicates whether the QoS constraint *j*, such as the minimum data rate, of service *s* can be satisfied by the network *n*. If the constraint can be satisfied (for example, the data rate provided by network *n* is equal to or greater than the minimum data rate required for supporting service *s*), $E_{s,j}^n$ will be equal to 1 and network *n* can be evaluated. Otherwise, the network *n* cannot be evaluated.

In [CLH04], Chen et al. propose a utility function based vertical handover network selection algorithm. They use a function to evaluate the quality of receiving each of the requested services from every network covering the user. The quality is represented by the result of the function: utility. The greater the utility is, the better quality can be perceived. The network, which produces the greatest utility, is selected to handover to. Assuming network *j* can satisfy the constraints of the requested service, its utility function, $Utility_j$, can be calculated as [CLH04]:

$$Utility_j = \sum_i w_i \times f_{i,j} \tag{3.4}$$

i is the index representing the parameters of the requested service and w_i is the weight specifying the importance of parameter *i* to the requested service. $f_{i,j}$ represents the QoS provided by network *j* for parameter *i* of the requested service. The value of $f_{i,j}$ is normalised to range from zero to one. When the value for parameter *i* provided by network *j* is equal to or less than the minimum requirement, the value of $f_{i,j}$ is zero. When the value for parameter i is equal to or greater than the maximum requirement, the value of $f_{i,j}$ is one. The normalisation process can be summarised as follows [CLH04]:

$$if\left(LB_i < R_{i,j} < UB_i\right) \Rightarrow f_{i,j} = \frac{R_{i,j} - LB_i}{UB_i - LB_i},$$
$$if\left(R_{i,j} \le LB_i\right) \Rightarrow f_{i,j} = 0,$$
$$if\left(R_{i,j} \ge UB_i\right) \Rightarrow f_{i,j} = 1$$

UB_i and LB_i are the upper bound and the lower bound requirements for the parameter i, respectively. $R_{i,j}$ is the real value for the parameter i provided by the network j.

In [OMM07], Ormond et al. take into account the transfer of non-real-time data files and present a study on designing the corresponding network selection strategy. The network selection strategy is user-centric and implemented in the user terminal. Ormond et al. propose a solution which is not only able to describe the user's preference on the service metrics, but also it is able to predict the throughput in each of the candidate RANs. Ormond et al. take advantage of microeconomic principles to enable the network selection decision to maximise the user's interests in a dynamic radio environment.

The network selection strategy proposed by Ormond et al. is characterised by its consideration towards the charging schemes employed by the candidate RANs. In Ormond's setting, each of the candidate RANs implements a set of price-per-byte charging scheme and differentiates between other rivals by charging a different price. Every user wishes to get the best value for his/her money by timely transferring the non-real-time data file at the lowest monetary cost. The network selection strategy is based on a utility function, which describes the relation between the user's willingness to pay (also it can be understood as the value of the service to the user) and the time spent in completing the file transfer. Generally speaking, the longer the transfer completion time is, the less the user is willing to pay. More precisely, the shape of the utility function depicts the user's preference on the transfer completion time and the monetary cost. In [OMM07], Ormond et al. present the design of the utility function, which is plotted in Figure 3-1.

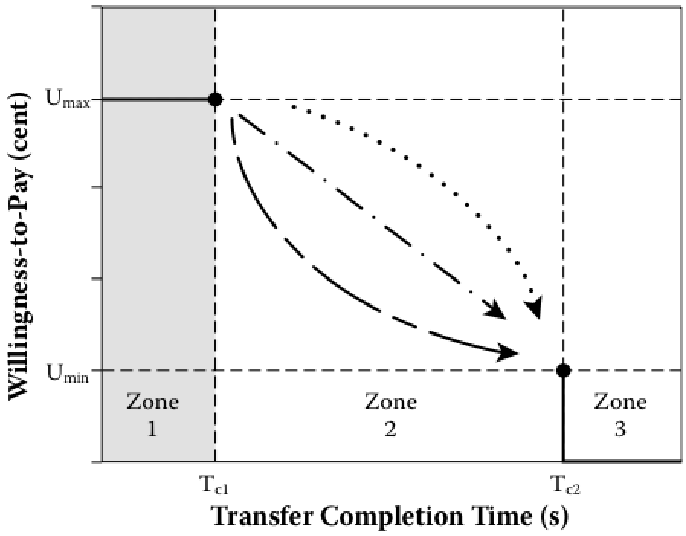


Figure 3-1 Design of Utility Function [OMM07]

The value of the utility function, $U_i(T_c)$, describes the user's willingness to pay in cents for transferring a file *i* in the completion time T_c. If the file *i* can be transferred within T_{c1}, the user is satisfied and willing to pay the highest price, U_{max}, for the service. The user is in the satisfaction zone, which is denoted as Zone 1 in Figure 3-1, and T_{c1} is threshold of the satisfaction zone. If the transfer completion time exceeds T_{c1}, the user's willingness to pay begins to decrease and drops to the lowest value, U_{min}, as the completion time reaches T_{c2}. In such case, the throughput in the RAN still meets or exceeds the minimum bandwidth threshold and the user is still tolerant of the service quality. The user is in the tolerance zone, which is denoted as Zone 2 in Figure 3-1. However, any transfer completion time longer than T_{c2} (Zone 3 in Figure 3-1) indicates that the throughput is below the minimum bandwidth threshold and leads to an unsatisfied user who is not willing to pay any money for the service.

The shape of the utility function in Zone 2 depends on the user's preference on the transfer completion time and the monetary cost. In [OMM07], Ormond et al. study three types of users, risk seeking, risk averse, and risk neutral, and plot the shapes of the utility functions for each user type in Figure 3-1. The risk seeking users are delay-sensitive and willing to pay more for shorter transfer completion time. The utility function for the risk seeking users has a concave curve in the tolerance zone. The risk averse users are cost-sensitive and willing to tolerate longer transfer completion time for less monetary cost. The utility function for the risk averse users presents a convex curve in the tolerance zone. The risk neutral users consider the relation between the transfer completion time and the monetary cost in an unbiased way. In the tolerance zone, the willingness to pay linearly decreases as the transfer completion time increases, and the shape of the utility function is a straight line. Ormond et al. present a solution to define the utility function for the risk neutral user as [OMM07]:

$$U_{1i}(T_c)=\begin{cases} U_{\max} & T_c < T_{c1} \\ \dfrac{(U_{\max}-U_{\min})\times(T_c-T_{c1})}{T_{c2}-T_{c1}} & T_{c1} < T_c < T_{c2} \\ 0 & T_c > T_{c2} \end{cases} \tag{3.5}$$

Furthermore, Ormond et al. also implement a microeconomic term, consumer surplus, in designing the network selection strategy. The consumer surplus is the difference between the value of the utility function (the value of the service to the user) and the actual price charged by the candidate RAN. The greater the consumer surplus is, the more satisfied the user is. The goal of the network strategy is to maximise the consumer surplus and the RAN that can produce the greatest consumer surplus will be selected.

Cost/Utility based algorithms provide simple and efficient solutions for network selection. Studies are continuously carried out in [LWZ07] [LJC08] [SZ08], etc.

3.2 Multiple Criteria Algorithms

The context information discussed in Chapter 2 makes the RAN selection a multiple criteria decision-making problem. *Analytic Hierarchy Process*, *Grey Relational Analysis*, and *Multiple Attribute Decision Making* are some of the techniques proposed to solve multiple criteria decision-making problems.

3.2.1 Analytic Hierarchy Process based Algorithm

Analytic Hierarchy Process (AHP) is a structured technique to decompose a complicated decision-making problem into a hierarchy of simpler and more easily comprehended sub-problems. The sub-problems are also called decision factors and each relates to or represents one aspect or attribute of the decision-making problem. The decision factors are weighted based on their relative importance between each other with respect to the decision-making problem. The hierarchy of AHP also includes a series of solution alternatives. AHP calculates the synthesised weights for each solution alternative and selects the solution which has the greatest weight.

The use of AHP can be summarised into five steps. The first step is to model the decision-making problem into a hierarchy containing decision goal, decision factors, and solution alternatives. For example, a user is trying to make a selection among three networks for making a voice call. The user criteria include Cost, Network Coverage Area, and Voice Quality. This problem can be decomposed into the hierarchy shown in Figure 3-2.

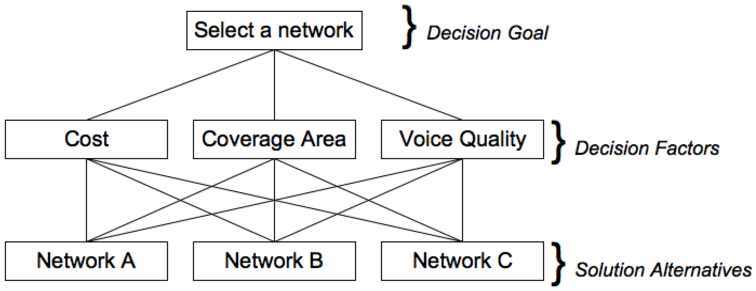


Figure 3-2 Hierarchy of the Decision-Making Problem [SJ05]

The decision goal is the objective of the decision factors. Each decision factor is an objective of the solution alternatives. The decision factors and solution alternatives are also called elements.

The second step is to establish the priorities among the elements at the same level of the hierarchy. The user judges the relative importance between the elements by making a series of pairwise comparisons. The judgements are converted into numerical values on a scale of 1 to 9 based on the contribution intensity of the elements to the objective. The scale is shown in Table 3.1.

Intensity	Definition	Explanation
1	Equal importance	Two elements equally contribute to the object
3	Moderate importance	Slightly prefer one element to another
5	Strong importance	Strongly prefer one element to another
7	Demonstrated importance	Very strongly prefer one element to another
9	Extreme importance	Extremely prefer one element to another
2, 4, 6, 8	Intermediate values	

Table 3.1 Scale of relative importance for pairwise comparison [SJ05]

The comparison results can be presented in a reciprocal matrix. Given the decision factors in Figure 3-2, the comparison matrix is shown in Figure 3-3. The matrix element m_{ij} on the *i*th row and the *j*th column illustrates the result of the comparison between decision factors f_i and f_j. For example, the value of matrix element m_{12} is 3, which means that, comparing Cost to Coverage Area, the user considers Cost slightly more important than Coverage Area when choosing a network. Correspondingly, the matrix element m_{21} represents the judgement of the importance of Coverage Area to Cost. m_{21} and m_{12} are correlated and the value of m_{21} can be derived as m_{21} = $1/m_{12}$. The diagonal elements of the matrix present the results of decision factors' self-comparisons and their values are 1.

	Cost	Coverage Area	Voice Quality
Cost	1	3	2
Coverage Area	1/3	1	2/3
Voice Quality	1/2	3/2	1

Figure 3-3 Comparison Matrix of the Decision Factors [SJ05]

The third step is to synthesise the comparison results and calculate the weights of each decision factor. For example, the weight of decision factor Cost (W_{COST}) with respect to the decision goal can be calculated as:

$$W_{COST} = \left(m_{11} \Big/ \sum_{i=1}^{3} m_{i1} + m_{12} \Big/ \sum_{i=1}^{3} m_{i2} + m_{13} \Big/ \sum_{i=1}^{3} m_{i3} \right) \Big/ 3 \qquad (3.6)$$

Based on equation 3.6, the weights of decision factors Coverage Area (W_{CA}) and Voice Quality (W_{VQ}) also can be obtained.

The fourth step is to examine the consistency of the user's judgements. By calculating the principal eigenvalue $x_{\max}$ of the comparison matrix, the consistency index *CI* can be obtained as [SJ05]:

$$CI = \frac{x_{\max} - n}{n - 1} \tag{3.7}$$

where *n* is the dimension of the comparison matrix. The value of the consistency index is non-negative and the closer it is to zero, the more consistent the user's judgements (represented by the comparison matrix) are. Then, the consistence ratio *CR* is calculated as the ratio of the consistency index *CI* to the corresponding random consistency index *RI*:

$$CR = \frac{CI}{RI(n)} \tag{3.8}$$

where *RI(n)* is the value of the random consistency index when the dimension of the comparison matrix is *n*. If the consistence ratio *CR* is smaller than or equal to 0.1, the user's judgements are considered as consistent and no further processes are required. Otherwise, the user's judgements need to be examined and adjusted.

The values of the random consistency index corresponding to different matrix dimensions are presented in Table 3.2 [SA00]. Each value represents the consistency of a randomly generated comparison matrix, which is a sample of 500 randomly generated matrices based on the scale of relative importance.

n	1	2	3	4	5	6	7	8	9	10
RI(n)	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

Table 3.2 Random Consistency Index [SA00]

The solution alternatives also will be processed in the way the decision factors are processed, including priority establishment, comparison result synthesis and weights calculation, and judgement consistency examination. Assuming the user's judgements are consistent, the weights of each solution alternatives with respect to each decision factor can be obtained. The final step of AHP is to calculate the synthesised weights for each solution alternative. Given the weights for Network A with respect to Cost

($W_{COST}^{N_A}$), Coverage Area ($W_{CA}^{N_A}$) and Voice Quality ($W_{VQ}^{N_A}$), the synthesised weight for network A, W^{N_A}, can be calculated as equation 3.9:

$$W^{N_A} = W_{COST} \times W_{COST}^{N_A} + W_{CA} \times W_{CA}^{N_A} + W_{VQ} \times W_{VQ}^{N_A} \tag{3.9}$$

Finally, the network which produces the greatest synthesised weight will be selected.

3.2.2 Grey Relational Analysis based Algorithm

Grey Relational Analysis (GRA) is a technique to analyse the relationship between discrete sequences. GRA defines a reference sequence representing the ideal situation, which includes the desirable quality entities. Each candidate network is represented by an analysing sequence. The relation between the reference sequence and an analysing sequence is represented by a *Grey Relational Coefficient* (GRC). The greater the GRC is, the closer the analysing sequence is to the reference sequence. The analysing sequence which has the greatest GRC will be considered as the most desirable one and the network which is represented by this analysing sequence will be selected.

The implementation of GRA can be summarised into three steps. The first step is to normalise the data of each analysing sequence. Supposing *n* sequences (X_1, X_2, …, X_n) are compared and each sequence contains *m* elements, for example $X_i = \{x_i(1), x_i(2), ..., x_i(m)\}$. The normalised value of a sequence element, $x_i^*(j)$, can be calculated based on three different expectancies [SJ05]: *larger-the-better*, *smaller-the-better*, and *nominal-the-better*. The use of the expectancies might depend on the characteristics of the considered elements and user expectations. For example, given the network selection problem presented in subsection 3.2.1, the network QoS conditions are listed in Table 3.3. GRA considers each candidate network as a sequence, which comprises elements representing the network's QoS conditions: the values of *Cost*, *Coverage Area* and *Voice Quality*. When normalising the element *Cost*, the expectancy *smaller-the-better* will be applied. When normalising the elements *Coverage Area* and *Voice Quality*, the expectancy *greater-the-better* is applied.

	Network A	Network B	Network C
Cost (dollar/s)	0.2	0.1	0.5
Coverage Area (km^2)	1	1.4	1.8
Voice Quality (star)	2	1	3

Table 3.3 Network QoS Conditions [SJ05]

The second step of implementing GRA is to define the reference sequence X_0. The reference sequence comprises elements with ideal values. For the elements normalised according to the expectancy *greater-the-better*, the ideal values are the greatest normalised values of the considered elements among all sequences. For the elements normalised according to the expectancy *smaller-the-better,* the ideal values are the smallest normalised values of the considered elements among all sequences. For the elements normalised according to the expectancy *nominal-the-better*, the ideal values are the moderate normalised values of the considered elements among all sequences.

The final step is to calculate the GRC. Assuming the GRC of sequence *i* is denoted as $\Gamma_{0,i}$, it is calculated as follows [SJ05]:

$$\Gamma_{0,i} = \frac{1}{m}\sum_{j=1}^{m}\frac{\Delta_{\min} + \Delta_{\max}}{\Delta_i(j) + \Delta_{\max}}$$

where *m* is the number of elements in sequence *i*, $\Delta_i(j)$ is the absolute difference of $x_0^*(j)$ and $x_i^*(j)$, $\Delta_{\max}$ is the maximum absolute difference of $x_0^*(j)$ and the normalised value of element *j* among all analysing sequences, and $\Delta_{\min}$ is the minimum absolute difference of value of $x_0^*(j)$ and the normalised value of element *j* among all analysing sequences. Then, the GRCs of all analysing sequences are compared and the sequence that has the greatest GRC is selected.

3.2.3 RAN Selection Algorithm based on the Combination of AHP and GRA

The GRA provides a solution for the network selection problem. However, it does not consider the user preference over certain aspects or attributes of network selection. In [SJ05], Song and Jamalipour combine AHP and GRA techniques and present a hybrid network selection mechanism. The hybrid network selection mechanism considers two networks, UMTS and WLAN. It collects two types of information which can influence the decision-making. One is the user's preference and requirements on the QoS parameters related to the requested service. The other one is the network performance with respect to the QoS parameters. The selection mechanism applies AHP technique to calculate the weight values over each QoS parameter based on the user's preference and requirements. It also implements GRA technique to decide which network best satisfies the requirements. The combination of AHP and GRA results in modifications to the calculation of GRC, which is presented as follows [SJ05]:

$$\Gamma_{0,i} = \frac{1}{m}\sum_{j=1}^{m}\frac{\Delta_{\min}+\Delta_{\max}}{W_j \times \Delta_i(j)+\Delta_{\max}} \tag{3.11}$$

where W_j is the weight value over QoS parameter *j*.

3.2.4 Multiple Attribute Decision Making based Algorithm

In [Zha04], the problem of selecting a network from a limited number of candidate networks (from various service providers and technologies) with respect to different criteria is regarded as a typical *Multiple Attribute Decision Making* (MADM) problem. The following scenario is used: a user is connected to a WCDMA cell (A_1) and he/she is covered by other three networks A_2, A_3 and A_4. A_3 is also a WCDMA cell but belongs to a different operator. A_2 and A_4 are WLAN hotspots. The user has to make a decision within these four candidate networks for handover. Handover criteria include price, bandwidth, signal noise ratio, sojourn time, seamlessness, and battery consumption, which are denoted as: X_1, X_2, X_3, X_4, X_5 and X_6. The capabilities of each candidate network are shown in matrix 3.12 where sojourn time and seamlessness are represented using linguistic terms.

$$D = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{array} \\ \begin{bmatrix} 10 & 30 & 80 & very\_long & seamless & 0.5 \\ 7 & 40 & 80 & very\_long & very\_bad & 0.5 \\ 1 & 80 & 20 & short & very\_good & 1 \\ 2 & 40 & 40 & short & good & 1 \end{bmatrix} \end{array} \tag{3.12}$$

The user is also supposed to have two running services: voice and file download. For each service, the preference on the handover criteria is modelled as weights allocated by the user. The vector 3.13 w_v is the user preference for the voice service and the vector 3.14 w_d is the user preference for the file download service [Zha04].

$$w_v = [medium \quad medium \quad low \quad low \quad high \quad low] \tag{3.13}$$

$$w_d = [high \quad high \quad low \quad low \quad medium \quad medium] \tag{3.14}$$

The fuzzy data in matrix 3.12 and vectors 3.13 and 3.14 are converted to crisp numbers based on a fuzzy scoring method [CHH92]. Then, classical MADM is applied to determine the ranking order of the alternatives. There are three methods considered in [Zha04], including *Simple Additive Weighting* Method (SAW), AHP, and *Technique for Order Preference by Similarity to Ideal Solution* (TOPSIS). The evaluation shows that the SAW and the TOPSIS produce reasonable results.

3.3 Computational Intelligence based Algorithms

Fuzzy Logic and *Neural Network* are the techniques commonly used in the computation intelligence based algorithms. Fuzzy Logic is part of the *Fuzzy Set Theory* and a form of multi-valued logic. Fuzzy Logic deals with fuzzy variables. Each fuzzy variable has continuous values between 0 (completely false) and 1 (completely true), which describe the degrees of membership to the fuzzy sets for a parameter being considered. Therefore, Fuzzy Logic can be used for reasoning with approximate information.

In [SM07], Stoyanova and Mähönen proposed a multi-criteria vertical handover decision algorithm based on the Fuzzy Logic technique. The algorithm uses a fuzzy logic system with a fuzzifier, a product-inference rule base, and a centre average defuzzifier [SM07]. The algorithm produces membership degrees of a user terminal to the currently serving network and the candidate networks. The handover will be triggered when the membership degrees satisfy a predefined threshold and the network, which produces the greatest membership degree, will be selected to handover to.

The algorithm considers six types of parameters as decision metrics: received signal strength (RSS), signal-to-interference ratio (SIR), cost (C), bit error rate (BER), latency (L), and data transmission rate (TR). The values of these parameters are grouped into a pattern vector *PV*. The pattern vector for the currently serving network is denoted as PV_C and for the candidate network (such as N_i) is denoted as PV_{N_i}:

$$PV_C = [RSS_C, SIR_C, C_C, BER_C, L_C, TR_C]$$

$$PV_{N_i} = [RSS_{N_i}, SIR_{N_i}, C_{N_i}, BER_{N_i}, L_{N_i}, TR_{N_i}]$$

The pattern vectors are fed into the fuzzifier which maps each element (denoted as *P*) of the pattern vectors to a fuzzy variable. This process is also called *Fuzzification* [Neg05]. The fuzzy variable, denoted as *PF*, is a three-tuple illustrating the membership degrees of the element to three fuzzy sets: LOW, MEDIUM, and HIGH, as shown in Figure 3-4.

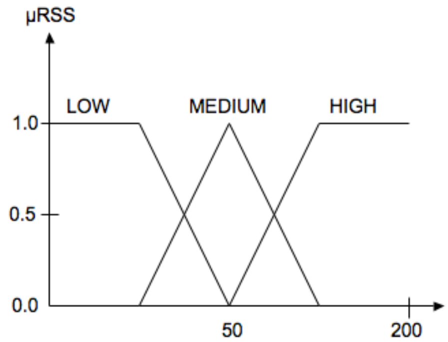


Figure 3-4 Membership Functions of Fuzzy Sets for Received Signal Strength [SM07]

The mapping between an element of a pattern vector and a fuzzy variable can be realised by using a membership functions μ_F. Take the *i*th element P_i of a pattern vector as an example, the corresponding fuzzy variable PF_i can be obtained as:

$$PF_i = \mu_F(P_i) \qquad (3.15)$$

The output of the fuzzifier is a fuzzy pattern vector. The fuzzy pattern vector for the currently serving network is denoted as PV_C^F and for the candidate network (such as N_i) is denoted as $PV_{N_i}^F$:

$$PV_C^F = \left[PF_{RSS_C}, PF_{SIR_C}, PF_{C_C}, PF_{BER_C}, PF_{L_C}, PF_{TR_C}\right]$$

$$PV_{N_i}^F = \left[PF_{RSS_{N_i}}, PF_{SIR_{N_i}}, PF_{C_{N_i}}, PF_{BER_{N_i}}, PF_{L_{N_i}}, PF_{TR_{N_i}}\right]$$

Then, the product-inference rules are applied to the fuzzy pattern vectors. This process is also called *Rule Evaluation* [Neg05]. Because the handover decision algorithm considers six metrics and each metric has three fuzzy sets, the maximum possible number of rules is $3^6 = 729$. The output of rule *l*, which is denoted as y_l, presents the membership degree of the user terminal to the currently serving network or a candidate network under rule *l*.

Then, the *Aggregation* process is applied to unify the outputs of all rules and combine them into a single fuzzy set [Neg05]. Finally, the *Defuzzification* process is carried out to covert the aggregate fuzzy set to a crisp number [Neg05]. In [SM07], the result of the *Defuzzification* process is the average membership degree (denoted as *M*) of the user terminal to the currently serving network or a candidate network. The defuzzifier employed by Stoyanova and Mähönen uses the centroid technique to calculate the centre of gravity of the aggregate fuzzy set, which indicates the average membership degree *M*. The calculation can be presented as [SM07]:

$$M = \sum_{l=1}^{729} y_l \left(\prod_{i=1}^{6} \mu_F(P_i) \right) \Big/ \sum_{l=1}^{729} \left(\prod_{i=1}^{6} \mu_F(P_i) \right) \qquad (3.16)$$

Supposing M_C is the average membership degree of the user terminal to the currently serving network, M_t is the average membership degree threshold, M_h is the average membership degree hysteresis, and M_{N_i} is the average membership degree of the user terminal to the candidate network N_i, the decision rule can be summarised as [SM07]:

IF ($M_C < M_t$) AND ($M_{N_i} - M_h > M_C$), THEN HANDOVER TO N_i

The fuzzy system and decision-making processes, *Fuzzification*, *Rule Evaluation*, *Aggregation*, and *Defuzzification*, also can be represented by a layered structure used in artificial neural networks. In [GAP-RSR08], Giupponi et al. propose a joint radio resource management scheme based on fuzzy neural methodology. The fuzzy neural scheme includes a neural network which has a five-layer structure and enables the use of reinforcement learning to adjust the membership functions. The fuzzy neural scheme is presented in Figure 3-5.

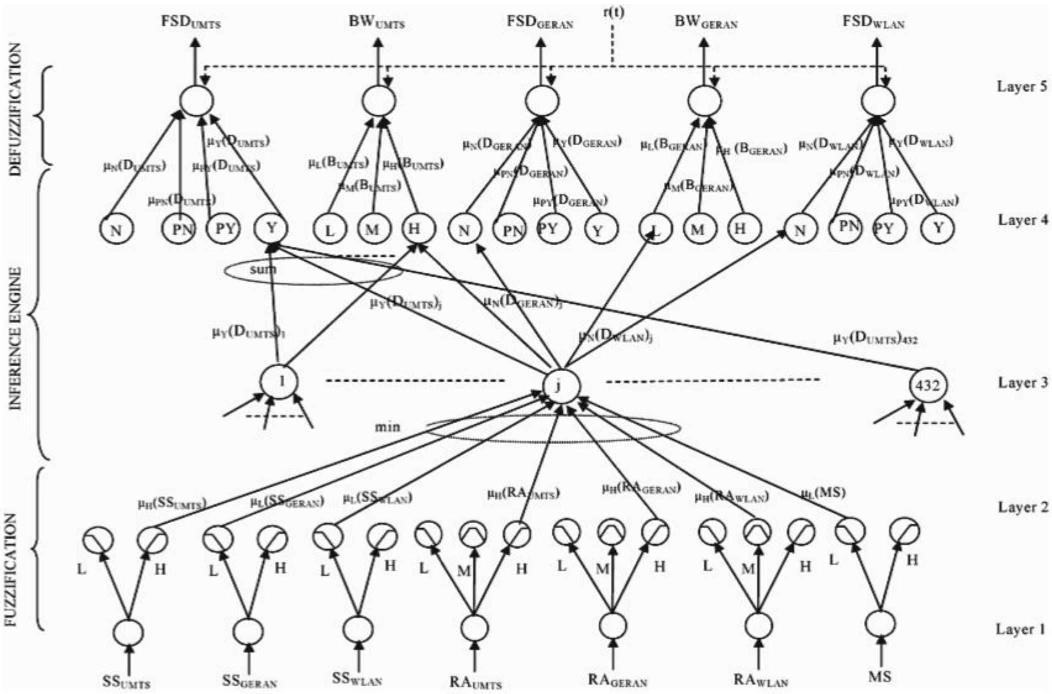


Figure 3-5 Fuzzy Neural Scheme [GAP-RSR08]

Supposing u_i^k is the ith input signal for the kth layer and p is the number of inputs connected to the node, each node combines the inputs by using an integration function $f\left(u_1^k, u_2^k, ..., u_p^k\right)$ and produces the output based on an activation function $a(f)$. Layer 1 of the fuzzy neural scheme has as many nodes as the number of inputs, and the nodes transmit the input values to the next layer. In [GAP-RSR08], seven types of input values are considered. Layer 2 corresponds to the *Fuzzification* process and the outputs of the nodes are membership degrees of the inputs to their fuzzy sets. Layer 3 corresponds to the rule base and each node applies one inference rule. In [GAP-RSR08], each node is linked to seven layer 2 nodes and the links are used to perform the precondition matching of inference rules. The results of layer 3 nodes are the outputs of the inference rules. The layer 4 nodes perform a fuzzy OR operation to aggregate the inputs from layer 3 nodes that have the same inference rule results. Then, the *Aggregation* results are transmitted to the next layer. Layer 5 corresponds to the *Defuzzification* process. In [GAP-RSR08], each node of layer 5 uses the centroid technique to defuzzify the inputs from layer 4 to crisp values which indicate the suitability for selecting each network and the bit rate allocated to each network.

The reinforcement learning scheme for a fuzzy neural system can be divided into two phases: self-organised learning and supervised learning [LL91]. The self-organised learning phase is implemented to define membership functions and set up inference rules. Once the layer structure is defined, the supervised learning phase will be used to optimise the membership functions for desired decision results. In order to initiate the supervised learning, training data will be provided outside system depending on the required output quality. The training data is fed to the system at layer 5 and propagate from up to down so as to adjust specific parameters of the membership function [GAP-RSR08]. The use of reinforcement learning enables online system adjustment and provides a means to design an adaptive decision-making system.

3.4 Adaptivity and Optimisation

In previous sections, algorithms based on different techniques have been presented. These algorithms provide basic and common solution for network selection in heterogeneous communication environments. They are designed to meet individual user's requirement and have less considerations for service adaptivity, network performance and resource utilisation optimisation.

In order to realise service convergence, joint resource management and adaptive quality of service in heterogeneous communication environments, adaptivity, optimisation, and low complexity are essential requirements in designing a successful network selection algorithm. In [JZCCL06], Jia et al. propose a centralised network selection scheme which aims to optimise user distribution among different networks and maximise global spectrum efficiency. Supposing R_n is the required bit rate for the user *n* and b_{nm} is the maximum transmission rate per hertz provided by network *m* for user *n*, the required bandwidth BR_{nm} is calculated as [JZCCL06]:

$$BR_{nm} = \frac{R_n}{b_{nm}} \tag{3.17}$$

Supposing B_m is the overall bandwidth of network *m*, and a_{nm} is the network allocation indicator whose value is 1 if user *n* is allocated to network *m*, or 0 if user *n* is not. The optimal user distribution among the different networks can be achieved by a user allocation scheme which consumes least bandwidth resources [JZCCL06]:

$$\min_{a_{nm}} \sum_{m=1}^{M}\sum_{n=1}^{N} a_{nm} BR_{nm} \tag{3.18}$$

subject to:

$$\sum_{n=1}^{N} a_{nm} BR_{nm} \leq B_m, \forall m .$$

Jia et al. consider the network selection problem as an integer linear programming problem and present a suboptimal algorithm to relax the computational complexity. The algorithm is divided into *N* iterations. Before performing the algorithm, the maximum and minimum required bandwidths for each user are obtained. In each iteration, the user who has the greatest maximum required bandwidth is extracted to select a network. The network, which can provide the minimum bandwidth to support the user's required bit rate, will be selected. The algorithm terminates when 1) all the users have been allocated, or 2) none of the network is available for the user. This algorithm is based on the idea that the network resource utilisation is mostly influenced by the users with higher bandwidth requirements. By selecting a network which provides the minimum bandwidth, the network resource utilisation can be reduced.

In [KAT07], Koundourakis et al. present a network-based algorithm for solving the access and interface selection problem. The proposed algorithm aims to optimise resource allocation while the minimum QoS constraints and thresholds of all users are met. The algorithm receives inputs from three sources. The first source is user data, which provides inputs including a list of visible networks by the user terminal, network signal qualities, and a list of requested services. The second source is the network data, which provides information containing network bandwidths and delays. The last source is the policy decision information, which defines the weight of each decision factor and the QoS and cost thresholds. The result of the algorithm is the allocation of each requested service to an appropriate network.

The benefit of the allocation is represented by the value of an Objective Function, *OF*, which can be presented as [KAT07]:

$$OF(\forall s \in S, \forall ap \in AP) = F(s, ap) + OF(\forall s' \in S, s' \neq s, \forall ap \in AP) \quad (3.19)$$

where *s* is the requested service, *s'* is the allocation of the existing services, *S* is the set containing all services, *ap* is a network access point visible by the user terminal, and *AP* represents the set containing all network access points. Function *F* is defined by two functions: function *Q* and function *PT*. Function *Q* evaluates the quality of receiving the requested service from the network access point *ap*. Function *PT* gauges the compatibility between the requested service and the RAT employed by the network access point *ap*, the degree of trust and the preference for the network operator owning *ap*, and the capability of the RAT employed by the *ap* to serve the user terminal. Function *F* can be presented as [KAT07]:

$$F = w_q Q + w_{pt} PT \quad (3.20)$$

where w_q and w_{pt} are the weights of function *Q* and function *PT*, respectively. Functions *Q* and *PT* can be presented as [KAT07]:

$$Q = w_{bi} BI + w_{di} DI + w_{sqi} SQI \quad (3.21)$$

$$PT = w_{cci} CCI + w_{npi} NPI + w_{tti} TTI \quad (3.22)$$

BI is the bandwidth indicator which estimates the bandwidth sufficiency at the wireless link. *DI* is the delay indicator which evaluates the delay at the link between the access router and the backbone network. *SQI* is the signal quality indicator which shows the quality of the received radio signal. *CCI* is the cost and compatibility indicator which determines the compatibility degree between a requested service and a chosen network, as well as the allocation cost. *NPI* is the network provider indicator which examines the trust degree and user preference to a specific network provider. *TTI* is the terminal type indicator which evaluates the capability of a radio access technology to serve a specific type of user terminal. w_{bi}, w_{di}, w_{sqi}, w_{cci}, w_{npi}, and w_{tti} are the respective weights of the former described factors.

The goal of the proposed algorithm is to maximise the objective function and the process can be summarised into ten steps [KAT07]:

1. Collect all required input information including user data, network data, and policy information.
2. Calculate all possible permutations of the set of services.
3. Select the first permutation item, which is a specific chain of services
4. Select the first service item.
5. For each network access point, calculate the value of the function *F* for providing the selected service. Compare all the values, allocate the selected service to the network access point which can provide the greatest value (denoted as *maxF*) and keep it as the temporary optimum.
6. Reduce the available bandwidth of the selected network access point and change the value of *BI*.
7. Select the next service item and return to step 5. If no service item remains, proceed to step 8.
8. Calculate the total objective function value *OF*, which represents the benefits from the service-to-network allocation, by summing up all *maxF* values. If the new *OF* is greater than current maximum objective function value *maxOF*, *maxOF* will be replaced by *OF* and the service-to-network allocation corresponding to *OF* will be kept as the temporary optimum solution.
9. Return the available bandwidths and *BI*s to initial values. Select the next permutation item and return to step 4. If no permutation item remains, proceed to step 10.
10. Consider current *maxOF* as the maximum objective function value and implement the corresponding service-to-network allocation as the optimum solution.

The objective function maximisation process proposed in [KAT07] tries to reduce the computational complexity and provide an efficient network selection scheme.

3.5 Concluding Remarks

This chapter presents several different decision-making mechanisms for developing RAN selection algorithms. The first three sections introduce cost/utility function based algorithms, multiple criteria based algorithms, and computational intelligence based algorithms. These algorithms relax the complexity in network selection with the consideration of context information. They consider and evaluate network conditions to select a network which best meets the requirements of the requesting user. Some of the algorithms, such as cost/utility function based algorithms, can be implemented at the network side or in the user terminals. Some of the algorithms, such as AHP based algorithms, are better fitted to work in the user terminals. Furthermore, the fuzzy neural scheme even demonstrates a solution enabling online adjustment. All of these algorithms concentrate on serving the requesting user, whereas the existing users and the utilisation of the network resources are not the main focus. However, the admission of a new user would inevitably affect the network resource status of the selected RAN and might impact the interests of the existing users. Without taking into account the existing users and network resource utilisation, the network selection decision might only be able to benefit a limited number of users while the satisfaction levels of some users are compromised and overlooked, and the network resources might be consumed in an ineffective way.

Section 3.4 studies the algorithms that take into account the amount of network resources consumed for supporting the requested service during decision-making. Each algorithm employs a particular network selection strategy which not only considers user requirements, but also optimises network resource utilisation. These algorithms present solutions for realising network cooperation, joint resource management, and adaptive quality of service. However, algorithms employing a single and particular network selection and optimisation strategy or policy are not well fitted to cope with the ever-changing radio environment, traffic load, and network resource availability. A novel algorithm is required, i.e. an algorithm able to manage multiple policies and adaptively implement an appropriate policy according to the current scenario and network conditions.

The next chapter presents the proposed architecture of the intelligent RAN selection and optimisation system.

Chapter 4 System Architecture

This chapter describes the reference architecture of the intelligent RAN selection and optimisation system proposed as a contribution in this thesis.

4.1 The Reference Interworking Architecture and RAN Selection and Optimisation System

In an environment with multiple RATs, the interworking and cooperation between RANs is a challenge. Considering that users can access a packet switched network (e.g. Internet) via different types of RAN, and currently wireless and fixed network systems are increasingly packet-based, a natural solution for the network interworking would be a packet switched, IP-based core network, as for example, the 3GPP's *System Architecture Evolution*. An IP packet presents no difference to the upper layer protocols, when the data link layer protocols have their own ways for addressing. Therefore, an IP backbone network becomes a suitable vehicle for network integration. The IP backbone network can be a dedicated network constructed for the interworking or it can be based on the Internet. For the latter approach, this network can be logically private to the Internet by using some security mechanisms such as IPSec and Virtual Private Network. The motivation for this separation is to reduce the risks of attacks. The intelligent RAN selection and optimisation system is based on the reference interworking architecture shown in Figure 4-1.

The reference interworking architecture is based on an IP backbone network. In Figure 4-1, several types of wireless systems are considered in the interworking architecture, including evolved UTRAN (such as Base Station Routers [BSMPR07]), legacy circuit-switched mobile network, WiFi, and WiMAX. The infrastructures of the above networks are maintained without modifications. The evolved UTRAN, the WiFi and the WiMAX networks can be directly connected to the IP backbone network. In some situations (e.g. for feasibility reasons [SAE08]), a gateway may be introduced between a non-3GPP packet data network, such as WiMAX, and the IP backbone network. RANs of the legacy circuit-switched mobile networks can access the IP backbone network via an access router or a gateway, for instance, a Serving

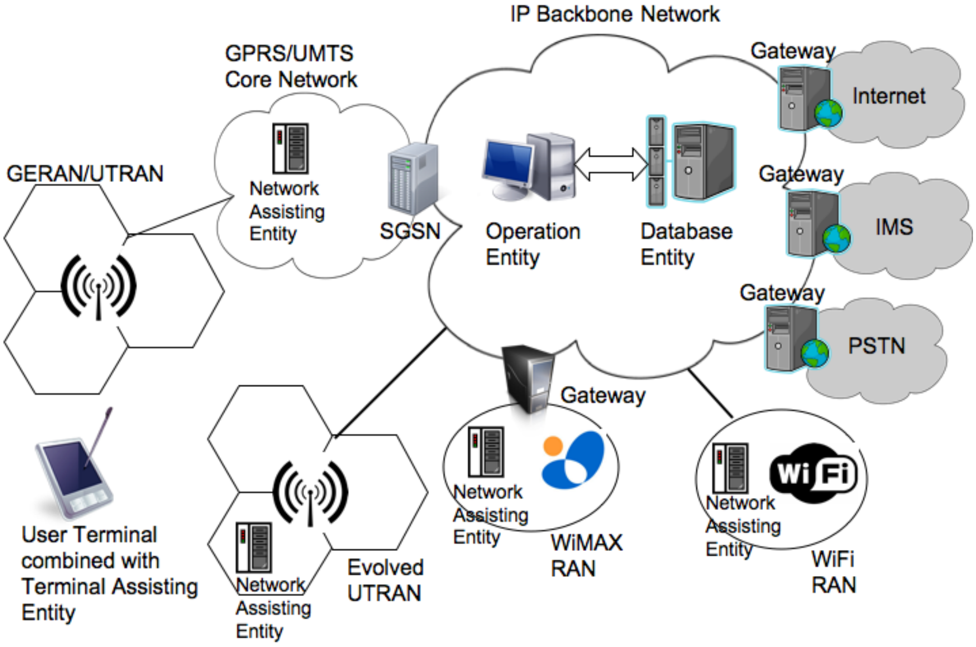


Figure 4-1 Reference Interworking Architecture

GPRS Service Node. The reference interworking architecture simplifies the complexity in network integration and enables diversified and flexible services to fulfil various user requirements [AIPN05].

In order to assist the RAN selection and optimisation system, the reference interworking architecture supports a signalling network which provides a set of physical media for delivering user service request and user context information, as well as decisions from the selection and optimisation system. The signalling network consists of two types of functional components: the *Initial Access Network* and the *Common Signalling Network*. The Initial Access Network provides a basic and ubiquitous medium to the idle users (including the users who just switch on their user terminals) for transmitting their service requests and context information to, and receiving decisions from the RAN selection and optimisation system. The Initial Access Network can help to reduce the need for thorough RAN scanning by the idle users to reduce user costs and terminal power consumption. Considering the importance of the Initial Access Network, the candidate medium is required to have wide area coverage, reliable link, adequate transmission rate, relatively low power consumption, and high popularity. Therefore, the reference interworking architecture proposes the PLMN as the Initial Access Network. For a user who has already been served by a RAN, his/her service request, context information, and the decision from the RAN selection and optimisation system can be exchanged via the existing connection as the serving RAN can act as the Common Signalling Network. The use of the Common Signalling Network provides a straightforward signalling medium for the active users and avoids overloaded signalling traffic over the Initial Access Network.

Based on the reference interworking architecture, the RAN selection and optimisation system is composed of four types of functional entities: the Operation Entity, the Database Entity, the Terminal Assisting Entity, and the Network Assisting Entity. The Operation Entity and the Database Entity coexist and are located in the IP backbone network. The functional configurations and interactions between the entities are shown in Figure 4-2. Both entities working together constitute the so-called *Cooperation Administrator* for the interworking architecture, which is responsible for the context information management and the RAN selection and optimisation. The Terminal Assisting Entity and the Network Assisting Entity interact with the Cooperation Administrator and provide assistance functionalities. The Terminal Assisting Entity is located in the user terminal. The Network Assisting Entity is located in the core and access network of each wireless system. The signalling

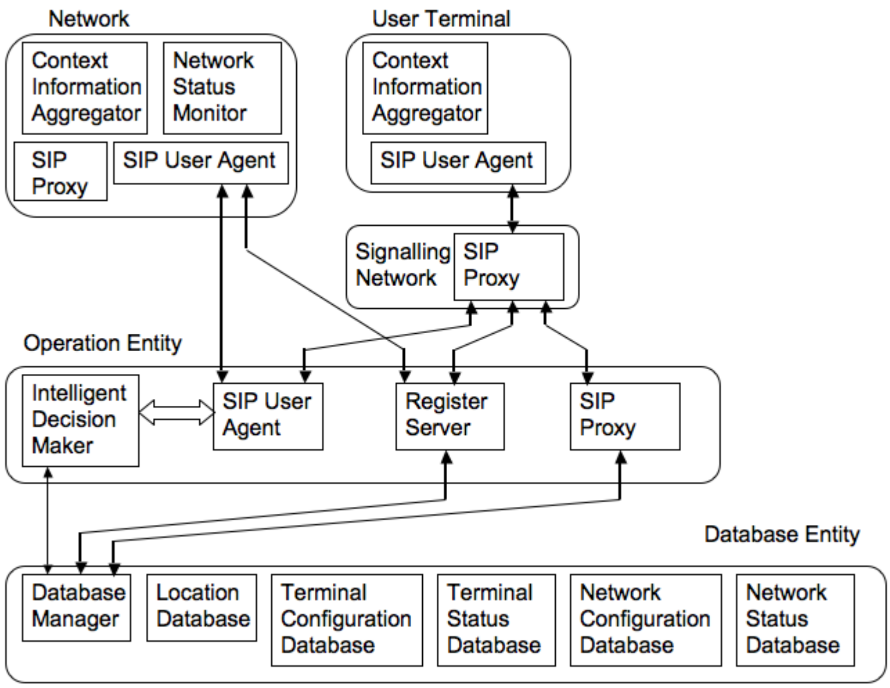


Figure 4-2 RAN Selection and Optimisation System Functional Configuration

network provides a set of physical media for exchanging service request, context information and decisions. However, the diversity of the context information requires a general-purpose and network independent mechanism for the description and packetisation. The SIP protocol is proposed as a solution due to its generic, extensible, and programmable characteristics. Therefore, according to the entities' functionalities, different SIP elements are applied in the RAN selection and optimisation system.

The Operation Entity comprises four modules: an SIP User Agent, a Register Server, an SIP Proxy, and an Intelligent Decision Maker. Based on the SIP elements, the Operation Entity can handle the tasks such as User Agent registration, update information forward, service request reception, reconfiguration request reception, and decision response. The Intelligent Decision Maker adopts an adaptive algorithm to dynamically implement the decision-making policies specified by network operators. It reasons on the context information in order to select the most appropriate RAN for the requesting user. It also provides optimal resource utilisation solutions for certain RANs. The Intelligent Decision Maker can be activated in five situations:

a) User Service Request Initiation

In this situation, a user initiates a service request by sending his/her user ID, user preference, location, and the type requested service via the Signalling Network to the SIP User Agent in the Operation Entity. Then, the SIP User Agent will activate the Intelligent Decision Maker, which in turn, studies the context information stored in the Database Entity and begins the RAN selection and optimisation process.

b) Remote Connection Initiation

In this situation, a user (the callee) is paged to receive a call from a peer user in a remote place (the caller). The SIP Proxy in the Operation Entity receives the connection request and activates the Intelligent Decision Maker. The Intelligent Decision Maker examines the network resource availability to set up the connection, selects the appropriate RAN and if necessary, provides an optimal resource utilisation solution for the target RAN as well.

c) Handover

In the reference interworking architecture, the mobility of an active user is mainly handled by the RRM entity of the serving wireless system. The activation is triggered only in case that the RRM entity of the serving wireless system is unable to cope with the handover and requests cooperation from other RANs. In this situation, the Network Assisting Entity of the serving wireless system transmits a service request to the SIP User Agent in the Operation Entity. The SIP User Agent activates the Intelligent Decision Maker to conduct the RAN selection and optimisation.

d) Service/Resource Reconfiguration

In this situation, the activation is triggered for service/resource reconfiguration purpose, such as load balancing, QoS enhancement, etc. The reconfiguration request is initiated by the RRM entity of a wireless system and transmitted by the Network Assisting Entity. The SIP User Agent in the Operation Entity receives the request and activates the Intelligent Decision Maker to study the context information stored in the Database Entity and carry out the RAN reselection.

e) Policy Alteration

In this situation, the activation is caused by the alteration of decision-making policies. The new policy implemented by the Intelligent Decision Maker may have a different guideline for service quality provisioning and network resource utilisation. For example, in emergency scenarios, the corresponding decision-making policy may require to have more network resources to cope with emergency communications. In this case, the Intelligent Decision Maker will study the context information stored in the Database Entity and perform the RAN reselection following the newly implemented policy.

The Database Entity is responsible for context information management and storage. There are six modules in the Database Entity: a Location Database, a Terminal Configuration Database, a Terminal Status Database, a Network Configuration Database, a Network Status Database, and a Database Manager. The Location Database stores identities and location information of networks and user terminals. The Configuration Databases are in charge of relatively stable information, such as the amount of overall resources of a RAN (e.g. the maximum transmission power of a CDMA base station), network and user terminal capabilities, available access interfaces on a user terminal, and so on. The Status Databases record more dynamic information, such as the amount of available resources of a RAN, network traffic load, network QoS parameters (e.g. interference level, carrier-to-interference ratio, bit error rate), user terminal battery status, RAN in use and active service types on a user terminal, etc. The dynamic information can be updated according to an update period parameter. If necessary, the Database Entity also can directly query context information from the context information collection mechanisms located in networks and terminals through the Operation Entity. The Database Manager administrates the above databases and interacts with the Operation Entity. It receives the context information obtained by the Operation Entity and stores them into appropriate databases. It also receives requests from the Intelligent Decision Maker, queries the databases and responds with the required information.

For reliability and scalability reasons, each Cooperation Administrator covers a certain area and cooperates with a limited number of RANs. If a Cooperation Administrator controls a vast area, it needs to provide services to a large amount of users. This may introduce a considerable volume of data and traffic, especially for signalling, to both the backbone network and the RAN selection and optimisation system. It also requires the Cooperation Administrator with high performance and large capacity. This may introduce reliability and scalability problems. In order to avoid the above problems, a vast area is separated into several sub-areas and one Cooperation Administrator is responsible for one sub-area.

The Network Assisting Entity is located in the core or access network of each wireless system. It consists of four modules: a Network Status Monitor, a Context Information Aggregator, an SIP Proxy, and an SIP User Agent. The Network Status Monitor keeps track of the dynamic parameters of the RAN. When the parameters cannot fulfil a given set of requirements (for example, the carrier-to-interference ratio is below a threshold, or the target RAN cannot support the handover request), the Network Status Monitor will trigger a service/resource reconfiguration request or a service request and transmit it to the Operation Entity. As presented in Chapter 2, different methods and mechanisms are proposed in the research field for context information acquisition. These solutions satisfy the requirements for providing context aware communication services. The reference interworking architecture is assumed to have a mechanism for collecting context information from RANs and user terminals. Therefore, the Context Information Aggregator is responsible for organising the collected context information and providing them to the Cooperation Administrator for service request, information update, or query response purposes. The interactions between the Cooperation Administrator and the Network Assisting

Entity modules are via the SIP User Agent. As discussed before, the reference interworking architecture supports a Signalling Network for exchanging context information, service requests and decisions. Therefore, the SIP proxy situated in the Network Assisting Entity provides the routing functionality to forward the service requests from users to the Operation Entity.

A Terminal Assisting Entity is situated in each user terminal. It includes two modules: a Context Information Aggregator and an SIP User Agent. The Context Information Aggregator functions in the same way as its peer module in the Network Assisting Entity. The SIP User Agent handles the tasks including context information transmission, service request transmission, and decision reception.

Comparing with the related research, such as CRE [DKKTOSPTVE-K03][DVE-KT04], the proposed Operation Entity and the Database Entity are centralised and situated in the IP backbone network. This feature helps to reduce signalling traffic. In CRE, the MS-CRE is responsible for describing network status. As shown in Figure 2-3, after receiving a trigger from a user for network selection, the serving MS-CRE will send the service request to the MS-CRE peers in other networks and wait for replies with the status of other networks. This procedure may introduce overhead in signalling when more than one request is being processed. The centralised Database Entity can solve this problem. Assuming two requests are in process, after making a reply for the first request, the Database Entity can dynamically update the context information and use the information for the second request. This can help to avoid signalling overhead for querying the network status information and decreases the response time as well.

4.2 Signalling and Protocol Message Exchange between a User Terminal and the Cooperation Administrator

In this section, the exchange of service requests, context information and decisions between a user terminal and the Cooperation Administrator are presented. The basic parameters included in the exchange can be:

A) User Identity

B) Update Period

C) User Location

D) Available Access Interface

E) Terminal Capability

F) Requested Service Type

G) User Preference

H) RAN in Use

I) Active Service Type

4.2.1 User Terminal Registration

When a user terminal is switched on, the first step is to register with the Cooperation Administrator. The terminal's SIP User Agent begins this registration procedure by initiating a Register request to the Register Server in the Operation Entity. The Register method is one of six original methods in SIP. It is used to register an SIP User Agent with a location service. The original Register method mainly contains two types of information. The first type of information includes the user's *Uniform Resource Indicator* (URI) and the current contact Uniform Resource Locator (such as the IP address). This type of information is stored in a location server and helps to locate a user terminal while setting up a session. The second type of information is the expiry time which indicates the valid period of the registration. Considering the context aware feature of the proposed system, more information is required for the registration procedure, as shown in Figure 4-3. For the user's preference as an

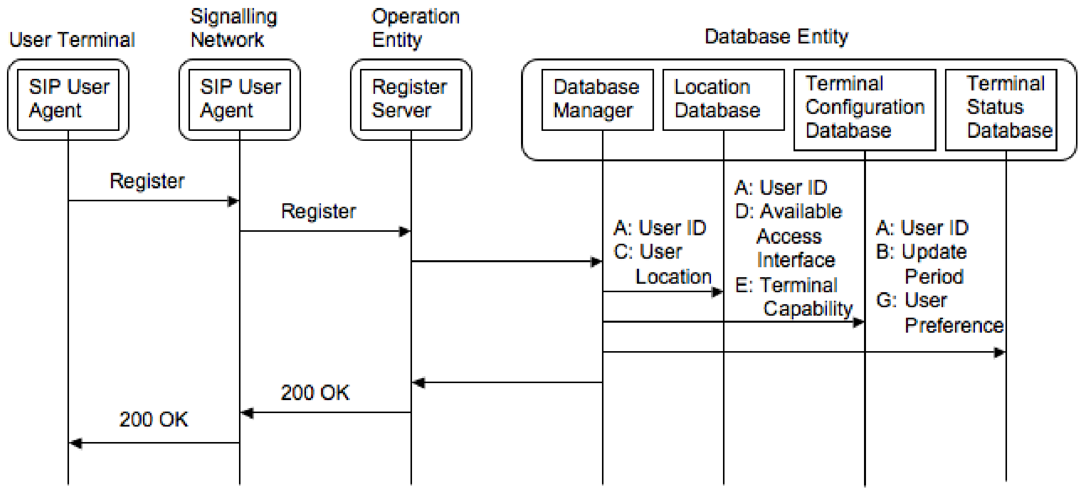


Figure 4-3 User Terminal Registration Signalling

instance, when the user is called by another peer user, the preference information stored in the database acts as the default values for the decision-making. In this work, the protocol used between the Register Server and the Database Manager is not specified in general SIP. Generic database query languages should be able to handle this extended interaction. After the registration procedure successfully finishes, the Register Server sends an OK response to the SIP User Agent in the user terminal.

4.2.2 User Terminal Information Update

A user terminal will update its information periodically or when some change occurs. The terminal SIP User Agent transmits parameters, including location, preference and status information, to the Database Manager in the Database Entity. The Database Manager will update the databases according to the received information. Actually, this update procedure can be regarded as a registration procedure. The difference is that it only updates active and dynamic information, as shown in Figure 4-4.

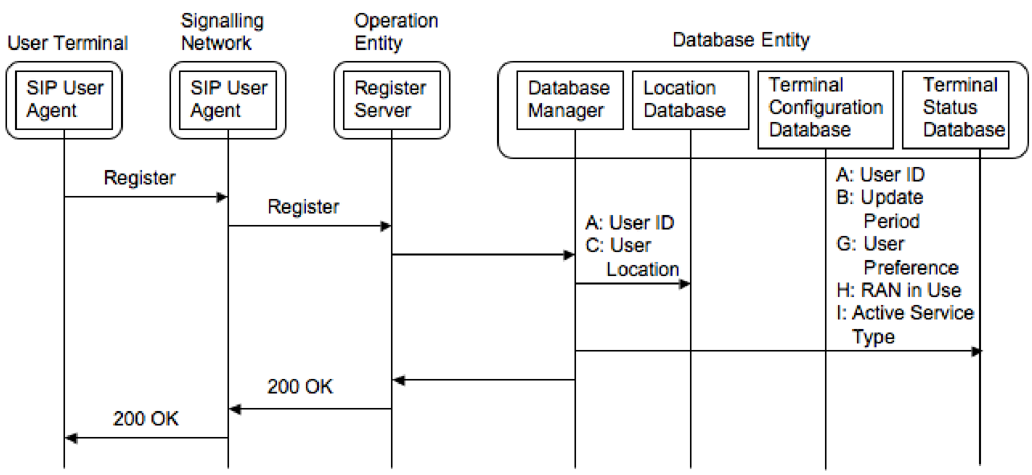


Figure 4-4 User Terminal Information Update Signalling

4.2.3 Service Request

When a user requests a service, the Operation Entity studies user and network context information to select an appropriate RAN for the requesting user and, if necessary, produce an optimal solution for network resource usage. The Terminal Assisting Entity firstly sends an INVITE request message, which contains the information about the type of the requested service, the user location, the terminal capability and the available access interfaces, and if necessary, the user's preference, to the SIP User Agent situated in the Operation Entity. Once the INVITE message is received, the Operation Entity SIP User Agent transmits the service request information to the Intelligent Decision Maker. Meanwhile, the Operation Entity SIP User Agent also sends an OK response back to the Terminal Assisting Entity to indicate that the request was successfully received. The Intelligent Decision Maker analyses the service request and studies relevant context information stored in the Database Entity, such as the amount of overall and available resources in the candidate networks, the network traffic loads, the network QoS parameters, and the network capability. Then, the Intelligent Decision Maker utilises its decision-making algorithm to select an appropriate RAN for the requesting user and, if necessary, produce an optimal solution for reconfiguring resource usage for the target network. After that, the Operation Entity SIP User Agent sends the decision back to the requesting user and, if necessary, the reconfiguration instructions to the target network by using the MESSAGE method [Joh04].

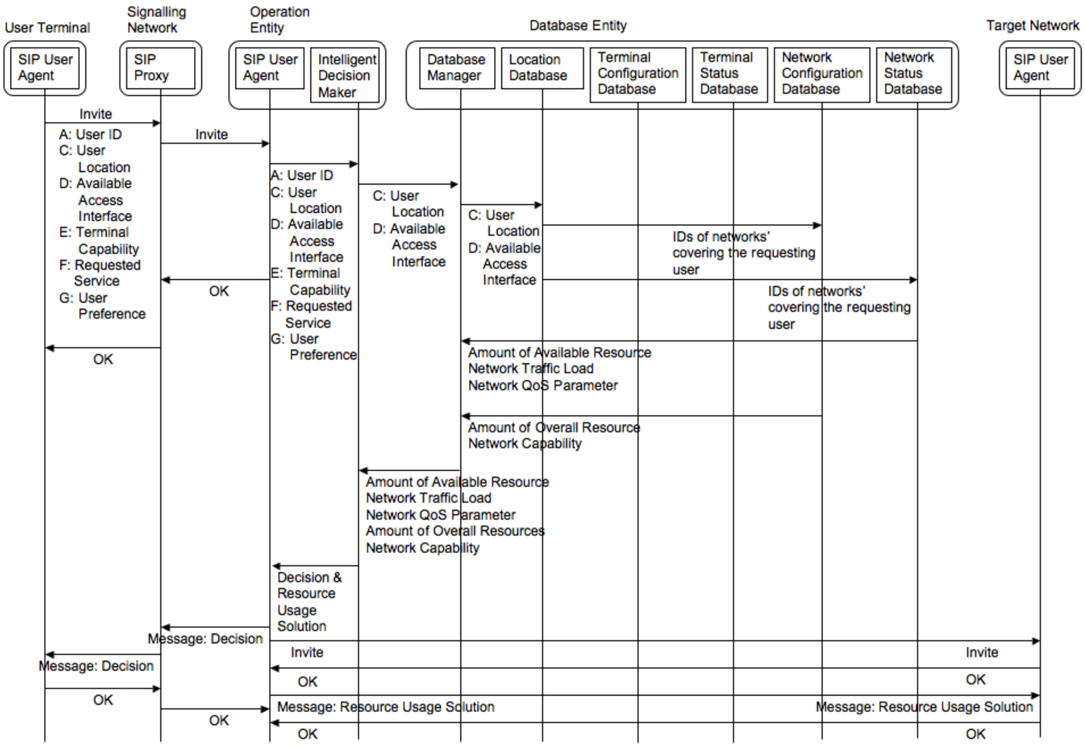


Figure 4-5 Service Request Signalling

4.2.4 Remote Connection Request

When a user (the callee) is paged to receive a call from a peer user in a remote place (the caller), the SIP Proxy in the Operation Entity receives the request. Before forwarding the request to the callee, the Intelligent Decision Maker is activated to examine the network resource availability for setting up the connection, to select the appropriate RAN and, if necessary, to produce an optimal resource usage solution for the target RAN. Then, the Operation Entity transmits the decision and forwards the connection request to the target RAN for establishing the connection. It also transmits the decision to the User terminal to activate the corresponding access interface.

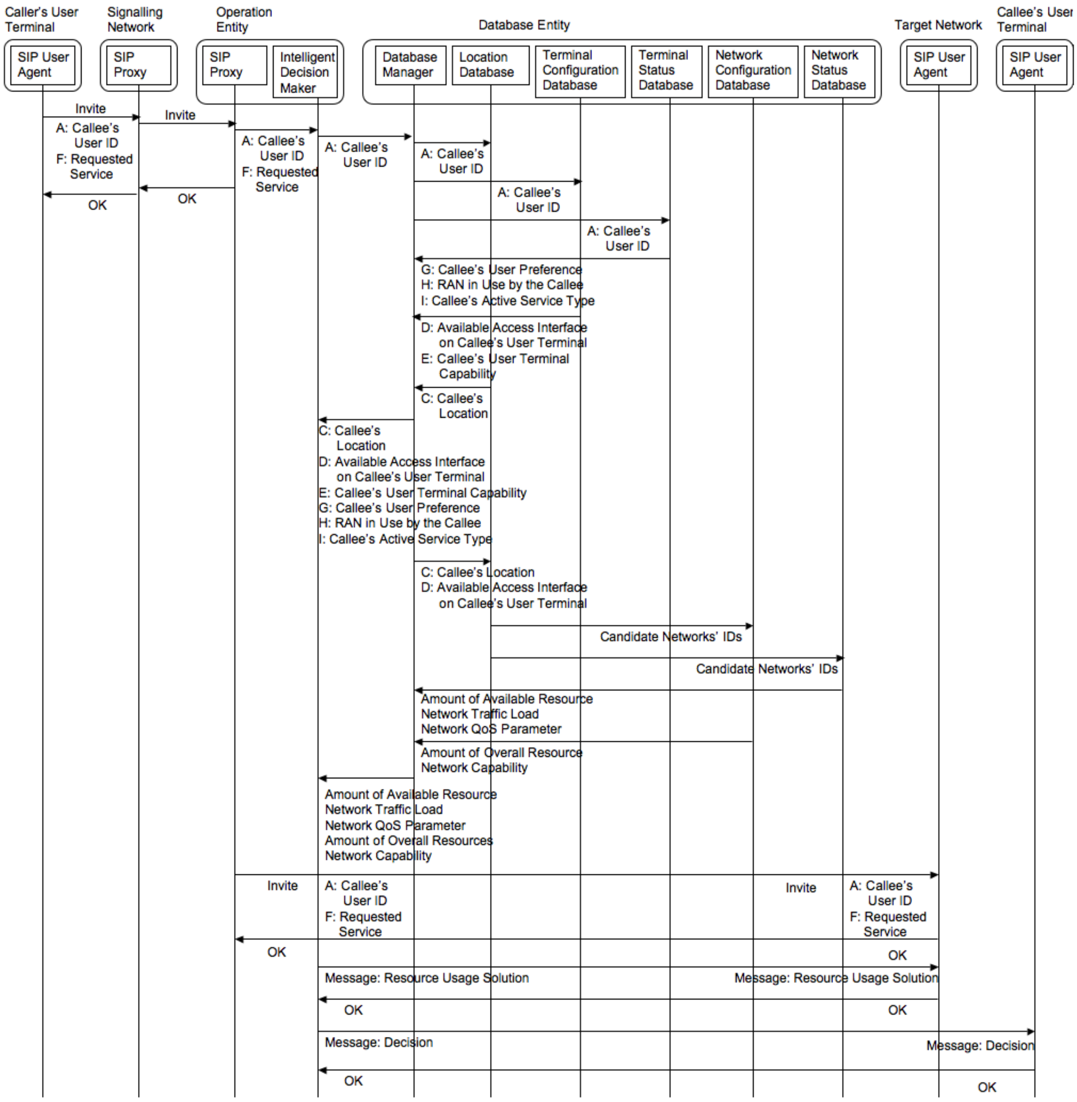


Figure 4-6 Remote Connection Request Signalling

4.3 Signalling and Protocol Message Exchange between a RAN and the Cooperation Administrator

In this section, the exchange of handover requests, service/resource configuration requests, context information and decision between a RAN and the Cooperation Administrator are presented. The basic parameters included in the exchange can be:

J) Network Identity

K) Network Location and Coverage

L) Network Capability

M) Amount of Overall Resources

N) Amount of Available Resources

O) Network Traffic Load

P) Network QoS Parameters (e.g. interference level, carrier-to-interference ratio)

Q) Update Period

4.3.1 Network Registration

After a new deployment or a major change, the network should re-register with the Operation Server. The network registration procedure is similar to the user registration procedure but with different context information, such as network location and coverage, network capability, and the amount of overall and available resources, network traffic load.

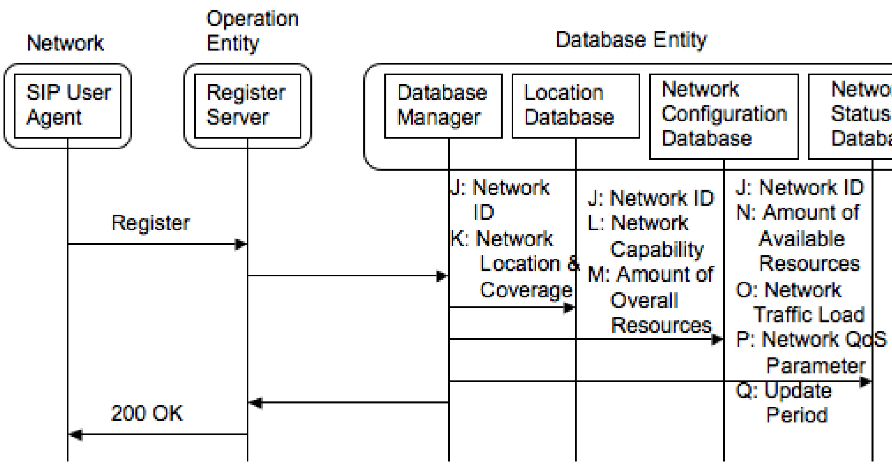


Figure 4-7 Network Registration Signalling

4.3.2 Network Information Update

A RAN will update its information periodically or when some changes occur. The Network Assisting Entity SIP User Agent transmits active and dynamic parameters, including the amount of available resources, network traffic load, and network QoS parameters, to the Database Entity.

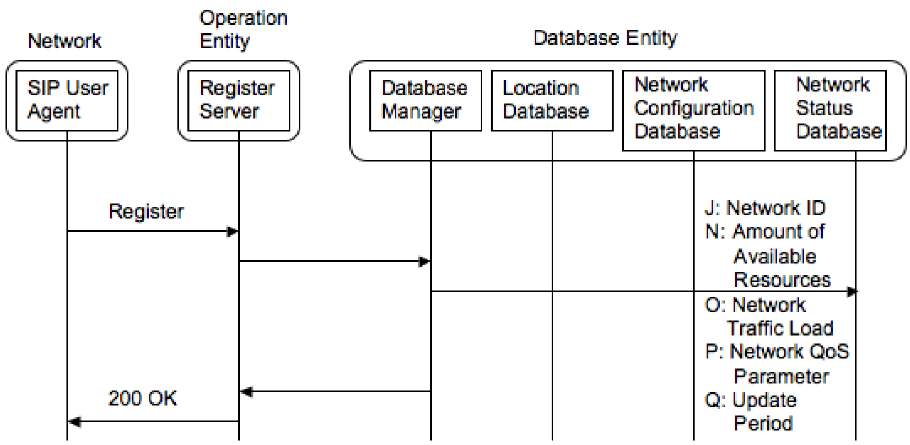


Figure 4-8 Network Information Update Signalling

4.3.3 Handover

The Network Assisting Entity of a serving wireless system transmits a service request to the Operation Entity for the active user who is leaving the coverage area of his/her serving RAN. The service request contains the information about the type of requested service, the user location, the available access interfaces on the user terminal, the terminal capability, and, if necessary, the user's preference. The SIP User Agent in the Operation Entity receives the request, returns an OK response back to the Network Assisting Entity and activates the Intelligent Decision Maker. The Intelligent Decision Maker analyses the service request and studies relevant user and network context information stored in the Database Entity. Then, the Intelligent Decision Maker utilises its decision-making algorithm to select an appropriate RAN for the active user and, if necessary, produce an optimal solution for reconfiguring resource usage in the target network. After that, the Operation Entity User Agent sends the decision back to the Network Assisting Entity and the reconfiguration instructions to the target network by using the MESSAGE method.

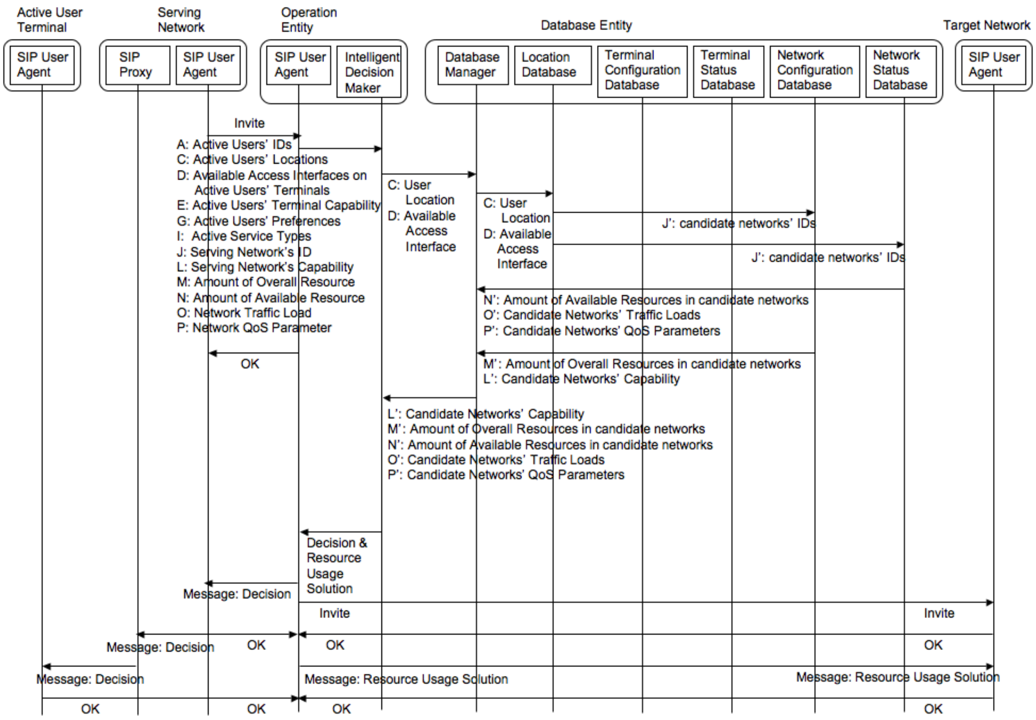


Figure 4-9 Service Request Signalling for Handover

4.3.4 Service/Resource Reconfiguration

The Network Assisting Entity of a serving wireless system transmits a service/resource reconfiguration request to the Operation Entity. The reconfiguration request contains the information about the amount of overall and available resources in the RAN, the network traffic load, the network QoS parameters, the locations of the active users, the active service types, the preferences of the active users, the available access interfaces on the user terminals, and the user terminals' capability. The SIP User Agent in the Operation Entity receives the request, returns an OK response back to the Network Assisting Entity and activates the Intelligent Decision Maker. The Intelligent Decision Maker analyses the reconfiguration request and studies relevant user and network context information stored in the Database Entity. Then, the Intelligent Decision Maker utilises its decision-making algorithm to produce an optimal solution for reconfiguring resource usage in the requesting network. In some cases, the reconfiguration decision may redirect some active users in the requesting network to another RAN. After that, the Operation Entity SIP User Agent sends the decision back to the Network Assisting Entity, certain active users, and the target RANs.

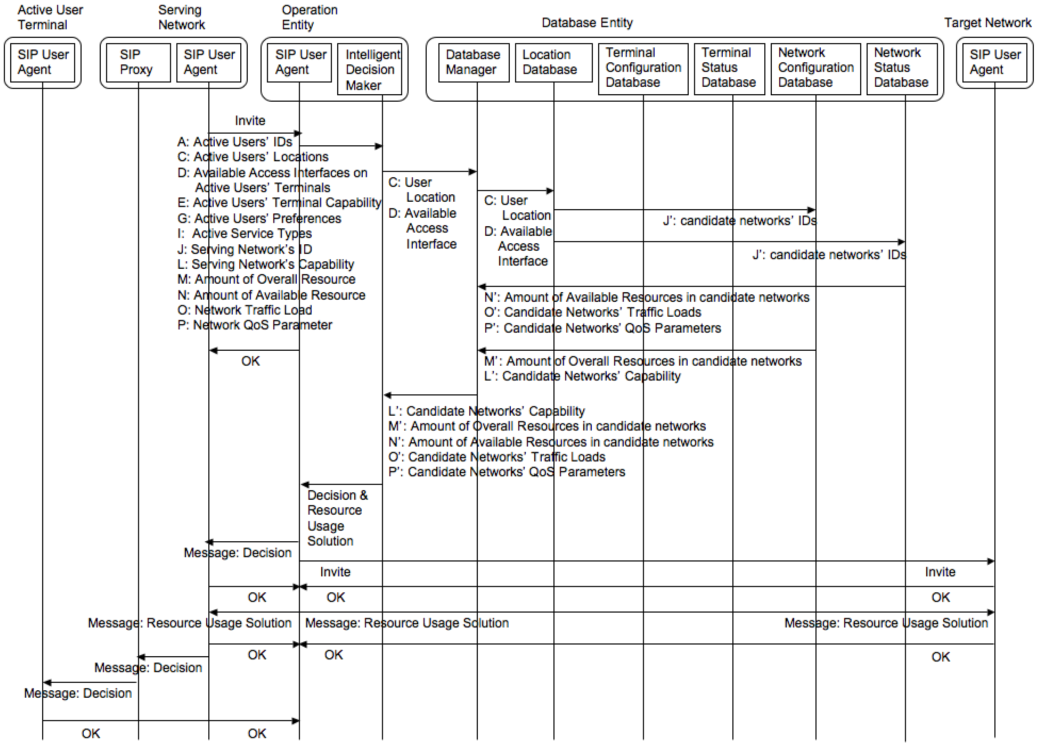


Figure 4-10 Service/Resource Reconfiguration Signalling

4.4 Concluding Remarks

This chapter introduces the architecture of the proposed intelligent RAN selection and optimisation system. It is the base for future research in integration, cooperation, and interworking of heterogeneous wireless systems. The next chapter presents the methods for evaluating network resource availability.

Chapter 5 Network Resource Availability Evaluation Models

The system architecture presented in chapter 4 enables a context-aware RAN selection and optimisation process. Network resource availability is an important attribute of context information. It is dynamic and dependent on the access technology and the service type. In order to develop a context-aware RAN selection and optimisation algorithm and investigate the algorithm's performance, a series of call level simulations were carried out. The call level simulations are based on the resource availability evaluation models of the analysed radio access networks. When a service request arrives, the evaluation model considers the characteristics of the requested service and the access technology, and efficiently obtains the network resource availability information, which can be used by the RAN selection and optimisation algorithm. Currently, two RAN technologies can be simulated: UTRAN and IEEE802.11a/b based WLAN.

5.1 UTRAN Resource Availability Evaluation Model

The UTRAN uses the W-CDMA technology and it is an interference-limited cellular network. The evaluation can be divided into two parts: uplink and downlink. For duplex services, such as speech and video call, both uplink and downlink should be evaluated. For asymmetrical services, e.g. video streaming and web browsing, the downlink traffic plays a major role and the evaluation neglects the uplink traffic and concentrates on the downlink.

5.1.1 Uplink

In the uplink, UTRAN resource availability is determined by two parameters: interference margin and noise rise [HT07].

The interference margin is a predefined parameter in the link budget. In a telecommunication system, a link budget depicts the detail information about all the gains and losses from the transmitter, via the medium, e.g. free space, cable, waveguide, fibre, etc., to the receiver. A link budget considers the attenuation of the transmitted signal caused by propagation, loss, or antenna gain. The calculation of a link budget does not take into account random attenuation, such as fading, which is assumed to be handled by diversity techniques. The value of the interference margin determines the loading of a cell. A typical value for the interference margin in coverage-limited cells ranges from 1 to 3 dB, which corresponds to 20 to 50 percent loading [HT07].

The noise rise measures the ratio of the base station total received interference to the base station thermal noise power. Assuming a new uplink connection is made to the network, it will increase the total interference received at the base station and lead to a new value of noise rise. If the noise rise value is less than or equal to the value of the interference margin, the uplink connection will be admitted. However, if the noise rise value is greater than the interference margin, the uplink connection will be rejected.

The method for calculating the received interference is presented as follows. E_b/N_o is an important factor used in the calculation. It is the ratio of the energy per bit to the noise spectral density. The E_b/N_o value is service type dependent, for instance, 5.0 dB for speech, 1.5 dB for 144 kbps real-time data, and 1.0 dB for 384 kbps non-real-time data [HT07]. The E_b/N_o value of the **admitted** user *j* can be derived as [HT07]:

$$(E_b / N_0)_j = processing\ gain\ of\ user\ j \times \frac{Received\ Signal\ Power\ of\ user\ j}{Total\ received\ power - Received\ Signal\ Power\ of\ user\ j}$$

It can be expressed as [HT07]:

$$(E_b/N_0)_j = \frac{W}{v_j \times R_j} \times \frac{i_j}{I_{total} - i_j} \tag{5.1}$$

W is the chip rate of W-CDMA; v_j is the activity factor of user *j*; R_j is the bit rate of user *j*; i_j is the received signal power from user *j* at the base station; I_{total} is the wideband power received at the base station, it includes the total received wideband signal power from the users who are being served in the cell of interest (intra-cell interference) and in the surrounding cells (inter-cell interference), as well as the thermal noise power at the base station.

The value of v_j is service type dependent and illustrated in Table 5.1:

Service Type	Value of the Activity Factor
Voice Service	0.67
Real time data service	1
Non real time data	0.2

Table 5.1 Values of the Activity Factor in W-CDMA [Che03][HT07]

Assuming the number of the admitted users in a UTRAN cell is *n*, the new user will be the *n+1*th user in the network. In order to calculate the E_b/N_o value of the *n+1*th user, which has not been admitted, equation 5.1 can be modified to:

$$(E_b/N_o)_{n+1} = \frac{W}{v_{n+1} \times R_{n+1}} \times \frac{i_{n+1}}{I_{total\_before\_n+1th\_user\_admitted}} \qquad (5.2)$$

i_{n+1} is the estimated signal power from the new connection. $I_{total\_before\_n+1th\_user\_admitted}$ is the wideband power received at the base station, before the *n+1*th user is admitted. The reason for such modification is that, before the request is admitted, the current wideband power at the base station includes the wideband signal power from the users who have already been admitted, but does not include the signal power from the new connection request. From the perspective of the new connection request, the current wideband power can be regarded as noise against its signal power at the base station. $I_{total\_before\_n+1th\_user\_admitted}$ can be calculated as:

$$I_{total\_before\_n+1th\_user\_admitted} = I_{thermal} + \sum_{k=1}^{n} i_k + I_{inter-cell} \qquad (5.3)$$

$I_{thermal}$ is the thermal noise power at the base station; i_k is the signal power from the *k*th admitted user in the cell of interest; $I_{inter-cell}$ is the inter-cell interference which includes the signal power from the users in the surrounding cells of the cell of interest.

During the development of the evaluation model, tests have shown that the inter-cell interference causes significant influence to the capacity of the interference-limited cellular network. However, the authors in [HT07] only assume the ratio of other cell to own cell interference to be 55% to 65%. In order to develop a more realistic evaluation model for UTRAN, in the uplink evaluation, the inter-cell interference is regarded as the signal power generated by the user terminals in the first tier of cells surrounding the cell of interest; in the downlink evaluation, the interference is regarded as the signal power generated by the base stations in the first and second tiers of cells surrounding the cell of interest. The layout of the cells is shown in Figure 5-1.

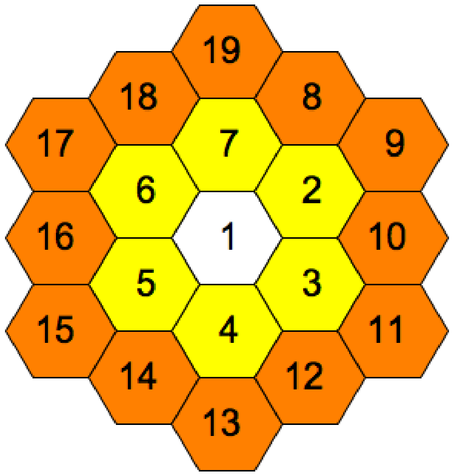


Figure 5-1 The Layout of the Cells

In Figure 5-1, the central cell denoted with a numerical value 1 is the cell of interest. The cells in yellow (cell 2 to cell 7) are the first tier of cells surrounding the cell of interest. The cells in orange (cell 8 to cell 19) are the second tier of cells surrounding the cell interest. In order to simplify the interference evaluation, the capacity in one cell is measured and all the surrounding cells are assumed to possess the same features as the cell of interest. They are identical in network configuration, link budget, user distribution, services to provide, power consumption, received interference and even the radio environment. Therefore, as a new connection request is made in the cell of interest, identical requests will correspondently occur in the surrounding cells and introduce new inter-cell interference. Similar assumptions were made in [Che03].

The inter-cell interference $I_{inter-cell}$ can be derived as:

$$I_{inter-cell} = \sum_{j=2}^{7}\sum_{k=1}^{n} p_k \times L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell} \tag{5.4}$$

p_k is the transmission power of the *k*th user; $L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell}$ is the radio propagation attenuation between the *k*th user in cell *j* and the base station in the cell of interest (the central cell).

p_k can be calculated as:

$$p_k = \frac{i_k}{L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \tag{5.5}$$

$L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}$ is radio propagation attenuation between the *k*th user in the central cell and the base station in the central cell.

Radio attenuation *L* is defined by [Vit98] as:

$$L = 10^{\zeta/10} \times path\_loss \tag{5.6}$$

ζ is the decibel attenuation due to shadowing. It follows a Gaussian distribution with zero mean and 6 dB standard deviation for micro cell and 8 or 10 dB for macro cell [Vit98]. The *path_loss* is the path loss between the base station and the user.

The path loss is defined by [ETSI98] as:

$$path\_loss = 40 \times \log_{10}(R) + 30 \times \log_{10}(f) + 49(dB) \tag{5.7}$$

or

$$path\_loss = 40 \times (1 - 4 \times 10^{-3} \times \Delta h_b)\log_{10}(R) - 18 \times \log_{10}(\Delta h_b) + 21 \times \log_{10}(f) + 80(dB) \tag{5.8}$$

In a micro cell environment, equation 5.7 is used, while equation 5.8 is used for macro cells. *R* (in kilometres) is the distance between the base station and the user; *f* is the carrier frequency; Δh_b is the base station antenna height.

Based on equation 5.2, i_{n+1} can be isolated as:

$$i_{n+1} = \frac{(E_b/N_o)_{n+1} \times v_{n+1} \times R_{n+1} \times I_{total\_before\_n+1th\_user\_admitted}}{W} \tag{5.9}$$

Therefore, the new total received wideband power after admitting the *n+1*th user can be estimated as:

$$I_{total\_after\_n+1th\_user\_admitted} = I_{total\_before\_n+1th\_user\_admitted} + i_{n+1} + I_{increased\_inter-cell\_interference} \tag{5.10}$$

$I_{increased\_inter-cell\_interference}$ is the increased inter-cell interference after the correspondent *n+1*th user is admitted at the surrounding cells. It can be calculated as:

$$I_{increased\_inter-cell\_interference} = \left(\frac{i_{n+1}}{L_{n+1th\_user\_in\_central\_cell,BS\_in\_central\_cell}} \right) \times \sum_{j=2}^{7} L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell} \tag{5.11}$$

From the perspective of the existing users, the new uplink connection increases the interference against their signal power received at the base station. In order to maintain their E_b/N_o values, they have to raise their signal power. For example, before the *n+1*th user is admitted, the signal power of the *n*th user is denoted as i_n. After the *n+1*th user is admitted, the signal power of the *n*th user has to be recalculated and is denoted as $i_n^{'}$. From equation 5.1, the E_b/N_o of the *n*th user can be calculated as:

$$\left(E_b/N_o\right)_n = \frac{W}{v_n \times R_n} \times \frac{i_n^{'}}{I_{total\_after\_n+1th\_user\_admitted} - i_n} \tag{5.12}$$

From equations 5.3, 5.4, 5.5, 5.10, and 5.11, the denominator $I_{total\_after\_n+1th\_user\_admitted} - i_n$ can be manipulated as:

$$I_{total\_after\_n+1th\_user\_admitted} - i_n = I_{thermal} + \sum_{k=1}^{n-1} i_k + i_{n+1} + \sum_{j=2}^{7}\sum_{k=1}^{n-1} \frac{i_k \times L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \sum_{j=2}^{7} \frac{i_{n+1} \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{n+1th\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \sum_{j=2}^{7} \frac{i_n^{'} \times L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \tag{5.13}$$

$\sum_{j=2}^{7} \frac{i_n^{'} \times L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}}$ is the inter-cell interference from the *n*th user correspondently residing in the surrounding cells, based on the estimated value $i_n^{'}$.

From equation 5.13, equation 5.12 can be changed as:

$$(E_b/N_o)_n = \frac{W}{v_n \times R_n} \times \frac{i_n^{'}}{\left(\begin{array}{l} I_{thermal} + \sum_{k=1}^{n-1} i_k + i_{n+1} + \sum_{j=2}^{7}\sum_{k=1}^{n-1} \frac{i_k \times L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \\ \sum_{j=2}^{7} \frac{i_{n+1} \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{n+1th\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \sum_{j=2}^{7} \frac{i_n^{'} \times L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \end{array} \right)}$$

Then, the above equation can be manipulated as:

$$i_n^{'} = \frac{(E_b/N_o)_n \times v_n \times R_n}{W} \times \left(\begin{array}{l} I_{thermal} + \sum_{k=1}^{n-1} i_k + i_{n+1} + \sum_{j=2}^{7}\sum_{k=1}^{n-1} \frac{i_k \times L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \\ \sum_{j=2}^{7} \frac{i_{n+1} \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{n+1th\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \sum_{j=2}^{7} \frac{i_n^{'} \times L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \end{array} \right)$$

$i_n^{'}$ can be isolated as:

$$i_n^{'} = \frac{\frac{(E_b/N_o)_n \times v_n \times R_n}{W} \times \left(\begin{array}{l} I_{thermal} + \sum_{k=1}^{n-1} i_k + i_{n+1} + \sum_{j=2}^{7}\sum_{k=1}^{n-1} \frac{i_k \times L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \\ \sum_{j=2}^{7} \frac{i_{n+1} \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{n+1th\_user\_in\_central\_cell,BS\_in\_central\_cell}} \end{array} \right)}{1 - \frac{(E_b/N_o)_n \times v_n \times R_n}{W} \times \sum_{j=2}^{7} \frac{L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}}}$$

The numerator of the above equation can be simplified as:

$$\begin{array}{l} I_{thermal} + \sum_{k=1}^{n-1} i_k + i_{n+1} + \sum_{j=2}^{7}\sum_{k=1}^{n-1} \frac{i_k \times L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \sum_{j=2}^{7} \frac{i_{n+1} \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{n+1th\_user\_in\_central\_cell,BS\_in\_central\_cell}} \\ = I_{thermal} + \sum_{k=1}^{n-1} i_k + i_n + i_{n+1} + \sum_{j=2}^{7}\sum_{k=1}^{n-1} \frac{i_k \times L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \sum_{j=2}^{7} \frac{i_n \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \\ \sum_{j=2}^{7} \frac{i_{n+1} \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{n+1th\_user\_in\_central\_cell,BS\_in\_central\_cell}} - i_n - \sum_{j=2}^{7} \frac{i_n \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \\ = I_{thermal} + \sum_{k=1}^{n} i_k + \sum_{j=2}^{7}\sum_{k=1}^{n} \frac{i_k \times L_{kth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{kth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + i_{n+1} + \sum_{j=2}^{7} \frac{i_{n+1} \times L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} - \\ i_n - \sum_{j=2}^{7} \frac{i_n \times L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \\ = I_{total\_before\_n+1th\_user\_admitted} + i_{n+1} + \frac{i_{n+1}}{L_{n+1th\_user\_in\_central\_cell,BS\_in\_central\_cell}} \times \sum_{j=2}^{7} L_{n+1th\_user\_in\_cell\_j,BS\_in\_central\_cell} - \\ i_n - \sum_{j=2}^{7} \frac{i_n \times L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \\ = I_{total\_after\_n+1th\_user\_admitted} - i_n - \sum_{j=2}^{7} \frac{i_n \times L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \end{array}$$

Finally, $i_n^{'}$ can be calculated as:

$$i_n^{'} = \frac{\frac{(E_b/N_o)_n \times v_n \times R_n}{W} \times \left(I_{total\_after\_n+1th\_user\_admitted} - i_n - \sum_{j=2}^{7} \frac{i_n \times L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \right)}{1 - \frac{(E_b/N_o)_n \times v_n \times R_n}{W} \times \sum_{j=2}^{7} \frac{L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}}} \tag{5.14}$$

The new value of the signal power generated by the *n*th user also increases the total received interference at the base station, which can be calculated as:

$$\begin{aligned} I_{total\_after\_received\_power\_of\_nth\_user\_recalculated} = I_{total\_before\_n+1th\_user\_admitted} - i_n + i_n^{'} - \\ \sum_{j=2}^{7} i_n \times \frac{L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} + \\ \sum_{j=2}^{7} i_n^{'} \times \frac{L_{nth\_user\_in\_cell\_j,BS\_in\_central\_cell}}{L_{nth\_user\_in\_central\_cell,BS\_in\_central\_cell}} \end{aligned} \tag{5.15}$$

In order to maintain the required E_b/N_o values, all the rest of users have to recalculate and raise their signal power, which further increase the total received interference. Then, considering the new value of the interference, the connection request has to recalculate its signal power and the total interference again based on equations 5.14 and 5.10. These recalculations will be repetitively carried out until the value of the total received interference stabilises and achieves an equilibrium state, or the noise rise value is greater than the interference margin. An iterative algorithm (Figure 5-2) was developed for calculating a stable value of the total received interference.

The algorithm is described as follows:

1. A new uplink connection request arrives.
2. Use equation 5.9 and 5.10 to estimate the initial new total interference at the base station.
3. If the noise rise introduced by the estimated initial new total interference exceeds the interference margin, go to step 8. Otherwise, go to step 4.
4. Use equations 5.14 and 5.15 to calculate the increased signal power from each existing user and the new total interference received at the base station.
5. If the noise rise generated by the new total interference is greater than the

interference margin, go to step 8. Otherwise, go to step 6.

6. Compare the value of the new total interference with the value of the previous total interference. If these two values are the same, the calculation has achieved equilibrium status and go to step 9. Otherwise, go to step 7.
7. Recalculate the signal power from each existing user and the total interference received at the base station, and go to step 3.
8. Reject the request and go to step 10.
9. Admit the request.
10. Stop the iterative algorithm

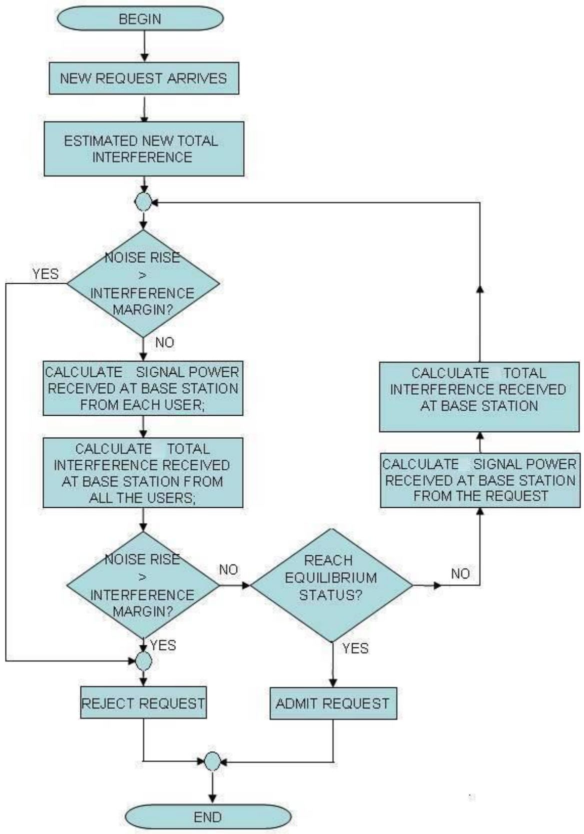


Figure 5-2 Flowchart of the UTRAN Uplink Iterative Algorithm

5.1.2 Downlink

In the downlink, the network resource availability is determined by the amount of base station transmission power being consumed and the maximum power that the base station can use. Assuming that the maximum value of the base station transmission power is 20 W [HT07], twenty percent of this power is used for signalling and the remaining, 16 W, is used for traffic [HT07]. Considering a new downlink connection request arrives, if the base station transmission power required for serving the existing connections and the new request is less than or equal to the maximum power for traffic (16 W), the request can be admitted. If not, it will be rejected. Similar to the uplink, the E_b/N_0 of the user *m* in the cell of interest (the central) can be calculated as:

$$(E_b/N_0)_m = Processing\ Gain\ of\ User\ m \times \frac{Received\ Signal\ Power\ from\ the\ Base\ Station\ to\ User\ m}{Total\ Intracell\ Interferece + Total\ Intercell\ Interferece + Thermal\ Noise\ at\ User\ m}$$

It can be expressed as [HT07]:

$$(E_b/N_0)_m = \frac{W}{v_m \times R_m} \times \frac{P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} \times L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}}{P_{BS\_in\_central\_cell} \times (1-\alpha) \times L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} + \sum_{k=2}^{19} P_{BS\_in\_cell\_k} \times L_{BS\_in\_cell\_k,mth\_user\_in\_central\_cell} + P_N} \tag{5.16}$$

$P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}$ is the transmission power of the base station in the central cell for the *m*th user; $L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}$ is the radio propagation attenuation between the base station in the central cell and the *m*th user; $P_{BS\_in\_central\_cell}$ is the total transmission power of the base station in the central cell; α is average orthogonality factor in the cell whose value ranges from 0.4 to 0.9 (1 means totally orthogonal); $P_{BS\_in\_cell\_k}$ is the total transmission power of the base station in cell *k*, which is in one of the cells (first and second tiers) surrounding the central cell; $L_{BS\_in\_cell\_k,mth\_user\_in\_central\_cell}$ is the radio propagation attenuation between the base station in cell *k* and the *m*th user in the central cell; P_N is the average thermal noise at the user terminal for the *m*th user.

As discussed before, all the surrounding cells possess the same features as the central cell. So, the values of the transmission power of every base station should be the same. $P_{BS\_in\_cell\_k}$ and $P_{BS\_in\_central\_cell}$ can be replaced by P_{BS} and equation 5.16 is modified to:

$$(E_b/N_0)_m = \frac{W}{v_m \times R_m} \times \frac{P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} \times L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}}{P_{BS} \times (1-\alpha) \times L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} + \sum_{k=2}^{19} P_{BS} \times L_{BS\_in\_cell\_k,mth\_user\_in\_central\_cell} + P_N}$$

Isolating $P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}$:

$$P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} = \frac{(E_b/N_0)_m \times v_m \times R_m}{W} \times \left\{ P_{BS} \times (1-\alpha) + \frac{\left[\sum_{k=2}^{19}\left(P_{BS} \times L_{BS\_in\_cell\_k,mth\_user\_in\_central\_cell}\right)\right] + P_N}{L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}} \right\}$$

(5.17)

Finally, by summing up the transmission power of the base station in the central cell for every individual user located in the central cell, such as $P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}$, the total transmission power of the base station in the central cell can be derived from equation 5.17 as:

$$P_{BS} = \sum_{m=1}^{n+1}\left\{ \frac{(E_b/N_0)_m \times v_m \times R_m}{W} \times \left\{ P_{BS} \times (1-\alpha) + \frac{\left[\sum_{k=2}^{19}\left(P_{BS} \times L_{BS\_in\_cell\_k,mth\_user\_in\_central\_cell}\right)\right] + P_N}{L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}} \right\}\right\} \quad (5.18)$$

Isolating P_{BS}:

$$P_{BS} = \frac{\sum_{m=1}^{n+1}\left[\frac{(E_b/N_0)_m \times v_m \times R_m}{W} \times \frac{P_N}{L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}} \right]}{1 - \sum_{m=1}^{n+1}\left[\frac{(E_b/N_0)_m \times v_m \times R_m}{W} \times \left[(1-\alpha) + \sum_{k=2}^{19} \frac{L_{BS\_in\_cell\_k,mth\_user\_in\_central\_cell}}{L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}} \right]\right]} \quad (5.19)$$

If the value of P_{BS} is less than or equal to the maximum power for traffic that the base station can use, the downlink connection request can be admitted. Otherwise, the downlink connection request will be rejected.

5.2 IEEE802.11a/b based WLAN Resource Availability Evaluation Model

IEEE802.11a/b based WLAN supports two access control schemes: *Distributed Coordination Function* (DCF) and *Point Coordination Function* (PCF). The DCF scheme, based on the *Carrier Sense Multiple Access with Collision Avoidance* (CSMA/CA) mechanism, is widely used and its basic operation is described in Appendix B. This resource availability evaluation model focuses on the DCF scheme.

The performance of IEEE802.11a/b based WLAN has been investigated in [CG00] and [Bia00]. The authors of these papers attempted to model the IEEE 802.11b backoff mechanism and derive the achievable throughput and maximum channel utilisation in different network conditions and configurations. The models presented in these two papers have been used as basic analytical methods for investigating IEEE802.11a/b based WLAN performance and have been cited in many publications on this field. The authors of [CG00] and [Bia00] assumed that the traffic in the IEEE802.11a/b based WLAN is inelastic and does not adjust its transmitting rate according to the available channel bandwidth. The authors also assumed that the senders operate in an asymptotic (or saturated) condition, which means that their traffic sources have unlimited amount of data and in each sender's queue always there is a packet ready to send. Furthermore, the traffic is assumed as uniformly distributed. That means each packet sent by one station is directed to another randomly selected station [BCB05].

However, these assumptions are not very consistent with the reality. Firstly, IEEE802.11a/b based WLAN primarily provides users with non-real-time data services, e.g. web browsing, FTP, etc. 95% of traffic in WLAN is composed of packets carried over the *Transport Control Protocol* (TCP) [BCB05]. The TCP traffic is elastic, because the TCP employs the flow control and the congestion control mechanisms to regulate its transmitting rate and avoid continuous packet transmission [BCB05]. Moreover, even for inelastic traffic, the senders may not operate in an asymptotic condition. For example, in *Voice over IP* (VoIP) services [WLL05], the voice data will be segmented into packets and these packets are not transmitted continuously, but they are transmitted at fixed intervals. The size of the voice data packet and the length of the transmission interval depend on the voice encoder being used. Secondly, most WLANs operate in an infrastructure-based mode, where the stations access network services through an AP. Therefore, the traffic is not uniformly distributed but all the packets are destined to/transmitted from the AP.

Bruno et al. pointed out the above inconsistencies in [BCB05] and investigated the performance of the TCP flows over the IEEE802.11 based WLAN in [BCB05][BCB06][BCB08][BCB09]. They assumed, the size of the congestion window of each TCP flow to be the same and, after an initial phase, the congestion window of each TCP flow grows to its maximum value. Therefore, each TCP flow reaches a stationary status and operates based on the TCP flow control algorithm. This assumption helps to simplify the complexity analysis and ensures that each TCP flow can have a fair access to the channel bandwidth. Bruno et al. also assumed that each station in the WLAN possess a single '*long-live*' [BCB05] TCP session which has an unlimited amount of data in the source and at least one packet to transmit. However, this assumption is inconsistent with the reality when considering the characteristics of different services. For example, a typical web browsing service session is not a '*long-live*' TCP session and it can be divided into ON/OFF periods [Der03]. The ON period represents web page download and consists of packet transmissions. The OFF period represents the intermediate reading time. Furthermore, real-time services, such as VoIP and Video Streaming, are becoming more popular and demanding evaluations for IEEE 802.11 network performance considering the coexistence of different traffic sources. Based on [BCB05], Bruno et al. present a throughput analysis of UDP and TCP flows in WLAN [BCB08]. However, they only consider the competition of multiple TCP downlink and uplink connections, with UDP uplink flows. Moreover, the UDP flows are still assumed to be saturated.

The assumptions of previous work are too stringent abstractions of real scenarios. Therefore, in this thesis, a simple but effective model was proposed and developed for evaluating resource availability in the IEEE802.11a/b based WLANs. The evaluation model tries to narrow the gap to more realistic network scenarios and analyses various service types, including VoIP, Video Call, Audio Streaming, Video Streaming, Web Browsing, and File Transfer. The first four services are real-time and UDP based. The others are non-real-time and TCP based.

The proposed evaluation model calculates the *expected number of contending packets* over the wireless channel, which is denoted by e_{ncp}. Assuming a new service request is made to the network, a new value of e_{ncp} is calculated based on the characteristics of the service request and the existing services. If the value of e_{ncp} is greater than 1, that means, on average, there is more than one packet in contention to access the network channel at any time. Before performing any action, the characteristics of the service request and the existing services within the network have to be considered. If the service request and the existing services are all UDP based or hybrid (coexistence of UDP and TCP based services), the request will be rejected. This is because channel contention will cause unacceptable delays and packet loss for the real-time UDP based services [WLL05]. If the service request and existing services are all TCP based, the analysis method proposed in [BCB05][BCB09] will be implemented to calculate the effective transmission rate (excluding the traffic and protocol overheads) of each packet generated by the service request and the existing services. For example, assuming the value of e_{ncp} is 2, on average there are two packets in contention to access the network channel at any time. In reality, not all services sessions behave as a *"long-live'* TCP session, such as web browsing. The packets transmitted over the network channel may be generated by more than two traffic sources. Therefore, these packets are *supposing* generated by ***two virtual*** TCP sessions coexisting in the network. These ***virtual*** TCP sessions always have packets in their queues and they are ready to send the packets at any time. Comparing with the *'long-live'* TCP session assumed in [BCB05][BCB09], these two ***virtual*** TCP sessions behave in the same way. Therefore, the analytical method presented in [BCB09] can be used to calculate the effective data transmission rate for each packet of the requested service and the existing services. These two ***virtual*** *'long-live'* TCP sessions can have a fair access to the IEEE802.11b based WLAN channel whose effective packet transmission rate is about 4500 kbps. This value is derived from the analytical method in [BCB09], which assumes the size of the congestion window of each TCP flow is the same and the size of a data packet is 1500 bytes (including the IP and TCP headers) and uses the typical values of the IEEE 802.11b DCF parameters (the values are presented in Appendix B). It is also validated through realistic discrete-event simulations in [BCB09]. Consequently, each ***virtual*** *'long-live'* TCP session can transmit each packet at an effective transmission rate of about 2250 kbps. That means the payload of each packet generated by these two ***virtual*** *'long-live'* TCP sessions can be transmitted at the data rate of 2250 kbps in the IEEE802.11b based WLAN. The calculation of e_{ncp} takes into account the requested service and the existing services in the network.

As mentioned before, not all service sessions behave as a *'long-live'* TCP session. Therefore, in a real scenario, the ***virtual*** *'long-live'* TCP session considered above may comprise several real TCP based service sessions. As a consequence, on average, each packet generated by each service session can be transmitted at an effective transmission rate of 2250 kbps. If the minimum QoS requirements (e.g. data rate) of the TCP based services still can be complied with, the service request will be admitted. Otherwise, the service request will be rejected.

In a situation where the value of e_{ncp} is equal to or less than 1, on average there is less than one packet in contention to access the network channel at any time. For UDP based real-time services in the network, such as VoIP, the data packets do not have to compete with other packets for channel access so that packet loss and severe delay can be avoided and the effective data rates are identical to the transmission rates of the encoders. For TCP based non-real-time services, without channel contention, each data packet can be transmitted at the maximum effective packet transmission rate (e.g. 4500 kbps in the IEEE802.11b based WLAN). If the minimum QoS requirements of the service request and the existing services can be satisfied, the request can be admitted. Otherwise, the service request will be rejected.

The value of the proposed e_{ncp} is defined as:

$$e_{ncp} = \sum_{i=0}^{N} i \times \binom{N}{i} \times {e_p}^{i} \times \left(1 - e_p\right)^{(N-i)} \tag{5.20}$$

N is the number of existing connections in the network plus the connection(s) introduced by the service request. For a VoIP service, two connections are considered. This is because the VoIP service is duplex and bi-directional. One connection is in the uplink and the other one is in the downlink. For an audio streaming, web browsing, or file transfer service, only one connection is considered. This is because the above services are asymmetrical and the downlink traffic plays a major role. Therefore, for simplicity reasons, the evaluation model neglects the uplink traffic of the latter services and only concentrates on the downlink. For a video call service, four connections are considered, because the video call service not only is duplex and bi-directional, but it also consists of two parts: voice and video. Similarly, a video streaming service introduces two connections.

e_p is the expected probability that the channel is occupied by a packet transmission. The value of e_p can be calculated as:

$$e_p = \sum_s (p\_on_s \times n_s / N) \tag{5.21}$$

s represents the service type and p_on_s is the probability that the channel is occupied by the transmission of the packets belonging to the service type *s*. n_s is the number of the connections of service type *s* and its value is service type dependent. The calculation of p_on_s depends on many factors, including data rate, packet size, packet interval, and service characteristics.

5.2.1 Calculation of *p_on*$_s$ for Real-time Services

In this subsection, the calculation of p_on_s for real-time services is presented. The real time services are UDP based and include VoIP, Video Call, Video Streaming, and Audio Streaming.

5.2.1.1 VoIP Service

For the VoIP service, p_on_{VoIP} can be derived as:

$$p\_on_{VoIP} = e\_t\_p_{VoIP} / in\_p_{VoIP} \tag{5.22}$$

in_p_{VoIP} represents the packet inter arrival time of the VoIP service. $e\_t\_p_{VoIP}$ is the expected time spent in completing the transmission for a packet. It consists of not only the time used for transmitting the packet payload and the headers, but also the overheads introduced by the CSMA/CA mechanism.

$e\_t\_p_{VoIP}$ can be calculated as follows:

$$\begin{aligned} e\_t\_p_{VoIP} = DIFS + e\_idle + phy\_mac\_hdr + \\ t\_b \times (payload + ip\_udp\_rtp\_hdr) + \\ delay + SIFS + delay + t\_ack \end{aligned} \tag{5.23}$$

DIFS represents the DCF interframe space and *SIFS* represents the short interframe space [WLAN07]. *e_idle* is the average backoff time [WLL05]. *phy_mac_hdr* is the time spent in transmitting the physical and MAC layer headers. *t_b* means the time used for transmitting a byte in the network. *payload* is the size of the service packet payload in bytes. *ip_udp_rtp_hdr* is the size of the IP/UDP/RTP header, which is 40 bytes. *delay* represents the maximum radio propagation between the terminal and the AP, which is about 1 μs [CG00]. Finally, *t_ack* is the time used to transmit an acknowledgement (ACK) packet.

Two service classes are considered: the *basic* class and the *premium* class. It is assumed that, for the *basic* class, the VoIP service employs G.723 as the codec. The G.723 codec is assumed to act as a constant bit rate (CBR) traffic source which generates 33 packets per second and whose data rate is 6.3 kbps. It means that the packet inter arrival time is 30 ms and the packet payload is 24 bytes. Therefore, in equation 5.22, the value of in_p_{VoIP} is 30 ms and, in equation 5.23, the value of *payload* is 24 bytes. As the IP/UDP/RTP header is 40 bytes, $e\_t\_p_{VoIP}$ can be calculated as follows:

$$\begin{aligned} e\_t\_p_{VoIP} = {} & DIFS + e\_idle + phy\_mac\_hdr + \\ & t\_b \times (24+40) + delay + SIFS + delay + t\_ack \end{aligned}$$

For the *premium* class, it is assumed that the VoIP service employs GSM610 as the codec. The GSM610 codec is assumed to generate 50 packets per second and whose data rate is 13.2 kbps. It means that the packet inter arrival time is 20 ms and the packet payload is 33 bytes. Therefore, in equation 5.22, the value of in_p_{VoIP} is 20 ms and, in equation 5.23, the value of *payload* is 33 bytes. As the IP/UDP/RTP header is 40 bytes, $e\_t\_p_{VoIP}$ can be calculated as follows:

$$\begin{aligned} e\_t\_p_{VoIP} = {} & DIFS + e\_idle + phy\_mac\_hdr + \\ & t\_b \times (33+40) + delay + SIFS + delay + t\_ack \end{aligned}$$

5.2.1.2 Video Call Service

The video call service is actually comprised of two parts: video and audio. Therefore, two probabilities should be considered: $p\_on_{video\_call\_video\_part}$ and $p\_on_{video\_call\_audio\_part}$. $p\_on_{video\_call\_video\_part}$ is the probability that the channel is occupied by the transmission of the video packets of the video call service.

$p\_on_{video\_call\_audio\_part}$ is the probability that the channel is occupied by the transmission of the audio packets of the video call service.

$p\_on_{video\_call\_video\_part}$ can be calculated as:

$$p\_on_{video\_call\_video\_part} = e\_t\_p_{video\_call\_video\_part} / in\_p_{video\_call\_video\_part} \tag{5.24}$$

It is assumed that the video component employs H.263 as the codec. The H.263 codec generates 10 frames per second and the frame inter arrival time is 100 ms. Therefore, the value of $in\_p_{video\_call\_video\_part}$ of equation 5.24 is 100 ms. Two service classes are considered: the *basic* class and the *premium* class. For the *basic* service class, the data rate is assumed to be 64 kbps, which means that the frame payload is 800 bytes. However, considering the instability of the radio channel, the size of a packet transmitted in a wireless network is suggested to be around 100 bytes [Wen03]. The evaluation model assumes that, for the H.263 codec, one video frame is divided into 8 slices. On average, the size of one slice is 100 bytes. $e\_t\_p_{video\_call\_video\_part}$ can be derived as:

$$e\_t\_p_{video\_call\_video\_part} = \begin{pmatrix} DIFS + e\_idle + phy\_mac\_hdr + \\ t\_b \times (100 + 40) + delay + SIFS + \\ delay + t\_ack \end{pmatrix} \times 8 \tag{5.25}$$

For the *premium* service class, the data rate is assumed to be 128 kbps, which means that the frame payload is 1600 bytes. It is assumed that one video frame is divided into 8 slices and on average the size of one slice is 200 bytes. Such division not only considers the instability of the radio channel and the suggested packet size, but also takes into account the packet transmission overhead introduced by the CSMA/CA mechanism. If the video frame is divided into too many slices, the overhead in transmitting these slices will intensify channel contention and deteriorate the performance of the WLAN. Considering that the size of one slice is 200 bytes, $e\_t\_p_{video\_call\_video\_part}$ can be derived as:

$$e\_t\_p_{video\_call\_video\_part} = \begin{pmatrix} DIFS + e\_idle + phy\_mac\_hdr + \\ t\_b \times (200 + 40) + delay + SIFS + \\ delay + t\_ack \end{pmatrix} \times 8 \tag{5.26}$$

For the audio part, $p\_on_{video\_call\_audio\_part}$ can be calculated as:

$$p\_on_{video\_call\_audio\_part} = e\_t\_p_{video\_call\_audio\_part} / in\_p_{video\_call\_audio\_part}$$

It is assumed that the audio part employs G.723 as the codec. The G.723 codec generates 33 packets per second and its data rate is 6.3 kbps. It means that, the packet inter arrival time is 30 ms and, on average, the packet payload is 24 bytes. Therefore, the value of $in\_p_{video\_call\_audio\_part}$ is 30 ms. $e\_t\_p_{video\_call\_audio\_part}$ can be derived as:

$$\begin{aligned} e\_t\_p_{video\_call\_audio\_part} = {} & DIFS + e\_idle + phy\_mac\_hdr + \\ & t\_b \times (24 + 40) + delay + SIFS + delay + t\_ack \end{aligned}$$

5.2.1.3 Audio Streaming Service

For the audio streaming service, $p\_on_{audio\_str}$ can be derived as:

$$p\_on_{audio\_str} = e\_t\_p_{audio\_str} / in\_p_{audio\_str} \tag{5.27}$$

It is assumed that the audio streaming service employs G.726 as the codec. The G.726 codec generates 50 packets per second, which means that the packet inter arrival time is 20 ms. Therefore, in equation 5.27, the value of $in\_p_{audio\_str}$ is 20 ms. Two service classes are considered: the *basic* class and the *premium* class. For the *basic* service class, the data rate is assumed to be 32 kbps, which means that the packet payload is 80 bytes. Therefore, $e\_t\_p_{audio\_str}$ can be calculated as:

$$\begin{aligned} e\_t\_p_{audio\_str} = {} & DIFS + e\_idle + phy\_mac\_hdr + \\ & t\_b \times (80 + 40) + delay + SIFS + delay + t\_ack \end{aligned}$$

For the *premium* service class, the data rate is assumed to be 64 kbps. Therefore, the size of the packet payload is 160 bytes and $e\_t\_p_{audio\_str}$ can be calculated as:

$$\begin{aligned} e\_t\_p_{audio\_str} = {} & DIFS + e\_idle + phy\_mac\_hdr + \\ & t\_b \times (160 + 40) + delay + SIFS + delay + t\_ack \end{aligned}$$

5.2.1.4 Video Streaming Service

Similar to the video call service, the video streaming service is also comprised of two parts: video and audio. Therefore, two probabilities should be considered: $p\_on_{video\_str\_video\_part}$ and $p\_on_{video\_str\_audio\_part}$. $p\_on_{video\_str\_video\_part}$ is the probability that the channel is occupied by the transmission of the video packets of the video streaming service. $p\_on_{video\_str\_audio\_part}$ is the probability that the channel is occupied by the transmission of the audio packets of the video streaming service.

$p\_on_{video\_str\_video\_part}$ can be calculated as:

$$p\_on_{video\_str\_video\_part} = e\_t\_p_{video\_str\_video\_part} / in\_p_{video\_str\_video\_part} \qquad (5.28)$$

It is assumed that the video part employs H.264 as the codec. The H.264 codec generates 30 frames per second, which means that the frame inter arrival time is 33 ms. Therefore, in equation 5.28, the value of $in\_p_{video\_str\_video\_part}$ is 33 ms. Two service classes are considered: the *basic* class and the *premium* class. For the *basic* service class, the video data rate is assumed to be 64 kbps, which means that, on average, the frame payload is 266 bytes. One frame is divided into 4 slices and the size of each slice is 66 bytes. Therefore, $e\_t\_p_{video\_str\_video\_part}$ can be calculated as:

$$e\_t\_p_{video\_str\_video\_part} = \begin{pmatrix} DIFS + e\_idle + phy\_mac\_hdr + t\_b \times (66 + 40) \\ +delay + SIFS + delay + t\_ack \end{pmatrix} \times 4$$

For the *premium* service class, the video data rate is assumed to be 128 kbps. On average, the frame payload is 533 bytes. One frame is divided into 8 slices and the size of each slice is 67 bytes. $e\_t\_p_{video\_str\_video\_part}$ can be calculated as:

$$e\_t\_p_{video\_str\_video\_part} = \begin{pmatrix} DIFS + e\_idle + phy\_mac\_hdr + t\_b \times (67 + 40) \\ +delay + SIFS + delay + t\_ack \end{pmatrix} \times 8$$

Similarly, for the audio component, $p\_on_{video\_str\_audio\_part}$ can be calculated as:

$$p\_on_{video\_str\_audio\_part} = e\_t\_p_{video\_str\_audio\_part} / in\_p_{video\_str\_audio\_part} \qquad (5.29)$$

It is assumed that the audio part employs G.726 as the codec. The G.726 codec generates 50 packets per second and its data rate is 32 kbps. It means that, the packet inter arrival time is 20 ms and the packet payload is 80 bytes. In equation 5.21, the value of $in\_p_{video\_str\_audio\_part}$ is 20 ms and $e\_t\_p_{video\_str\_audio\_part}$ can be calculated as:

$$e\_t\_p_{video\_str\_audio\_part} = DIFS + e\_idle + phy\_mac\_hdr + t\_b \times (80 + 40) + delay + SIFS + delay + t\_ack$$

5.2.1.5 Accepting/Rejecting Real-time Services

The proposed evaluation model calculates value of e_{ncp} based on the characteristics of the requested service and the existing services. If value of e_{ncp} is smaller than or equal to 1 and the requested service and the existing services are all UDP based or hybrid, the service request will be accepted. Otherwise, the service request will be rejected.

In order to clarify the decision-making process, a scenario considering a variety of real-time services is studied. The services include VoIP with premium service class, Video Call with premium service class, Audio Streaming with premium and basic service classes, and Video Streaming with premium and basic service classes. Based on the calculation presented in subsections 5.2.1.1, 5.2.1.2, 5.2.1.3, and 5.2.1.4, the probability that the channel is occupied by the transmission of the packets belonging to the each type of service is listed in Table 5.2.

Service Type	***p_on***
VoIP with premium service class	$p\_on_{VoIP\_premium} = 0.044$
Video Call with premium service class: video part	$p\_on_{video\_call\_video\_part\_premium} = 0.0801$
Video Call with premium service class: audio part	$p\_on_{video\_call\_audio\_part} = 0.0291$
Audio Streaming with premium service class	$p\_on_{audio\_str\_premium} = 0.048609$
Audio Streaming with basic service class	$p\_on_{audio\_str\_basic} = 0.0457$
Video Streaming with premium service class: video part	$p\_on_{video\_str\_video\_part\_premium} = 0.219284$
Video Streaming with basic service class: video part	$p\_on_{video\_str\_video\_part\_basic} = 0.1096$
Video Streaming: audio part	$p\_on_{video\_str\_audio\_part} = 0.0457$

Table 5.2 Channel Occupation Probability of Each Type of Service

Assuming that the number of each service type is 1, these services will introduce 12 connections to the WLAN, including: two connections from the premium VoIP service, four connections from the premium video call service, one connection from the premium audio streaming service, one connection from the basic audio streaming service, two connections from the premium video streaming service, and two connections from the basic video streaming service. Therefore, based on equation 5.21 and the values of p_on_{st} of each type of service, the expected probability that the channel is occupied by a packet transmission (e_p) can be calculated as:

$$
\begin{aligned}
e_p &= \sum_s \left(p\_on_s \times n_s / N\right) \\
&= p\_on_{VoIP} \times \frac{n_{VoIP}}{N} + \\
&\quad p\_on_{video\_call\_video\_part} \times \frac{n_{video\_call\_video\_part}}{N} + p\_on_{video\_call\_audio\_part} \times \frac{n_{video\_call\_audio\_part}}{N} + \\
&\quad p\_on_{audio\_str\_premium} \times \frac{n_{audio\_str\_premium}}{N} + p\_on_{audio\_str\_basic} \times \frac{n_{audio\_str\_basic}}{N} + \\
&\quad p\_on_{video\_str\_video\_part\_premium} \times \frac{n_{video\_str\_video\_part\_premium}}{N} + p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{N} \\
&\quad p\_on_{video\_str\_video\_part\_basic} \times \frac{n_{video\_str\_video\_part\_basic}}{N} + p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{N} + \\
&= 0.044 \times \frac{2}{12} + 0.0801 \times \frac{2}{12} + 0.0291 \times \frac{2}{12} + 0.048609 \times \frac{1}{12} + 0.0457 \times \frac{1}{12} + \\
&\quad 0.219284 \times \frac{1}{12} + 0.0457 \times \frac{1}{12} + 0.1096 \times \frac{1}{12} + 0.0457 \times \frac{1}{12} \\
&= 0.0684125
\end{aligned}
$$

The value of e_{ncp} can be calculated by using equation 5.20 as follows:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times \left(1 - e_p\right)^{(N-i)} \\
&= \sum_{i=0}^{12} i \times \binom{12}{i} \times 0.0684125^i \times (1 - 0.0684125)^{(12-i)} = 0.8209502
\end{aligned}
$$

The value of e_{ncp} is smaller than 1. That means, on average, there is less than one packet in contention to access the network channel. The WLAN is estimated to be able to support these services. Assuming that a premium video call service request arrives at the WLAN, the number of connections would increase to 16 and the value of e_p can be calculated as:

$$
\begin{aligned}
e_p &= \sum_s \left(p\_on_s \times n_s / N\right) \\
&= p\_on_{VoIP} \times \frac{n_{VoIP}}{N} + \\
&\quad p\_on_{video\_call\_video\_part} \times \frac{n_{video\_call\_video\_part}}{N} + p\_on_{video\_call\_audio\_part} \times \frac{n_{video\_call\_audio\_part}}{N} + \\
&\quad p\_on_{audio\_str\_premium} \times \frac{n_{audio\_str\_premium}}{N} + p\_on_{audio\_str\_basic} \times \frac{n_{audio\_str\_basic}}{N} + \\
&\quad p\_on_{video\_str\_video\_part\_premium} \times \frac{n_{video\_str\_video\_part\_premium}}{N} + p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{N} \\
&\quad p\_on_{video\_str\_video\_part\_basic} \times \frac{n_{video\_str\_video\_part\_basic}}{N} + p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{N} + \\
&= 0.044 \times \frac{2}{16} + 0.0801 \times \frac{4}{16} + 0.0291 \times \frac{4}{16} + 0.048609 \times \frac{1}{16} + 0.0457 \times \frac{1}{16} + \\
&\quad 0.219284 \times \frac{1}{16} + 0.0457 \times \frac{1}{16} + 0.1096 \times \frac{1}{16} + 0.0457 \times \frac{1}{16} \\
&= 0.06496075
\end{aligned}
$$

Therefore, the value of e_{ncp} can be obtained as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times \left(1 - e_p\right)^{(N-i)} \\
&= \sum_{i=0}^{16} i \times \binom{16}{i} \times 0.06496075^i \times \left(1 - 0.06496075\right)^{(16-i)} = 1.0394
\end{aligned}
$$

e_{ncp} exceeds 1 to be 1.0394 and the increased channel contention will cause unacceptable delays and packet loss to the real-time UDP based service. In order to maintain the service quality, the video call service request should be rejected.

5.2.2 Calculation of p_on_s for Non-real-time Services

The non-real-time services include File Transfer and Web Browsing and they are TCP based. The network resource availability evaluation model focuses on the performance of the radio access network of the IEEE802.11a/b based WLAN. The evaluation model assumes that the FTP and HTTP servers are collocated with the AP. After receiving a TCP acknowledgement packet from a user, the FTP or HTTP server can immediately get ready to transmit the next packet at the AP side. Therefore, the influence of the core network delay on the radio access network performance can be neglected. Furthermore, the evaluation model also assumes that the size of the congestion window of each TCP flow is the same and, after an initial phase, the congestion window of each TCP flow grows to its maximum value. This assumption ensures each active TCP flow can have a fair access to the channel bandwidth.

5.2.2.1 File Transfer Service

For an active File Transfer service session, the user receives data packets from the FTP server and replies with TCP acknowledgement packets. Before the file is completely transferred, there is a packet, which can be a data packet or a TCP acknowledgement packet, in contention to access the network channel at any time. Therefore, the value of $p\_on_{file\_transfer}$ is 1 and an active File Transfer service session can be considered as one *'long-live'* TCP session.

5.2.2.2 Web Browsing Service

The Web Browsing service is based on the HTTP protocol. A Web Browsing service session can be divided into ON/OFF periods, which are the result of human interaction. An ON period represents the download of a web page, which can be referred as a "*packet call*" shown in Figure 5-3. An OFF period represents the intermediate reading time for digesting the downloaded web page.

During the ON periods, the user receives data packets from the HTTP server and replies with TCP acknowledgement packets. Before the web page is completely downloaded, there is a packet, which can be a data packet or a TCP acknowledgement packet, in contention to access the network channel at any time. Therefore, during the ON period, the Web Browsing service session can be considered as one *'long-live'* TCP session.

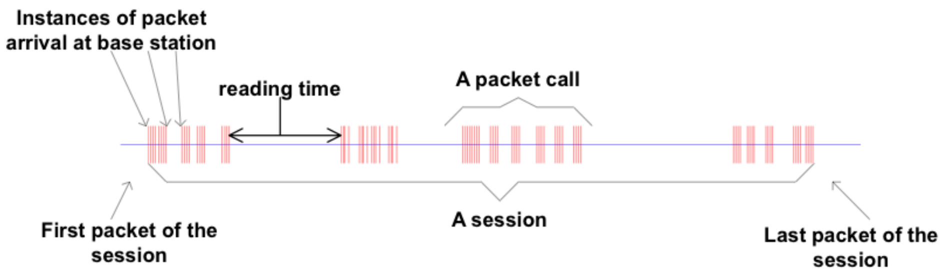


Figure 5-3 Packet Trace of a Typical Web Browsing Service Session [Der03]

During the OFF periods, the user reads the web page and no data transmissions are between the user and the HTTP server. $p\_on_{web\_browsig}$ can be calculated as:

$$p\_on_{web\_browsig} = \frac{e\_t\_p_{web\_page}}{e\_t\_p_{web\_page} + avg\_reading\_time} \qquad (5.30)$$

$e\_t\_p_{web\_page}$ is the expected time spent in completing the transmission for a web page. *avg_reading_time* is the average time spent in reading a downloaded web page, which is assumed to be 30 seconds [Der03].

Depending on the traffic conditions in the WLAN, the value of $e\_t\_p_{web\_page}$ can be firstly estimated in two ways. If the value of e_{ncp} is smaller than the number of active File Transfer service sessions plus one, $e\_t\_p_{web\_page}$ can be approximated as:

$$e\_t\_p_{web\_page} = \frac{avg\_size\_web\_page \times 8}{wlan\_data\_rate/(n\_ftp\_session+1)} \qquad (5.31a)$$

avg_size_web_page is the average size of a web page in bytes, which is assumed to be 312000 bytes [Kin08]. *wlan_data_rate* is the effective TCP packet transmission rate in the WLAN (e.g. 4500 kbps for the IEEE802.11b based WLAN). *n_ftp_session* is the number of active File Transfer service sessions in the WLAN.

Equation 5.31a is derived considering that the Web Browsing service sessions in the WLAN are not able to constitute a ***virtual*** '*long-live*' TCP session. As illustrated in subsection 5.2.2.1, an active File Transfer service session can be considered as one '*long-live*' TCP session. As the value of e_{ncp} is smaller than the number of File Transfer service sessions plus one, it means that, by neglecting the File Transfer service sessions, the Web Browsing service sessions cannot produce an e_{ncp} whose value is greater than or equal to 1. Such result indicates that, on average, there is less than one Web Browsing service packet in contention to access the network channel at any time. Therefore, from a Web Browsing service session's point of view, during the ON period, the major competition it will face is from the active File Transfer service sessions rather than the peer Web Browsing service sessions. Therefore, the effective packet transmission rate for the Web Browsing service session can be calculated as *wlan_data_rate/*(*n_ftp_session*+1).

If the value of e_{ncp} is greater than or equal to the number of active File Transfer service session plus one, $e\_t\_p_{web\_page}$ can be approximated as:

$$e\_t\_p_{web\_page} = \frac{avg\_size\_web\_page \times 8}{wlan\_data\_rate / e_{ncp}} \quad (5.31b)$$

Equation 5.31b is derived considering that the Web Browsing service sessions in the WLAN can constitute ***virtual*** *'long-live'* TCP session(s). As the value of e_{ncp} is greater than or equal to the number of File Transfer service sessions plus one, it means that, by neglecting the File Transfer service sessions, the Web Browsing service sessions can produce an e_{ncp} whose value is greater than or equal to 1. Such result indicates that, on average, there is at least one Web Browsing service packet in contention to access the network channel at any time. Therefore, from a Web Browsing service session's point of view, during the ON period, the major competition it will face is not only from the active File Transfer service sessions but also from the peer Web Browsing service sessions. Therefore, the effective packet transmission rate for the Web Browsing service session can be calculated as $wlan\_data\_rate/e_{ncp}$.

By observing equations 5.30, 5.31a, and 5.31b, one can note that the value of $p\_on_{web\_browsig}$ is dependent on the value of e_{ncp}. In return, the new value of $p\_on_{web\_browsig}$ also produces a new value of e_{ncp}, which again may affect the value of $p\_on_{web\_browsig}$. Therefore, an iterative process may be required to explore the value of e_{ncp} and the detailed calculation process is explained in subsection 5.2.2.3.

5.2.2.3 Accepting/Rejecting Non-real-time Services

In order to clarify the decision-making process for accepting or rejecting non-real-time services, two scenarios considering a variety of types of non-real-time services are studied. The services include File Transfer service and Web Browsing service. These services can belong to two different service classes, *basic* and *premium*, which differ from each other by the required data rate. The required data rates for each service class of the File Transfer and Web Browsing services are presented in Table 5.3. The values of the data rate are minimum requirements for each service class.

Service Type	Web Browsing	File Transfer
Required Data Rate for Basic Service Class	128 kbps	64 kbps
Required Data Rate for Premium Service Class	256 kbps	128 kbps

Table 5.3 Required Data Rates for the Basic/Premium Service Class of Web Browsing and File Transfer Services

Scenario 1

Assuming there are two active File Transfer service sessions in the IEEE802.11b based WLAN and no other services are running, the number of service connections is 2 and the value of e_{ncp} is 2. When a Web Browsing service request arrives at the WLAN, the number of connections increases to 3 and the effective TCP packet transmission rate for the requested Web Browsing service can be approximated as 1500 kbps based on equation 5.31a. Comparing with the required data rates presented in Table 5.3, the approximated effective packet transmission rate is significantly greater than the requirements. Therefore, the Web Browsing service request is considered to be acceptable and the further calculation of the value of e_{ncp} can be conducted.

Based on equation 5.31a, the expected time spent in completing the transmission for a web page ($e\_t\_p_{web\_page}$) can be calculated as:

$$
\begin{aligned}
e\_t\_p_{web\_page} &= \frac{avg\_size\_web\_page \times 8}{wlan\_data\_rate/(n\_ftp\_session+1)} \\
&= \frac{312000 \times 8}{4500000/3} \\
&= 1.664
\end{aligned}
$$

Assuming the average reading time is 30 seconds, $p\_on_{web\_browsnig}$ can be obtained based on equation 5.30:

$$p\_on_{web\_browsing} = \frac{e\_t\_p_{web\_page}}{e\_t\_p_{web\_page} + avg\_reading\_time}$$
$$= \frac{1.664}{1.664+30} = 0.05255$$

The expected probability that the channel is occupied by a packet transmission (e_p) can be calculated as:

$$\begin{aligned}
e_p &= \sum_s \left(p_on_s \times n_s / N\right) \\
&= p_on_{file_transfer} \times \frac{n_{file_transfer}}{N} + p_on_{web_browsing} \times \frac{n_{web_browsing}}{N} \\
&= 0.05255 \times \frac{1}{3} + 1 \times \frac{2}{3} \\
&= 0.68418
\end{aligned}$$

After that, the value of e_{ncp} can be calculated by using equation 5.20 as follows:

$$\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1-e_p)^{(N-i)} \\
&= \sum_{i=0}^{3} i \times \binom{3}{i} \times 0.68418^i \times (1-0.68418)^{(12-i)} = 2.052552
\end{aligned}$$

The value of e_{ncp} is smaller than the number of File Transfer service session plus one. It means that the admitted Web Browsing service session is unable to constitute a ***virtual*** '*long-live*' TCP session in the WLAN and no further calculation for the value of e_{ncp} is required. Then, the effective packet transmission rate for the File Transfer service sessions can be derived as:

$$wlan\_data\_rate / e_{ncp} = \frac{4500kbps}{2.052552} = 2192kbps$$

The effective packet transmission rate for the File Transfer service session is greater than the required data rate in Table 5.3. Therefore, the Web Browsing service request can finally be accepted.

Scenario 2

Assuming there are two active File Transfer service sessions and twenty-two Web Browsing service sessions WLAN, the number of service connections is 24 and the value of e_{ncp} is 3.245335. When a Web Browsing service request arrives the WLAN, the effective TCP packet transmission rate for the request Web Browsing service can be approximated as 1386 kbps based on equation 5.31b. Comparing with the required data rates presented in Table 5.3, the approximated effective packet transmission rate is greater than the requirements. Therefore, the Web Browsing service request is estimated to be acceptable and the further calculation of the value of e_{ncp} can be conducted. Based on equation 5.31b, the expected time spent in completing the transmission for a web page ($e\_t\_p_{web\_page}$) can be calculated as:

$$\begin{aligned} e\_t\_p_{web\_page} &= \frac{avg\_size\_web\_page \times 8}{wlan\_data\_rate / e_{ncp}} \\ &= \frac{312000 \times 8}{4500000/3.245335} \\ &= 1.8 \end{aligned}$$

Assuming the average reading time is 30 seconds, $p\_on_{web\_browsing}$ can be obtained based on equation 5.30:

$$\begin{aligned} p\_on_{web\_browsing} &= \frac{e\_t\_p_{web\_page}}{e\_t\_p_{web\_page} + avg\_reading\_time} \\ &= \frac{1.8}{1.8+30} = 0.0566 \end{aligned}$$

The expected probability that the channel is occupied by a packet transmission (e_p) can be calculated as:

$$\begin{aligned} e_p &= \sum_s \left(p\_on_s \times n_s / N\right) \\ &= p\_on_{file\_transfer} \times \frac{n_{file\_transfer}}{N} + p\_on_{web\_browsig} \times \frac{n_{web\_browsig}}{N} \\ &= 0.0566 \times \frac{23}{25} + 1 \times \frac{2}{25} 0.052072 \\ &= 0.132072 \end{aligned}$$

After that, the value of e_{ncp} can be calculated by using equation 5.20 as follows:

$$e_{ncp} = \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1-e_p)^{(N-i)}$$
$$= \sum_{i=0}^{25} i \times \binom{25}{i} \times 0.132072^i \times (1-0.132072)^{(25-i)} = 3.3018$$

The value of e_{ncp} is greater than the number of File Transfer service sessions plus one. It means that the admitted Web Browsing service session can constitute a ***virtual*** '*long-live*' TCP session in the WLAN. As discussed in subsection 5.2.2.2, this new value of e_{ncp} will affect the value of $p\_on_{web\_browsing}$, which, in return, will again affect the calculation of e_{ncp}. Therefore, in order to explore the value of e_{ncp}, an iterative process is implemented and presented in Figure 5-4.

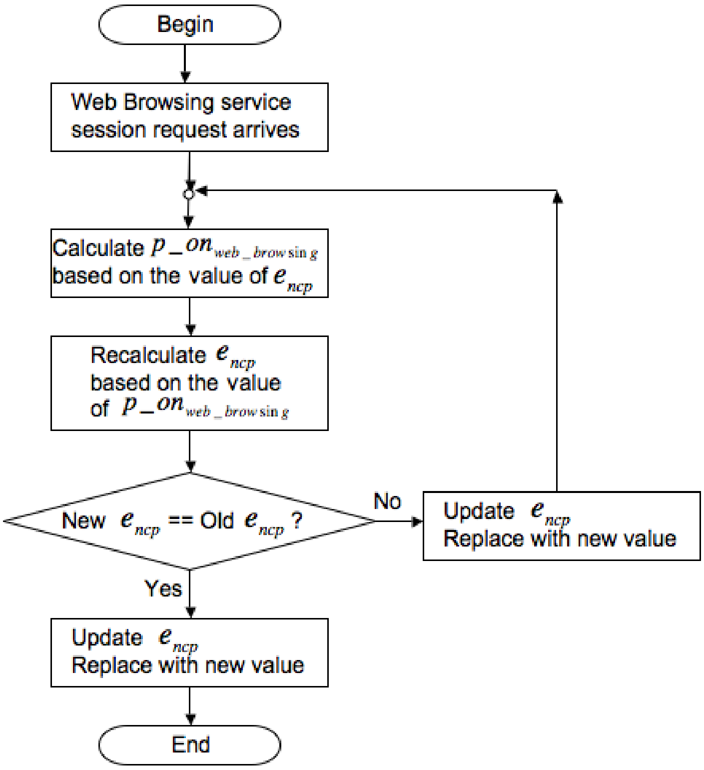


Figure 5-4 Flowchart of the Iterative Algorithm for Calculating e_{ncp}

The algorithm is described as follows:

1. A Web Browsing service session request arrives.
2. Based on the value of e_{ncp}, use equations 5.30 and 5.31b to calculate $p\_on_{web\_browsing}$

3. Based on the value of $p\_on_{web\_browsing}$, use equations 5.20 and 5.21 to recalculate e_{ncp}.
4. Compare the new value of e_{ncp} and the old value of e_{ncp}. If they are the same, it means that the equilibrium state is achieved and then go to step 6. Otherwise, go to step 5.
5. Update e_{ncp} with the new value and go to step 2.
6. Update e_{ncp} with the new value and go to step 7.
7. End the iterative algorithm.

After implementing the iterative algorithm, the final value of e_{ncp} is calculated to be 3.336321. Then, the effective packet transmission rate for both of the File Transfer service and the Web Browsing service can be derived as:

$$wlan\_data\_rate / e_{ncp} = \frac{4500kbps}{3.336321} = 1349kbps$$

The effective packet transmission rate for the File Transfer service and Web Browsing service is greater than the required data rate in Table 5.3. Therefore, the Web Browsing service request can finally be accepted.

5.3 Overview of Simulations

The proposed UTRAN and WLAN resource availability evaluation models enable the implementation of call level simulations for comparing and evaluating the performances of RAN selection and optimisation algorithms. The simulations take into account two types of RANs, a UTRAN cell and a WLAN, and no background traffic is assumed. In the simulations, all the users are uniformly distributed at random locations in the UTRAN cell and also covered by the WLAN with acceptable radio channel conditions. In the simulations, the resource availability evaluation models facilitate the understanding of the condition of network resource usage. When a service request arrives, considering the service type and the RAN being evaluated, the values of required noise rise, base station power, or the expected number of contention packets will be calculated by using the resource availability evaluation models. If the values of these parameters are within the constraints, such as the maximum transmission power that the base station can use for traffic, it means that the resources of the network being evaluated are sufficient to admit the service request. Otherwise, the network resources are considered to be insufficient to admit the service request. Such information will be provided to the RAN selection and optimisation algorithms and enable the algorithms to understand the conditions of network resource usage and availability. Then, the RAN selection and optimisation algorithms may perform further operations to adjust service requirements and optimise resource usage for network selection and admission control. The feasibility of these operations (whether the network resources are sufficient) is also verified by the resource availability evaluation models. Due to user arrival and departure, network resource availability may vary frequently. In the simulations, such variability is considered and the network resource availability is updated and examined by using the proposed evaluation models. For example, the admission of a Video Call in the UTRAN cell will increase the noise rise and the base station power. In the simulations, the network resource availability evaluation model will be used to recalculate and update the value of noise rise and base station power to reflect such alteration. The proposed evaluation models provide a means of understanding the conditions of the network resource availability and facilitate adjustment and optimisation in network resource usage.

5.4 Concluding Remarks

This chapter presents the models for evaluating network resource availability. Although the evaluation models are abstractions of real networks, they are based on appropriate assumptions and effectively obtain important network context information which helps the development and evaluation of effective RAN selection and optimisation algorithms. The next chapter validates the network resource availability evaluation models.

Chapter 6 Validation of Network Resource Availability Evaluation Models

Chapter 5 presents two models for evaluating network resource availability in UTRAN and IEEE802.11a/b based WLAN. These models are used to effectively obtain important network context information and enable the implementation of call level simulations for developing and evaluating RAN selection and optimisation algorithms. In order to validate the network resource availability evaluation models, a series of simulations were carried out through different scenarios. By comparing the evaluation results and the simulation results, the evaluation models are assessed and validated.

6.1 Validation of UTRAN Network Resource Availability Evaluation Model

In order to validate the proposed UTRAN network resource availability evaluation model, simulations are carried out to compare the evaluation results from the proposed network resource availability evaluation model and the model in [HT07]. Before demonstrating the simulation results, the network resource availability evaluation model developed based on [HT07] should be presented.

In [HT07], the inter-cell interference is assumed to be proportional to the intra-cell interference. For example, in a macro cell with an omnidirectional antenna, the ratio of the inter-cell interference to the intra-cell interference is assumed as 55%. In the uplink, assuming the number of admitted users in a UTRAN cell is n, equation 5.4 can be modified to estimate the inter-cell interference $I_{inter-cell}$ as:

$$I_{inter-cell} = \theta \times \sum_{k=1}^{n} i_k \tag{6.1}$$

where θ is the ratio of the inter-cell interference to the intra-cell interference in the uplink. Similarly, after admitting the $n+1$th user, equation 5.11 is modified to calculate the increased inter-cell interference $I_{increased\_inter-cell\_interference}$ as:

$$I_{increased\_inter-cell\_interference} = \theta \times i_{n+1} \tag{6.2}$$

From the perspective of an existing user, such as the *n*th user, the new uplink connection increases the interference against his/her signal power received at the base station. In order to maintain the target E_b/N_0, the *n*th user has to raise his/her signal power received at the base station from i_n to $i_n^{'}$. For serving the *n*th user, the total interference and thermal noise received at the base station $I_{total\_after\_n+1th\_user\_admitted} - i_n$ can be estimated by modifying equation 5.13 as:

$$I_{total\_after\_n+1th\_user\_admitted} - i_n = I_{thermal} + \sum_{k=1}^{n-1} i_k + i_{n+1} + \theta \times \left(\sum_{k=1}^{n-1} i_k + i_{n+1} \right) \quad (6.3)$$

From equation 6.3, equation 5.12 can be expanded into:

$$(E_b/N_o)_n = \frac{W}{v_n \times R_n} \times \frac{i_n^{'}}{I_{thermal} + \sum_{k=1}^{n-1} i_k + i_{n+1} + \theta \times \left(\sum_{k=1}^{n-1} i_k + i_{n+1} \right)}$$

The above equation can be manipulated as:

$$i_n^{'} = \frac{(E_b/N_o)_n \times v_n \times R_n}{W} \times \left[I_{thermal} + \sum_{k=1}^{n-1} i_k + i_{n+1} + \theta \times \left(\sum_{k=1}^{n-1} i_k + i_{n+1} \right) \right]$$

The new value of the signal power generated by the *n*th user also increases the total interference received at the base station, which can be calculated by modifying equation 5.15 as:

$$\begin{aligned} I_{total\_after\_received\_power\_of\_nth\_user\_recalculated} = {} & I_{total\_after\_n+1th\_user\_admitted} \\ & - (1+\theta) \times i_n + (1+\theta) \times i_n^{'} \end{aligned} \quad (6.4)$$

Finally, the iterative algorithm presented the subsection 5.1.1 is performed to obtain a stable value of the total received interference.

In the downlink, denoting σ as the ratio of the inter-cell interference to the intra-cell interference received at the user terminal, equation 5.16 can be modified as:

$$\begin{aligned} (E_b/N_0)_m = {} & \frac{W}{v_m \times R_m} \times \\ & \frac{P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} \times L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}}{\left[\begin{aligned} & P_{BS} \times (1-\alpha) \times L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} + \\ & P_{BS} \times (1-\alpha) \times L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} \times \sigma + P_N \end{aligned} \right]} \end{aligned} \quad (6.5)$$

Therefore, the transmission power of the base station in the central cell for the *m*th user $P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}$ can be obtained as:

$$P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell} = \frac{(E_b/N_0)_m \times v_m \times R_m}{W} \times \left\{ P_{BS} \times (1-\alpha) \times (1+\sigma) + \frac{P_N}{L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}} \right\} \tag{6.6}$$

By summing up the transmission power of the base station in the central cell for every individual user located in the central cell, such as $P_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}$, the total transmission power of the base station in the central cell can be derived from equation 6.6 as:

$$P_{BS} = \sum_{m=1}^{n+1} \left\{ \frac{(E_b/N_0)_m \times v_m \times R_m}{W} \times \left\{ P_{BS} \times (1-\alpha) \times (1+\sigma) + \frac{P_N}{L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}} \right\} \right\} \tag{6.7}$$

Isolating P_{BS}:

$$P_{BS} = \frac{\sum_{m=1}^{n+1} \left[\frac{(E_b/N_0)_m \times v_m \times R_m}{W} \times \frac{P_N}{L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}} \right]}{1 - \sum_{m=1}^{n+1} \left[\frac{(E_b/N_0)_m \times v_m \times R_m}{W} \times (1-\alpha) \times (1+\sigma) \right]} \tag{6.8}$$

In the simulation, both evaluation models are implemented to evaluate the resource availability of a UTRAN cell. The radius of the UTRAN cell is 1000 meters and an omnidirectional antenna is used. The maximum base station transmission power is 20 W. Twenty percent of this power is used for signalling and the remaining, 16 W, is used for traffic. The average orthogonal factor in the cell (α) is assumed as 0.6. The interference margin is defined as 3 dB. The average thermal noise at the base station ($I_{thermal}$) and the user terminal (P_N) are -103.2 dBm and -169.0 dBm, respectively. Both of the ratios of the inter-cell interference to the intra-cell interference in the uplink (θ) and in the downlink (σ) are assumed to be 55% [HT07].

The simulation considers a speech service. The service data rate is 12.2 kbps and the activity factor (v) is defined as 0.67. The values of the target E_b/N_o for the speech service in the uplink and downlink are defined as 5 dB and 10 dB, respectively. The simulation considers 50 speech users, which are randomly distributed in the UTRAN cell. In the simulation, the users start the services gradually. When a service request is received, the evaluation models investigate the availability of the network resources in the uplink and downlink.

The simulation results are presented in Figure 6-1 and Figure 6-2. Figure 6-1 portraits and compares the total received interference at the base station by implementing both evaluation models.

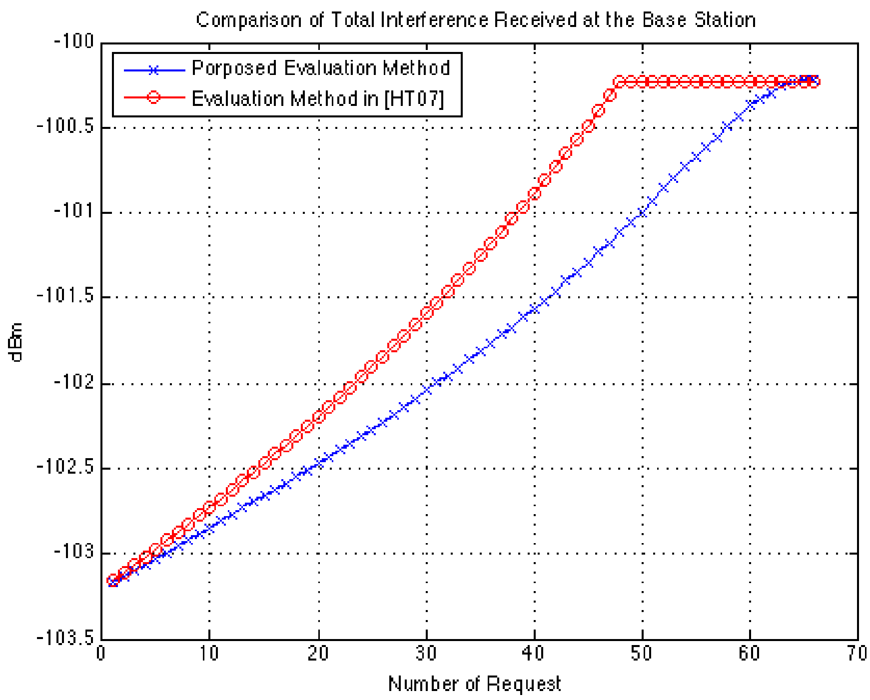


Figure 6-1 Comparison of the Total Interference Received at the Base Station

The simulation results presented in Figure 6-1 show that, at the early stage of the simulation, the values of the total received interference generated by both evaluation models are very close. The difference in these values begins to widen when the 10th service request is admitted. When the evaluation model based on [HT07] is employed, the value of the total received interference increases until the 49th service request arrives and achieves about -100.2 dBm. This is because the evaluation model determines that the resources in the uplink are not sufficient to admit the 49th service request. Without admitting new users, the total received interference remains the same value. In contrast, when the proposed evaluation model is used, the value of the total received interference increases until the 66th service request arrives.

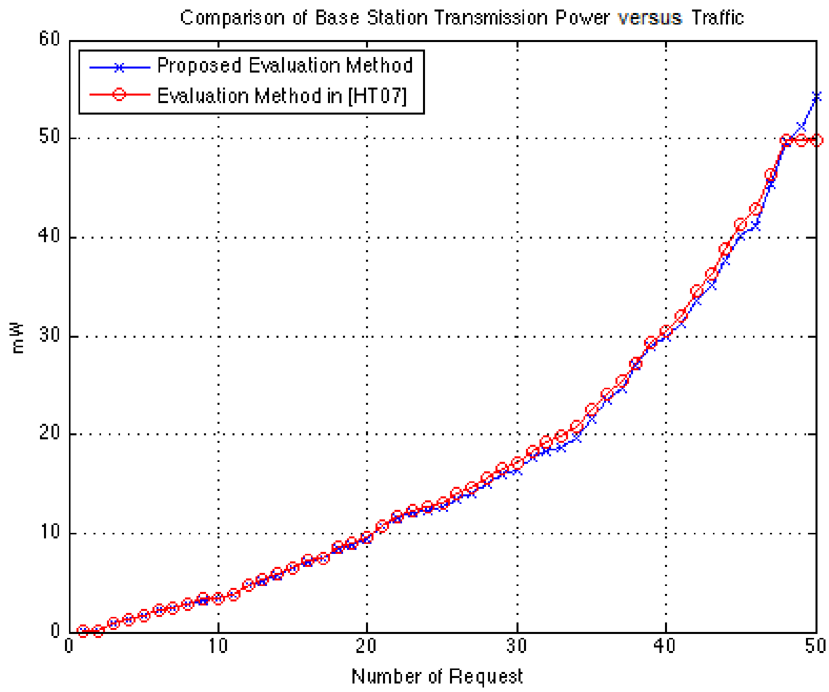


Figure 6-2 Comparison of Base Station Transmission Power versus Traffic

Figure 6-2 demonstrates and compares the total base station transmission power for traffic. The simulation results demonstrate that the values of the base station transmission power for traffic produced by both evaluation models are very close to each other throughout the simulation. The value generated by implementing the evaluation model based on [HT07] is slightly greater than the one produced by employing the proposed evaluation model since the admission of the 28th service request. Finally, when the evaluation model in [HT07] is used, the base station transmission power for traffic increases until the 49th service request arrives. This is because the evaluation model determines that the resources in the uplink are not sufficient to admit the 49th service request. Considering that the speech service is duplex, the 49th service request is not admitted in the downlink either.

All in all, the simulation results validate the proposed network resource availability evaluation model in UTRAN. The proposed evaluation model provides a solution to evaluate total received interference and base station transmission power without overestimating the inter-cell interference.

6.2 Validation of WLAN Network Resource Availability Evaluation Model

The simulations for validating WLAN network resource availability evaluation model are conducted by using OPNET version 10.5. The network topology is depicted in Figure 6-3. In the simulations, the wireless stations (denoted as *STA_1*, *STA_3*, etc) are distributed in the same *Basic Service Set*. These wireless stations communicate with an access point, *AP_1*. The *AP_1* is connected to an Ethernet switch, *Switch_1*. The *Switch_1* communicates with the wired stations, which are denoted as *CN_E_2*, *CN_E_4*, etc. The wired stations are acting as the correspondent nodes to the wireless stations. For example, *CN_E_2* is the correspondent node to *STA_1*.

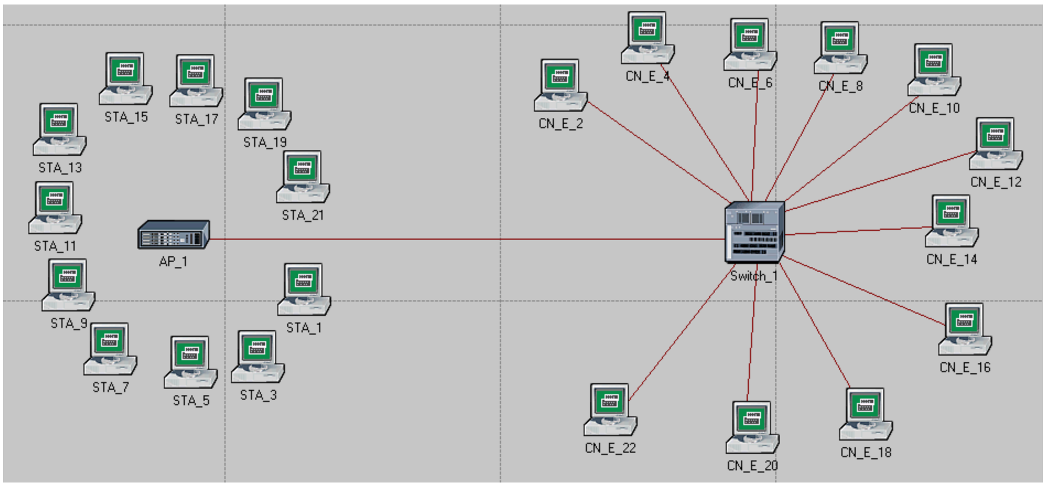


Figure 6-3 Network Topology and Architecture of the Simulations

For conversational services, such as VoIP and video call, both of the wireless and wired stations are acting as packet transmitters and receivers. For streaming services, such as video streaming and audio streaming, only the downlink is simulated and the wired stations act as the packet transmitters and the wireless stations act as the packet receivers. After the simulation starts to run, the starting time of each traffic source is randomly distributed within 0.1 to 1 second. The size and the inter arrival time of each packet depend on the characteristics of the simulated service.

6.2.1 Scenarios of the VoIP Service in the IEEE802.11b based WLAN

The evaluation model considers that the VoIP service has two service classes: basic and premium.

6.2.1.1 Evaluation Result of Resource Availability (Basic VoIP Service)

For the basic service class, the VoIP service employs G.723 as the codec and the data rate is 6.3 kbps. The G.723 codec acts as a CBR traffic source which generates 33 packets per second, which means that the packet inter arrival time is 30 ms and the packet payload is 24 bytes. Based on equation 5.23, $e\_t\_p_{VoIP\_basic}$ can be calculated as:

$$\begin{aligned} e\_t\_p_{VoIP\_basic} = {} & DIFS + e\_idle + phy\_mac\_hdr + \\ & t\_b \times \left(payload + ip\_udp\_rtp\_hdr\right) + \\ & delay + SIFS + delay + t\_ack \\ = {} & 0.00005 + 0.0003 + \left(0.000192 + 272/11000000\right) + \\ & \left(8/11000000\right) \times \left(24 + 40\right) + 0.000001 + 0.00001 + \\ & 0.000001 + 0.000248 \\ = {} & 0.00087327 \end{aligned}$$

Then, $p\_on_{VoIP\_basic}$ can be derived based on equation 5.22 as:

$$\begin{aligned} p\_on_{VoIP\_basic} &= e\_t\_p_{VoIP} / in\_p_{VoIP} \\ &= 0.00087327/0.03 \\ &= 0.0291 \end{aligned}$$

Based on equation 5.21, assuming only the VoIP service users are in the network, the value of e_p can be obtained:

$$\begin{aligned} e_p &= \sum_s p\_on_s \times n_s / N \\ &= p\_on_{VoIP\_basic} \times n_{VoIP\_basic} / N \\ &= p\_on_{VoIP\_basic} \\ &= 0.0291 \end{aligned}$$

Because this scenario only considers the VoIP service, N (N is the number of existing connections in the network plus the new requested connection) is equal to n_s (n_s is the number of connections of service type s). Therefore, e_p is equal to $p\_on_{VoIP\_basic}$. Then, e_{ncp} can be calculated by using equation 5.20. Assuming there are 17 basic VoIP users, the number of user connections in the WLAN is 34 and e_{ncp} can be derived as:

$$\begin{aligned} e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1-e_p)^{(N-i)} \\ &= \sum_{i=0}^{34} i \times \binom{34}{i} \times 0.0291^i \times (1-0.0291)^{(34-i)} \\ &= 0.9897091 \end{aligned}$$

e_{ncp} is smaller than 1, which means, on average, there is less than one packet in contention to access the network channel. Therefore, when there are 17 users, the contention to access network channel is still low and the quality requirements for the VoIP service can be satisfied. Assuming the 18th basic VoIP user is connected to the WLAN, the number of user connections is increased to 36 and e_{ncp} can be derived as:

$$\begin{aligned} e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1-e_p)^{(N-i)} \\ &= \sum_{i=0}^{36} i \times \binom{36}{i} \times 0.0291^i \times (1-0.0291)^{(36-i)} \\ &= 1.047927 \end{aligned}$$

e_{ncp} slightly exceeds 1. As a result, the WLAN should be able to support 17 basic VoIP users.

6.2.1.2 OPNET Simulation Configuration and Results (Basic VoIP Service)

For duplex services, both of the wireless and wired stations are acting as packet transmitters and receivers. Therefore, considering G.723 is used as the voice codec, each wireless and wired station will transmit a packet every 30 ms. The size of each packet is 64 bytes, involving the IP/UDP/RTP headers which are 40 bytes. The traffic generation parameters are shown in Table 6.1.

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (60)
OFF State Time (seconds)	constant (0)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.03)
Packet Size (bytes)	constant (64) (*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.1 Basic VoIP Traffic Generation Parameters

The OPNET simulation results are shown in Figure 6-4 to Figure 6-7. Figures 6-4 and 6-5 present the average delay of the packets received by a wireless station and the delay variation when there are 17 basic VoIP users and the value of e_{ncp} is 0.9897091.

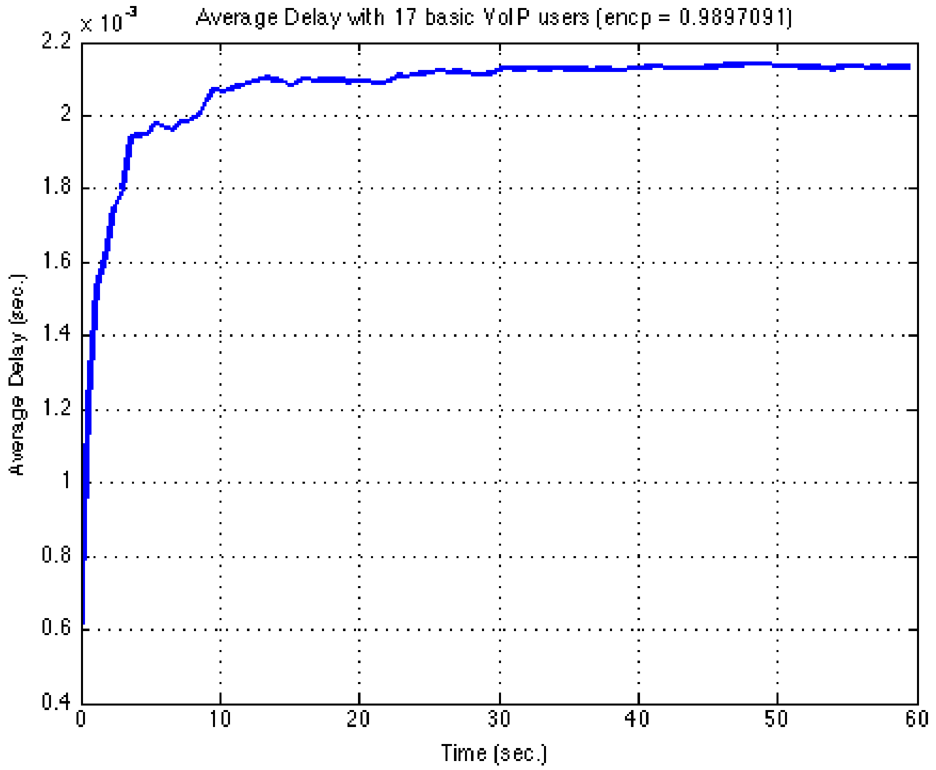


Figure 6-4 Average Delay with 17 basic VoIP users (e_{ncp} = 0.9897091)

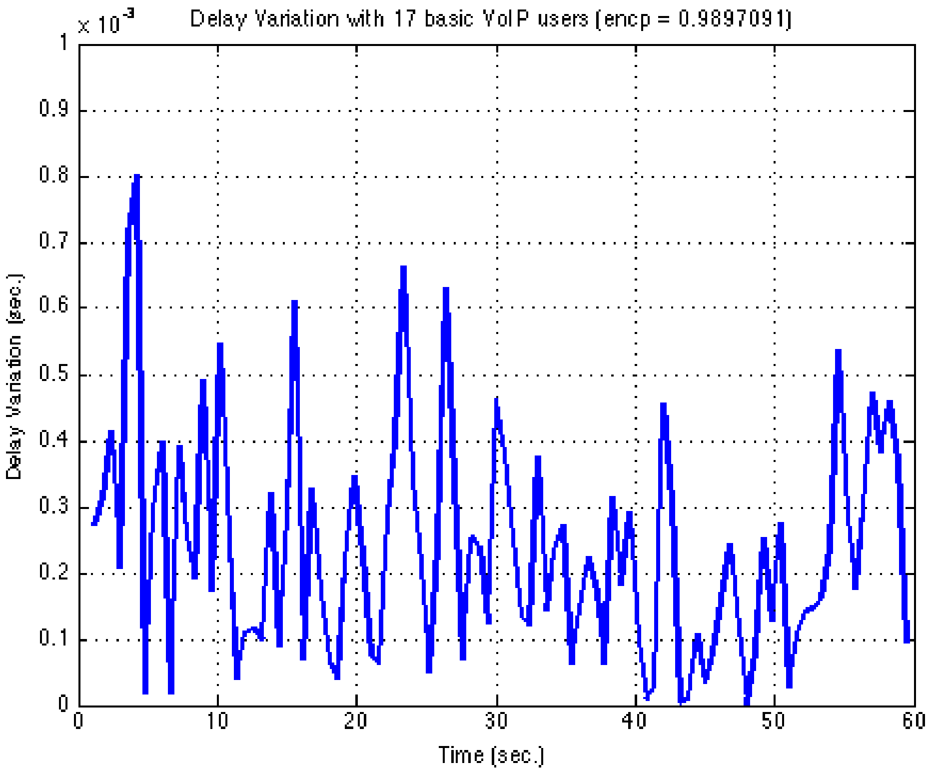


Figure 6-5 Delay Variation with 17 basic VoIP users (e_{ncp} = 0.9897091)

When there are 17 basic VoIP users in the WLAN, Figure 6-4 shows that the average packet delay is about 4 ms. Subjective evaluations have shown that, for conversational class services, the end-to-end delay should be less than 400 ms [Ser08]. Therefore, this value is acceptable for the real-time VoIP service. Figure 6-5 depicts fluctuations in packet delay and the delay variation is less than 0.8 ms. The delay variation is under the QoS constraint, which requires that the delay variation should be less than 1 ms [Ser08]. Therefore, it is acceptable. These results show that the WLAN network is able to provide acceptable quality of service when 17 basic VoIP users are in the network.

Figures 6-6 and 6-7 present the average delay of the packets received by a wireless station and the delay variation when there are 18 basic VoIP users and the value of e_{ncp} is 1.047927.

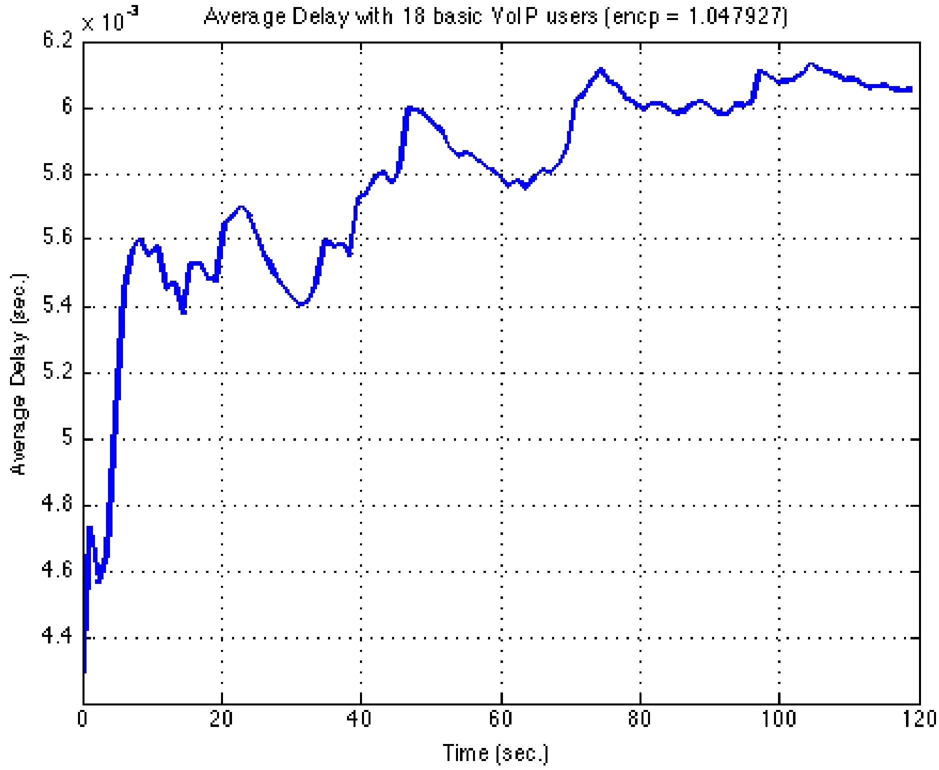


Figure 6-6 Average Delay with 18 basic VoIP users (e_{ncp} = 1.047927)

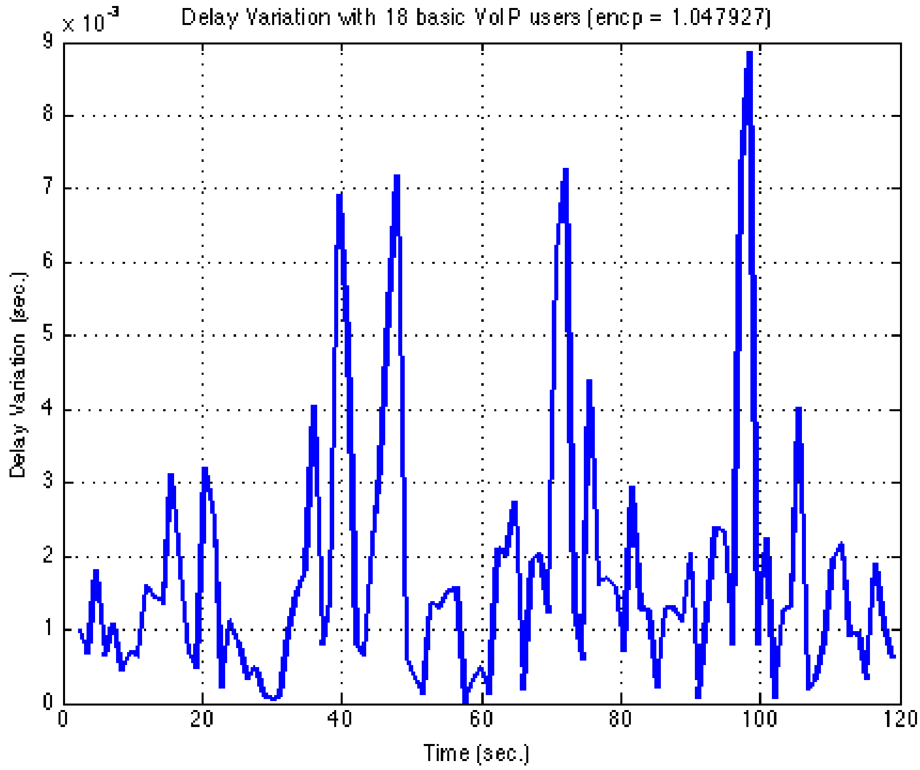


Figure 6-7 Delay Variation with 18 basic VoIP users (e_{ncp} = 1.047927)

When there are 18 basic VoIP users in the WLAN, Figure 6-6 shows that, after the fluctuations in the early stage of the simulation, the average packet delay stabilises at around 6.1 ms. This value is acceptable for the real-time VoIP service [Ser08]. However, Figure 6-7 depicts fluctuations in packet delay and the delay variation is greater than 1 ms and even reaches 9 ms. The delay variation exceeds the QoS constraint of 1 ms [Ser08] and is unacceptable. The WLAN network is unable to provide acceptable quality of service when 18 basic VoIP users are in the network and should not admit more than 17 basic VoIP users.

6.2.1.3 Evaluation Result of Resource Availability (Premium VoIP Service)

For the premium class, the VoIP service employs GSM610 as the codec and the data rate is 13.2 kbps. The GSM610 codec acts as a CBR traffic source and generates 50 packets per second, which means that the packet inter arrival time is 20 ms and the packet payload is 33 bytes. Therefore, $e\_t\_p_{VoIP\_premium}$ can be calculated as:

$$\begin{aligned}
e_t_p_{VoIP_premium} &= DIFS + e_idle + phy_mac_hdr + \\
&\quad t_b \times (payload + ip_udp_rtp_hdr) + \\
&\quad delay + SIFS + delay + t_ack \\
&= 0.00005 + 0.0003 + (0.000192 + 272/11000000) + \\
&\quad (8/11000000) \times (33 + 40) + 0.000001 + 0.00001 + \\
&\quad 0.000001 + 0.000248 \\
&= 0.0008788
\end{aligned}$$

Then, $p\_on_{VoIP\_premium}$ can be derived based on equation 5.22 as:

$$\begin{aligned}
p_on_{VoIP_premium} &= e_t_p_{VoIP_premium} / in_p_{VoIP_premium} \\
&= 0.0008788/0.02 \\
&= 0.04394
\end{aligned}$$

Assuming only the premium VoIP service users are in the network, the value of e_p can be obtained:

$$
\begin{aligned}
e_p &= \sum_s p\_on_s \times n_s / N \\
&= p\_on_{VoIP\_premium} \times n_{VoIP\_premium} / N \\
&= p\_on_{VoIP\_premium} \\
&= 0.04394
\end{aligned}
$$

Because this scenario only considers the VoIP service, N (N is the number of existing connections in the network plus the new requested connection) is equal to n_s (n_s is the number of connections of service type s). Therefore, e_p is equal to $p\_on_{VoIP\_premium}$. Then, e_{ncp} can be calculated by using equation 5.20. Assuming there are 11 premium VoIP users, the number of user connections in the WLAN is 22 and e_{ncp} can be derived as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1 - e_p)^{(N-i)} \\
&= \sum_{i=0}^{22} i \times \binom{22}{i} \times 0.04394^i \times (1 - 0.04394)^{(22-i)} \\
&= 0.9678
\end{aligned}
$$

e_{ncp} is smaller than 1, which means, on average, there is less than one packet in contention to access the network channel. Therefore, when there are 11 users, the contention to access network channel is still low and the quality requirements for the VoIP service can be satisfied. Assuming the 12th premium VoIP user is connected to the WLAN, the number of user connections is increased to 24 and e_{ncp} can be derived as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1 - e_p)^{(N-i)} \\
&= \sum_{i=0}^{24} i \times \binom{24}{i} \times 0.04394^i \times (1 - 0.04394)^{(24-i)} \\
&= 1.055782
\end{aligned}
$$

e_{ncp} slightly exceeds 1. As a result, the WLAN should be able to support 11 premium VoIP users.

6.2.1.4 OPNET Simulation Configuration and Results (Premium VoIP Service)

For duplex services, both of the wireless and wired stations are acting as packet transmitters and receivers. Considering GSM6.10 is used as the voice codec, each wireless and wired station will transmit a packet every 20 ms. The size of each packet is 73 bytes, involving the IP/UDP/RTP headers which are 40 bytes. The traffic generation parameters are shown in Table 6.2.

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (60)
OFF State Time (seconds)	constant (0)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.02)
Packet Size (bytes)	constant (73) (*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.2 Premium VoIP Traffic Generation Parameters

The OPNET simulation results are shown in Figure 6-8 to Figure 6-11. Figures 6-8 and 6-9 present the average delay of the packets received by a wireless station and the delay variation when there are 11 premium VoIP users and the value of e_{ncp} is 0.9678.

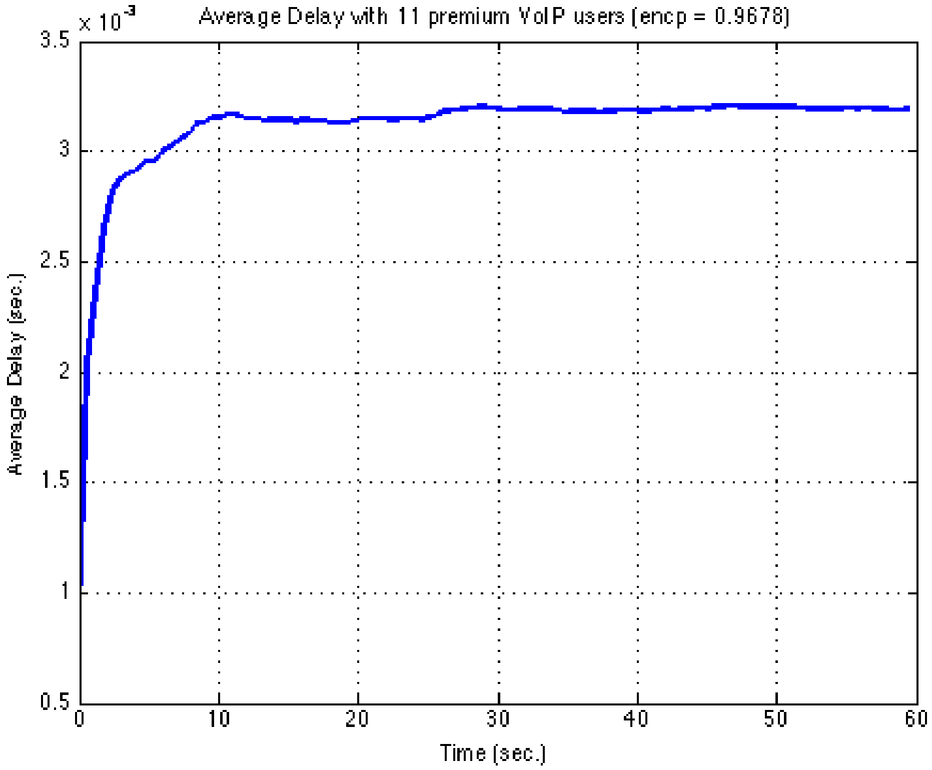


Figure 6-8 Average Delay with 11 premium VoIP users (e_{ncp} = 0.9678)

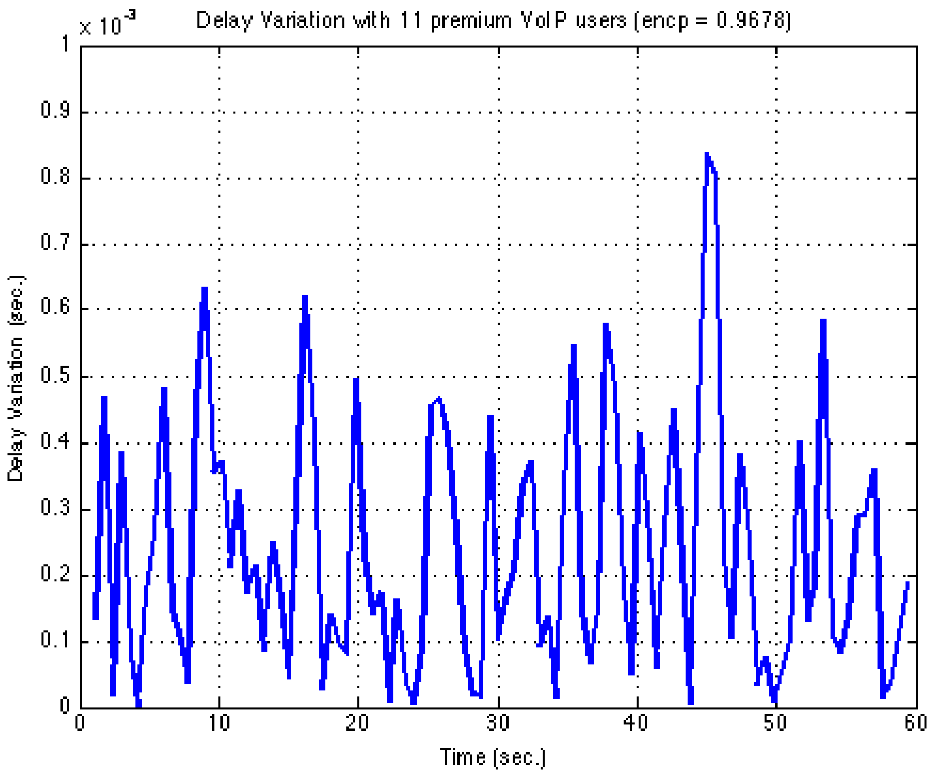


Figure 6-9 Delay Variation with 11 premium VoIP users (e_{ncp} = 0.9678)

When there are 11 premium VoIP users in the WLAN, Figure 6-8 shows that the average packet delay is about 3.2 ms, which is under the constraint of 400 ms [Ser08]. Therefore, this value is acceptable for the real-time VoIP service. Moreover, Figure 6-9 depicts fluctuations in packet delay and the delay variation is less than 0.85 ms, which is under the constraint of 1 ms [Ser08]. These results show that the WLAN network is capable of providing acceptable quality of service when 11 premium VoIP users are in the network.

Figures 6-10 and 6-11 present the average delay of the packets received by a wireless station and the delay variation at a wireless station when there are 12 premium VoIP users and the value of e_{ncp} is 1.055782.

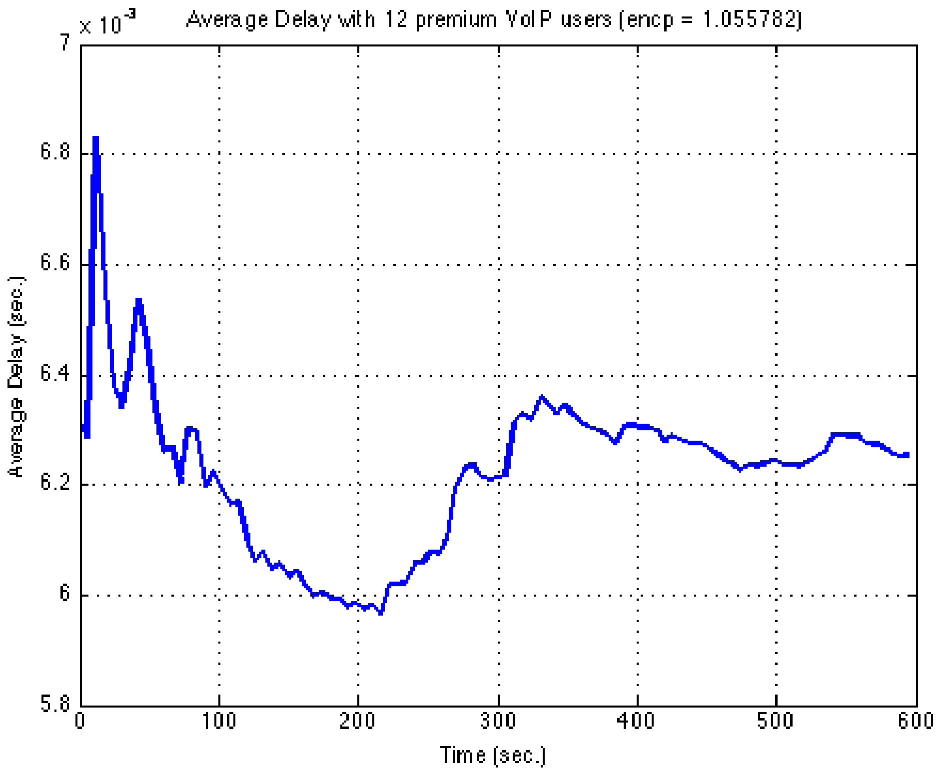


Figure 6-10 Average Delay with 12 premium VoIP users (e_{ncp} = 1.055782)

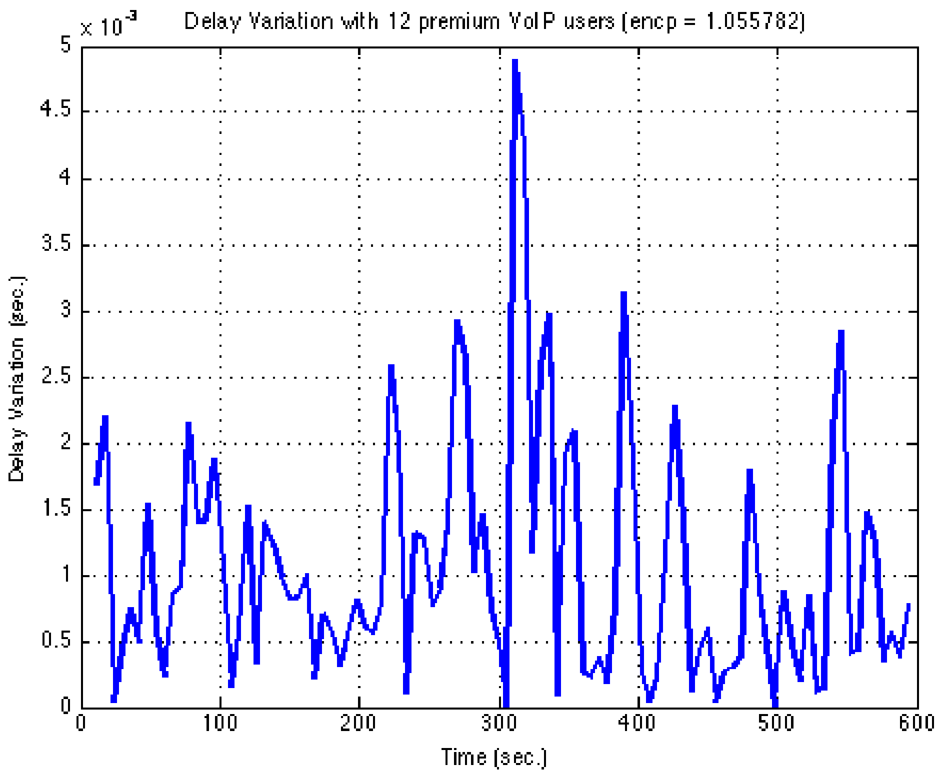


Figure 6-11 Delay Variation with 12 premium VoIP users (e_{ncp} = 1.055782)

When there are 12 premium VoIP users in the WLAN, Figure 6-10 shows that, after the fluctuations in the early stage of the simulation, the average packet delay stabilises at around 6.2 ms. This value is acceptable for the real-time VoIP service [Ser08]. However, Figure 6-11 depicts fluctuations in packet delay and the delay variation is greater than 1 ms. The delay variation exceeds the constraint of 1 ms [Ser08] and is unacceptable. The WLAN network is unable to provide acceptable quality of service when 12 premium VoIP users are in the network and should not admit more than 11 premium VoIP users.

6.2.2 Scenarios of the Video Call Service in the IEEE802.11b based WLAN

The video call service comprises two parts: video and audio. Two service classes, basic and premium, are considered in the evaluation model.

6.2.2.1 Evaluation Results of Resource Availability (Basic Video Call Service)

For the basic service class, the evaluation model assumes that the video part of the video call service employs H.263 as the codec. The H.263 codec acts as a CBR traffic source and generates 10 frames per second and the frame inter arrival time is 100 ms. The video data rate is 64 kbps, which means that the frame payload is 800 bytes. However, considering the instability of the radio channel, the size of a packet transmitted in a wireless network is suggested to be around 100 bytes [Wen03]. The evaluation model assumes that, for the H.263 codec, one video frame is divided into 8 slices. On average, the size of one slice is 100 bytes. Therefore, for the video part of the basic video call service, based on equation 5.24, $e\_t\_p_{video\_call\_video\_part\_basic}$ can be calculated as:

$$
\begin{aligned}
e\_t\_p_{video\_call\_video\_part\_basic} &= \begin{pmatrix} DIFS + e\_idle + phy\_mac\_hdr + \\ t\_b \times \left(payload + ip\_udp\_rtp\_hdr \right) + \\ delay + SIFS + delay + t\_ack \end{pmatrix} \times 8 \\
&= \begin{pmatrix} 0.00005 + 0.0003 + \left(0.000192 + 272/11000000 \right) + \\ \left(8/11000000 \right) \times \left(100 + 40 \right) + 0.000001 + 0.00001 + \\ 0.000001 + 0.000248 \end{pmatrix} \times 8 \\
&= 0.007428
\end{aligned}
$$

Then, $p\_on_{video\_call\_video\_part\_basic}$ can be derived based on equation 5.24 as:

$$
\begin{aligned}
p\_on_{video\_call\_video\_part\_basic} &= e\_t\_p_{video\_call\_video\_part\_basic} / in\_p_{video\_call\_video\_part\_basic} \\
&= 0.0074284/0.1 \\
&= 0.074284
\end{aligned}
$$

The evaluation model assumes that G.723 is used as the codec for the audio part of the basic video call service. The G.723 codec generates 33 packets per second and its data rate is 6.3 kbps. That means the packet inter arrival time is 30 ms and, on average, the packet payload is 24 bytes. Therefore, $e\_t\_p_{video\_call\_audio\_part}$ can be calculated as:

$$
\begin{aligned}
e\_t\_p_{video\_call\_audio\_part} &= DIFS + e\_idle + phy\_mac\_hdr + \\
&\quad t\_b \times (payload + ip\_udp\_rtp\_hdr) + \\
&\quad delay + SIFS + delay + t\_ack \\
&= 0.00005 + 0.0003 + (0.000192 + 272/11000000) + \\
&\quad (8/11000000) \times (24 + 40) + 0.000001 + 0.00001 + \\
&\quad 0.000001 + 0.000248 \\
&= 0.000873
\end{aligned}
$$

Then, $p\_on_{video\_call\_audio\_part}$ can be derived as:

$$
\begin{aligned}
p\_on_{video\_call\_audio\_part} &= e\_t\_p_{video\_call\_audio\_part} / in\_p_{video\_call\_audio\_part} \\
&= 0.000873/0.03 \\
&= 0.029109
\end{aligned}
$$

Based on equation 5.21, assuming only the video call service users are in the network, e_p can be calculated as:

$$
\begin{aligned}
e_p &= \sum_s p\_on_s \times n_s / N \\
&= p\_on_{video\_call\_audio\_part} \times \frac{n_{video\_call\_audio\_part}}{n_{video\_call\_audio\_part} + n_{video\_call\_video\_part\_basic}} + \\
&\quad p\_on_{video\_call\_video\_part\_basic} \times \frac{n_{video\_call\_video\_part\_basic}}{n_{video\_call\_audio\_part} + n_{video\_call\_video\_part\_basic}} \\
&= \frac{p\_on_{video\_call\_audio\_part} + p\_on_{video\_call\_video\_part\_basic}}{2} \\
&= 0.0516965
\end{aligned}
$$

$n_{video\_call\_audio\_part}$ is the number of the connections introduced by the audio part of the video call service. $n_{video\_call\_video\_part\_basic}$ is the number of the connections introduced by the basic video part of the video call service. Therefore, considering that the video call service is duplex and consists of two parts, one video call user will introduce four connections to the WLAN. Then, based on equation 5.20, e_{ncp} can be calculated. Assuming there are 4 basic video call users, the number of user connections in the WLAN is 16 and e_{ncp} can be derived as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times \left(1-e_p\right)^{(N-i)} \\
&= \sum_{i=0}^{16} i \times \binom{16}{i} \times 0.0516965^i \times (1-0.0516965)^{(16-i)} \\
&= 0.827142
\end{aligned}
$$

e_{ncp} is smaller than 1, which means, on average, there is less than one packet in contention to access the network channel. Therefore, when there are 4 users, the contention to access network channel is still low and the quality requirements for the video call service can be satisfied. Assuming the 5th basic video call user is connected to the WLAN, e_{ncp} can be derived as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times \left(1-e_p\right)^{(N-i)} \\
&= \sum_{i=0}^{20} i \times \binom{20}{i} \times 0.0516965^i \times (1-0.0516965)^{(20-i)} \\
&= 1.033927
\end{aligned}
$$

e_{ncp} slightly exceeds 1. The WLAN is estimated as being capable to support 4 basic video call users.

6.2.2.2 OPNET Simulation Configuration and Results (Basic Video Call Service)

The video call service is a duplex service which consists of the two parts: video and audio. Therefore, two types of packet transmitters are simulated. The audio packet transmitter uses the G.723 codec and transmits a packet every 30 ms. The size of each packet is 64 bytes, involving the IP/UDP/RTP headers which are 40 bytes. The video packet transmitter uses the H.263 codec and generates a frame every 100 ms. Each frame will be divided into 8 slices. Each slice will be encapsulated into a packet, whose size is 140 bytes, including the IP/UDP/RTP headers. The video packet transmitter has two traffic generation states: ON and OFF. The ON state lasts 50 ms and the OFF state lasts 50 ms. The video packets will only be transmitted during the ON state and the inter transmission time is 6 ms. The traffic generation parameters are shown in Table 6.3 and Table 6.4.

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (60)
OFF State Time (seconds)	constant (0)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.03)
Packet Size (bytes)	constant (64) (*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.3 Audio Traffic Generation Parameters for the Basic Video Call Service

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (0.05)
OFF State Time (seconds)	constant (0.05)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.006)
Packet Size (bytes)	constant(140)(*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.4 Video Traffic Generation Parameters for the Basic Video Call Service

The OPNET simulation results are shown in Figure 6-12 to Figure 6-15. Figures 6-12 and 6-13 present the average delay of the packets received by a wireless station and the delay variation at a wireless station when there are 4 basic video call users and the value of e_{ncp} is 0.8271418.

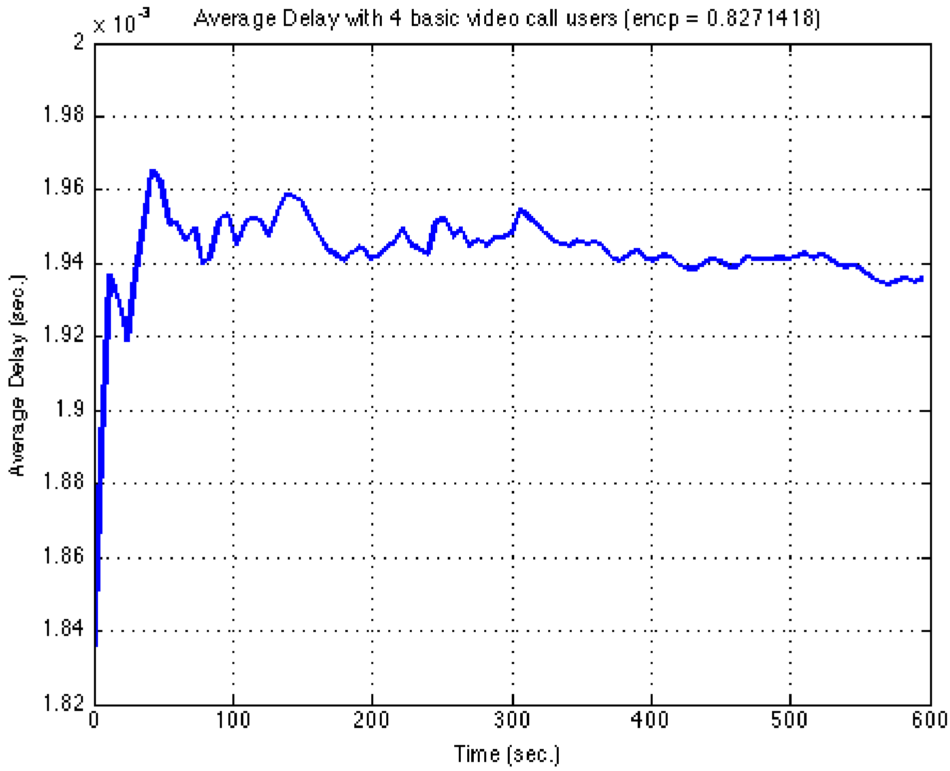


Figure 6-12 Average Delay with 4 basic video call users (e_{ncp} = 0.8271418)

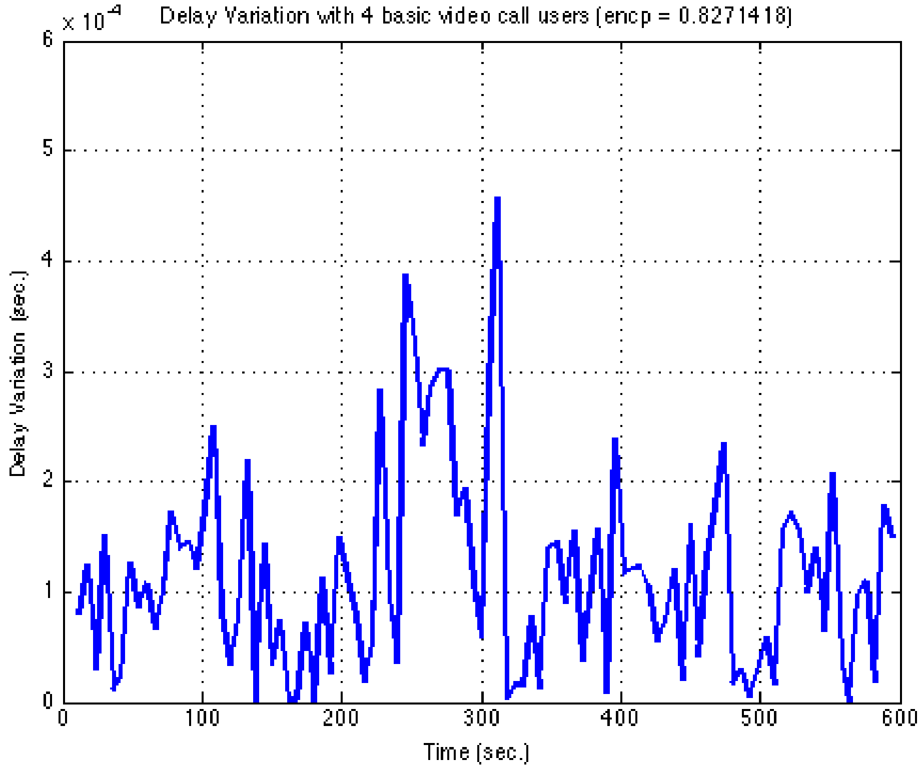


Figure 6-13 Delay Variation with 4 basic video call users (e_{ncp} = 0.8271418)

When there are 4 basic video call users in the WLAN, Figure 6-12 shows that the average packet delay is about 1.94 ms, which is under the constraint of 400 ms [Ser08]. Therefore, this value is acceptable for the real-time video call service. Moreover, Figure 6-17 depicts fluctuations in packet delay and the delay variation is less than 0.5 ms, which is under the constraint of 1 ms [Ser08]. These results show that the WLAN is capable of providing acceptable quality of service when there are 4 basic video call users.

Figures 6-14 and 6-15 present the average delay of the packets received by a wireless station and the delay variation at a wireless station when there are 5 basic video call users and the value of e_{ncp} is 1.033927.

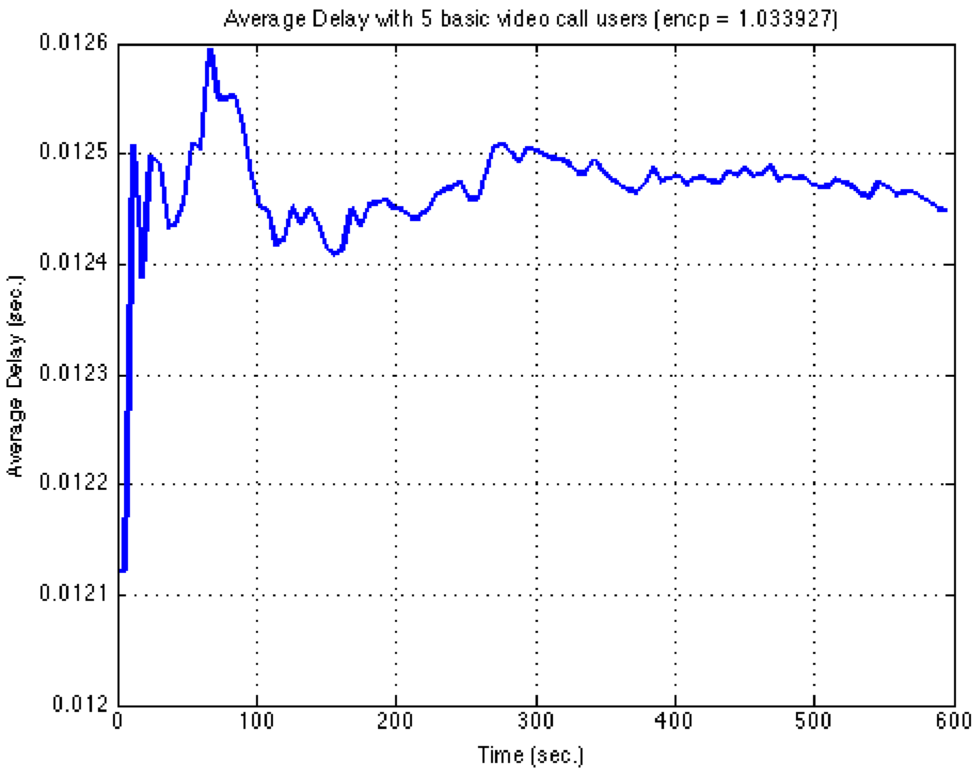


Figure 6-14 Average Delay with 5 basic video call users (e_{ncp} = 1.033927)

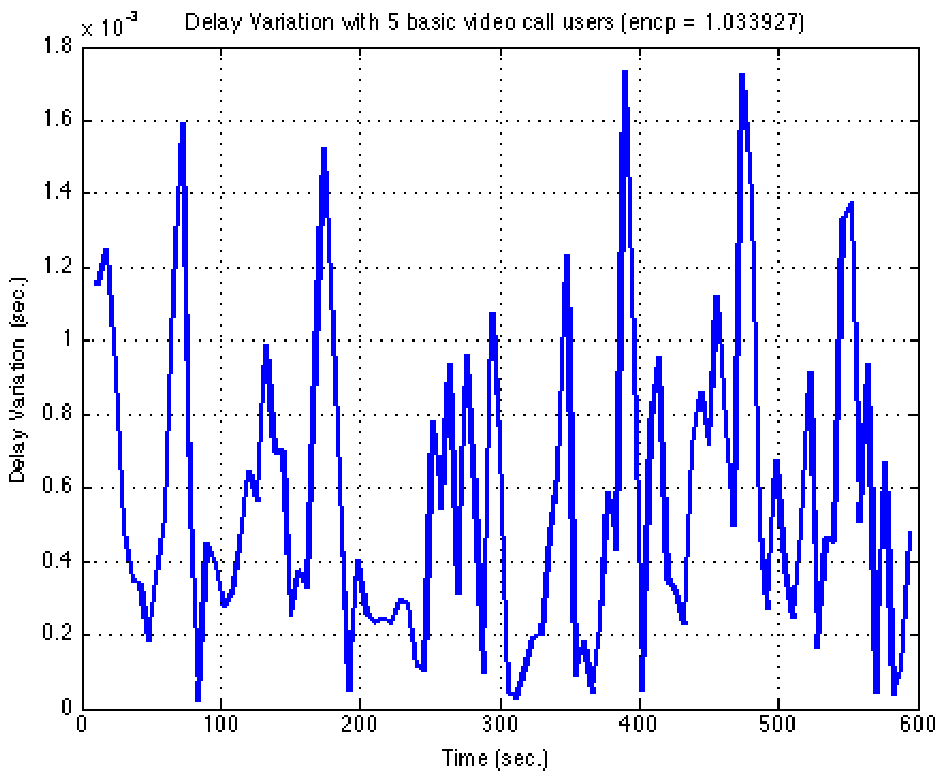


Figure 6-15 Delay Variation with 5 basic video call users (e_{ncp} = 1.033927)

When there are 5 basic video call users in the WLAN, Figure 6-14 shows that, after the fluctuations in the early stage of the simulation, the average packet delay stabilises at around 12.45 ms. This value is acceptable for conversational class services which require that the end-to-end delay should be less than 400 ms [Ser08]. However, Figure 6-15 depicts fluctuations in packet delay and the delay variation is greater than 1 ms. This value is unacceptable for conversational class services which require that the delay variation should be less than 1 ms [Ser08]. These results show that the WLAN is unable to provide acceptable quality of service when 5 basic video call users are in the network and it should be admit more than 4 basic video call users.

6.2.2.3 Evaluation Results of Resource Availability (Premium Video Call Service)

For the premium service class, the evaluation model assumes that the video part employs H.263 as the codec. The H.263 codec acts as a CBR traffic source and generates 10 frames per second and the frame inter arrival time is 100 ms. The video data rate is 128 kbps, which means that the frame payload is 1600 bytes. It is assumed that one video frame is divided into 8 slices and on average the size of one slice is 200 bytes. Therefore, for the video part of the premium video call service, $e\_t\_p_{video\_call\_video\_part\_premium}$ can be calculated as:

$$\begin{aligned}
e_t_p_{video_call_video_part_premium} &= \begin{pmatrix} DIFS + e_idle + phy_mac_hdr + \\ t_b \times (payload + ip_udp_rtp_hdr) + \\ delay + SIFS + delay + t_ack \end{pmatrix} \times 8 \\
&= \begin{pmatrix} 0.00005 + 0.0003 + (0.000192 + 272/11000000) + \\ (8/11000000) \times (200 + 40) + 0.000001 + 0.00001 + \\ 0.000001 + 0.000248 \end{pmatrix} \times 8 \\
&= 0.00801
\end{aligned}$$

Then, $p\_on_{video\_call\_video\_part\_premium}$ can be derived based on equation 5.24 as:

$$\begin{aligned}
p_on_{video_call_video_part_premium} &= \frac{e_t_p_{video_call_video_part_premium}}{in_p_{video_call_video_part_premium}} \\
&= 0.00801/0.1 \\
&= 0.0801
\end{aligned}$$

Similar to the basic service class, the evaluation model assumes that G.723 is used as the codec for the audio part of the premium video call service. The G.723 codec generates 33 packets per second and its data rate is 6.3 kbps. Based on the value of $p\_on_{video\_call\_audio\_part}$ calculated in subsection 6.2.2.1 and equation 5.21, assuming only the video call service users are in the network, e_p can be calculated as:

$$\begin{aligned}
e_p &= \sum_s p_on_s \times n_s / N \\
&= p_on_{video_call_audio_part} \times \frac{n_{video_call_audio_part}}{n_{video_call_audio_part} + n_{video_call_video_part_premium}} + \\
&\quad p_on_{video_call_video_part_premium} \times \frac{n_{video_call_video_part_premium}}{n_{video_call_audio_part} + n_{video_call_video_part_premium}} \\
&= \frac{p_on_{video_call_audio_part} + p_on_{video_call_video_part_premium}}{2} \\
&= 0.05460545
\end{aligned}$$

$n_{video\_call\_audio\_part}$ is the number of the connections introduced by the audio part of the video call service. $n_{video\_call\_video\_part\_premium}$ is the number of the connections introduced by the premium video part of the video call service. Therefore, considering that the video call service is duplex and consists of two parts, one video call user will introduce four connections to the WLAN. Then, based on equation 5.20, e_{ncp} can be calculated. Assuming there are 4 premium video call users, the number of user connections in the WLAN is 16 and e_{ncp} can be derived as:

$$\begin{aligned} e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times \left(1-e_p\right)^{(N-i)} \\ &= \sum_{i=0}^{16} i \times \binom{16}{i} \times 0.05460545^i \times \left(1-0.05460545\right)^{(16-i)} \\ &= 0.8736873 \end{aligned}$$

e_{ncp} is smaller than 1, which means, on average, there is less than one packet in contention to access the network channel. Therefore, when there are 4 users, the contention to access network channel is still low and the quality requirements for the video call service can be satisfied. Assuming the 5th premium video call user is connected to the WLAN, e_{ncp} can be derived as:

$$\begin{aligned} e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times \left(1-e_p\right)^{(N-i)} \\ &= \sum_{i=0}^{20} i \times \binom{20}{i} \times 0.05460545^i \times \left(1-0.05460545\right)^{(20-i)} \\ &= 1.092109 \end{aligned}$$

e_{ncp} slightly exceeds 1. The WLAN is estimated as being able to support 4 premium video call users.

6.2.2.4 OPNET Simulation Configuration and Results (Premium Video Call Service)

The video call service is a duplex service which consists of the two parts: video and audio. The audio packet transmitter uses the G.723 codec and transmits a packet every 30 ms. The size of each packet is 64 bytes, involving the IP/UDP/RTP headers which are 40 bytes. The video packet transmitter uses the H.263 codec and generates a frame every 100 ms. Each frame will be divided into 8 slices. Each slice will be encapsulated into a packet, whose size is 240 bytes, including the IP/UDP/RTP headers. The video packet transmitter has two traffic generation states: ON and OFF. The ON state lasts 50 ms and the OFF state lasts 50 ms. The video packets will only be transmitted during the ON state and the inter transmission time is 6 ms. The traffic generation parameters are shown in Table 6.5 and Table 6.6.

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (60)
OFF State Time (seconds)	constant (0)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.03)
Packet Size (bytes)	constant (64) (*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.5 Audio Traffic Generation Parameters for the Premium Video Call Service

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (0.05)
OFF State Time (seconds)	constant (0.05)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.006)
Packet Size (bytes)	constant(240)(*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.6 Video Traffic Generation Parameters for the Premium Video Call Service

The OPNET simulation results are shown in Figure 6-16 to Figure 6-19. Figures 6-16 and 6-17 present the average delay of the packets received by a wireless station and the delay variation when there are 4 premium video call users and the value of e_{ncp} is 0.8736873.

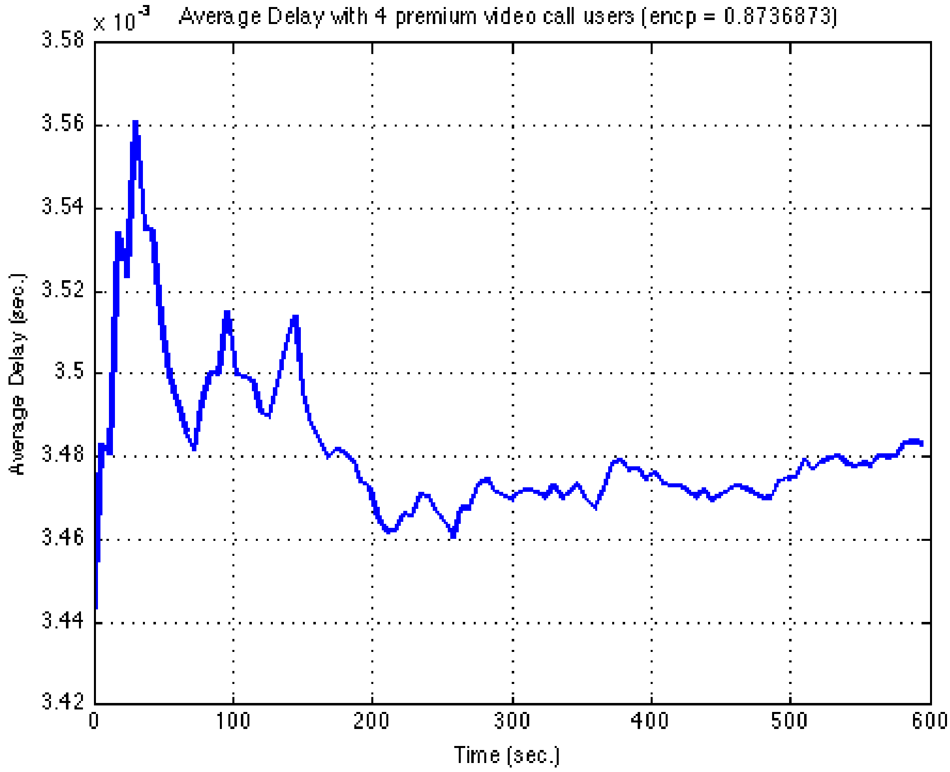


Figure 6-16 Average Delay with 4 premium video call users (e_{ncp} = 0.8736873)

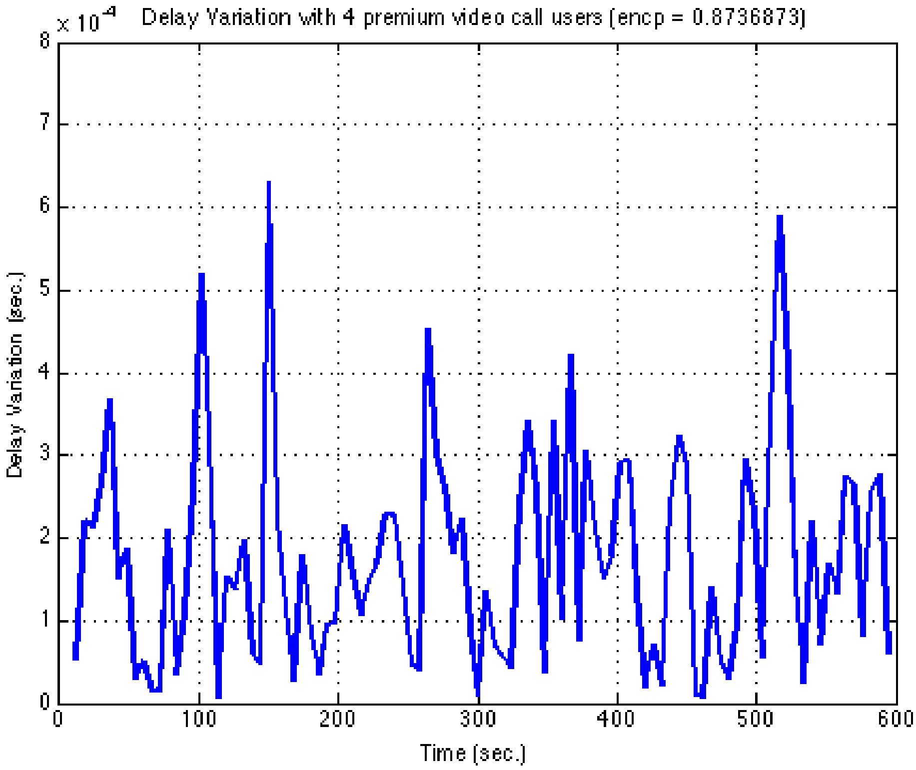


Figure 6-17 Delay Variation with 4 premium video call users (e_{ncp} = 0.8736873)

When there are 4 premium video call users in the WLAN, Figure 6-16 shows that, after fluctuations in the early stage of the simulation, the average packet delay stabilises at 3.48 ms, which is under the constraint of 400 ms [Ser08]. Therefore, this value is acceptable for the real-time video call service. Moreover, Figure 6-17 depicts fluctuations in packet delay and the delay variation is under 0.7 ms. These results show that the WLAN is able to provide acceptable quality of service when 4 premium video call users are in the network.

Figures 6-18 and 6-19 present the average delay of the packets received by a wireless station and the delay variation when there are 5 premium video call users and the value of e_{ncp} is 1.092109.

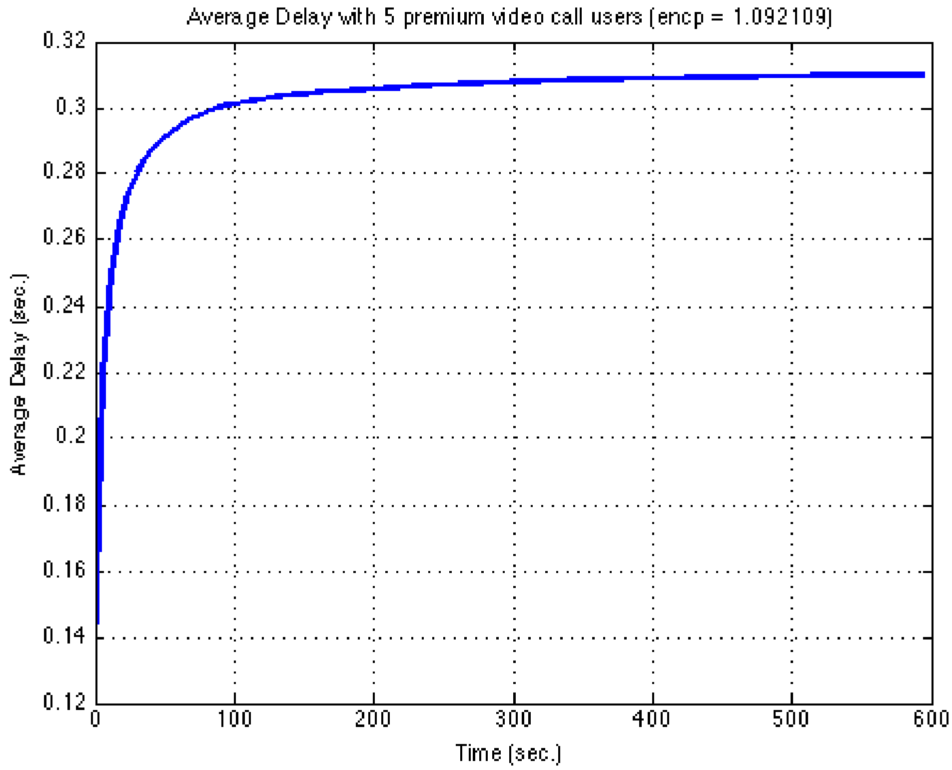


Figure 6-18 Average Delay with 5 premium video call users (e_{ncp} = 1.092109)

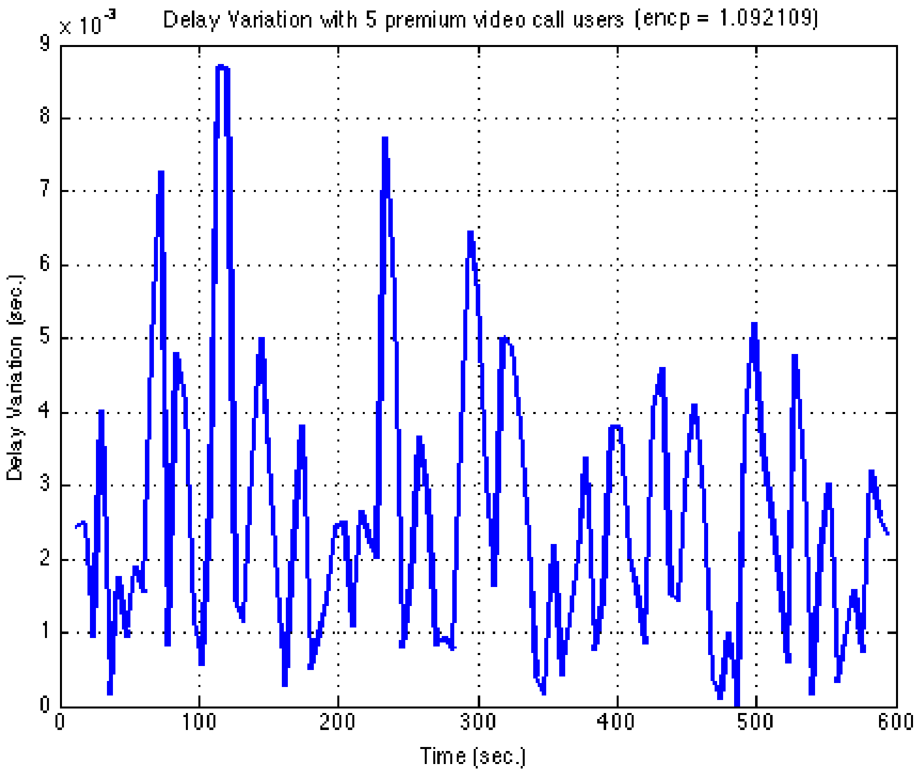


Figure 6-19 Delay Variation with 5 basic video call users (e_{ncp} = 1.092109)

When there are 5 premium video call users in the WLAN, Figure 6-18 shows that the average packet delay increases dramatically as simulation runs and achieves about 0.31 seconds, which exceeds the constraint of 400 ms [Ser08]. This value is unacceptable for the real-time video call service. Therefore, when there are 5 premium video call users, the WLAN cannot provide acceptable quality of service and should not admit more than 4 premium video call users.

6.2.3 Scenarios of the Audio Streaming Service in the IEEE802.11b based WLAN

The evaluation model considers that the Audio Streaming service has two service classes: basic and premium.

6.2.3.1 Evaluation Results of Resource Availability (Basic Audio Streaming Service)

The evaluation model assumes that the basic audio streaming service employs G.726 as the codec. The G.726 codec generates 50 packets per second, which means that the packet inter arrival time is 20 ms. For the basic service class, the data rate is assumed as 32 kbps, which means that the packet payload is 80 bytes. Therefore, $e\_t\_p_{audio\_str\_baisc}$ can be calculated as:

$$\begin{aligned} e\_t\_p_{audio\_str\_basic} &= DIFS + e\_idle + phy\_mac\_hdr + \\ &\quad t\_b \times (payload + ip\_udp\_rtp\_hdr) + \\ &\quad delay + SIFS + delay + t\_ack \\ &= 0.00005 + 0.0003 + (0.000192 + 272/11000000) + \\ &\quad (8/11000000) \times (80 + 40) + 0.000001 + 0.00001 + \\ &\quad 0.000001 + 0.000248 \\ &= 0.000914 \end{aligned}$$

Then, $p\_on_{audio\_str\_basic}$ can be derived based on equation 5.27as:

$$\begin{aligned} p\_on_{audio\_str\_baisc} &= e\_t\_p_{audio\_str\_basic} / in\_p_{audio\_str} \\ &= 0.000914/0.02 \\ &= 0.0457 \end{aligned}$$

Based on equation 5.21, assuming only the basic audio streaming service users are in the network the value of e_p can be obtained:

$$\begin{aligned} e_p &= \sum_s p\_on_s \times n_s / N \\ &= p\_on_{audio\_str\_basic} \times n_{audio\_str\_basic} / N \\ &= p\_on_{audio\_str\_baisc} \\ &= 0.0457 \end{aligned}$$

e_{ncp} can be calculated by using equation 5.20. Assuming there are 21 basic audio streaming users, the number of user connections in the WLAN is 21 and e_{ncp} can be derived as:

$$\begin{aligned} e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1 - e_p)^{(N-i)} \\ &= \sum_{i=0}^{21} i \times \binom{21}{i} \times 0.0457^i \times (1 - 0.0457)^{(21-i)} \\ &= 0.9597 \end{aligned}$$

e_{ncp} is smaller than 1, which means, on average, there is less than one packet in contention to access the network channel. Therefore, when there are 21 users, the contention to access the network channel is still low and the quality requirements for the audio streaming service can be satisfied. Assuming the 22nd basic audio streaming user is connected to the WLAN, e_{ncp} can be calculated as:

$$\begin{aligned} e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1 - e_p)^{(N-i)} \\ &= \sum_{i=0}^{22} i \times \binom{22}{i} \times 0.0457^i \times (1 - 0.0457)^{(22-i)} \\ &= 1.0054 \end{aligned}$$

e_{ncp} slightly exceeds 1. The WLAN is estimated as being capable to support 21 basic audio streaming users.

6.2.3.2 OPNET Simulation Configuration and Results (Basic Audio Streaming)

The OPNET simulation focuses on the downlink traffic between the wired stations and the wireless stations. The wired stations act as the audio packet transmitters and the wireless stations are the receivers. Considering G.726 is used as the audio codec, each wired station will transmit a packet every 20 ms. The size of each packet is 120 bytes, involving the IP/UDP/RTP headers which are 40 bytes. The traffic generation parameters are shown in Table 6.7.

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (60)
OFF State Time (seconds)	constant (0)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.02)
Packet Size (bytes)	constant(120)(*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.7 Traffic Generation Parameters for the Basic Audio Streaming Service

The OPNET simulation results are shown in Figure 6-20 to Figure 6-25. Figures 6-20, 6-21 and 6-22 present the average delay of the packets received by a wireless station, the delay variation and the packet arrival rate at a wireless station when there are 21 basic audio streaming users and the value of e_{ncp} is 0.9597.

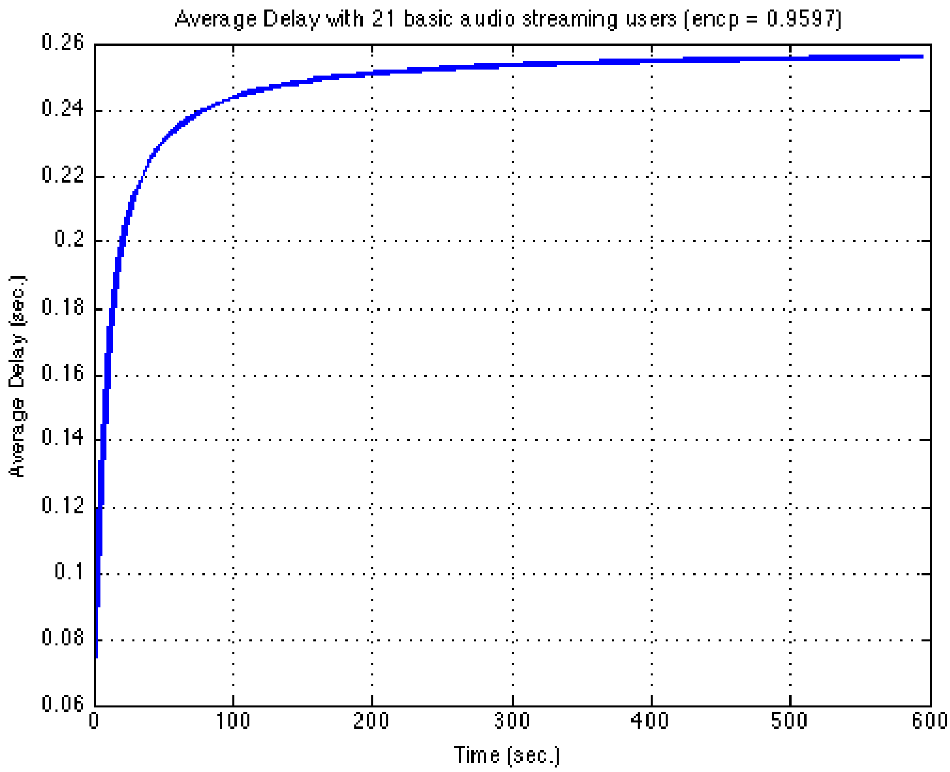


Figure 6-20 Average Delay with 21 basic audio streaming users (e_{ncp} = 0.9597)

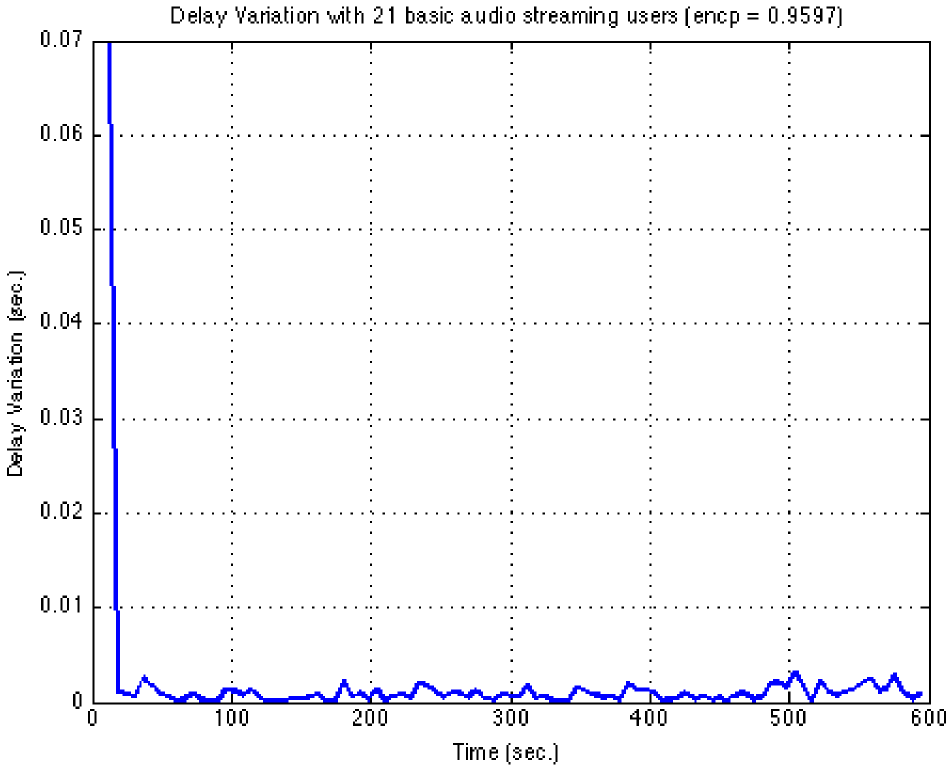


Figure 6-21 Delay Variation with 21 basic audio streaming users (e_{ncp} = 0.9597)

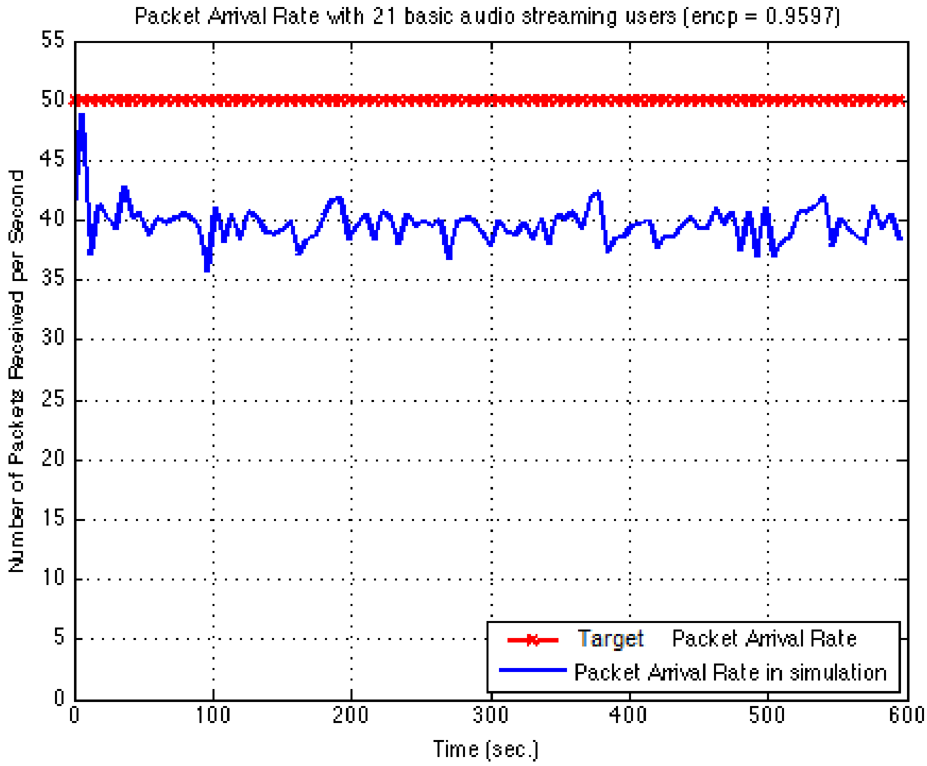


Figure 6-22 Packet Arrival Rate with 21 basic audio streaming users (e_{ncp} = 0.9597)

When there are 21 basic audio streaming users in the WLAN, Figure 6-20 shows that the average packet delay increases as simulation runs and achieves about 0.26 seconds. Figure 6-21 shows that the delay variation is mainly less than 20 ms. Comparing to conversational class services, the streaming class services have the same requirements for the bandwidth but can tolerate some delay variations by implement a dejitter buffer at the receiver [HT07]. [Ser08] requires that, for streaming class services, the start-up delay should be less than 10 seconds and the delay variation should be less than 2 seconds. Although the above simulation results in Figure 6-20 and Figure 6-21 comply with such constraints, Figure 6-22 depicts significant fluctuations in the packet arrival rates. Furthermore, the packet arrival rate does not achieve their target values. Based on the codecs implemented for the basic audio streaming service, the required packet arrival rate is 50 packets per second. Because Figure 6-30 shows that the packet arrival rate does not achieve the required values, it indicates that packet losses occur in the network. Figure 6-22 shows that the actual packet arrival rate is about 40 packets per second, which indicates that the packet loss rate is about 20%. [Ser08] requires that the packet loss rate for audio streaming service should be less than 1%. The simulation results demonstrate that, when there are 21 basic audio streaming users, the WLAN cannot provide acceptable quality of service.

Considering that the WLAN cannot support 21 basic audio streaming users, the number of users is reduced to be 20 and the corresponding value of e_{ncp} is 0.914. The simulation results are presented in Figures 6-23, 6-24 and 6-25.

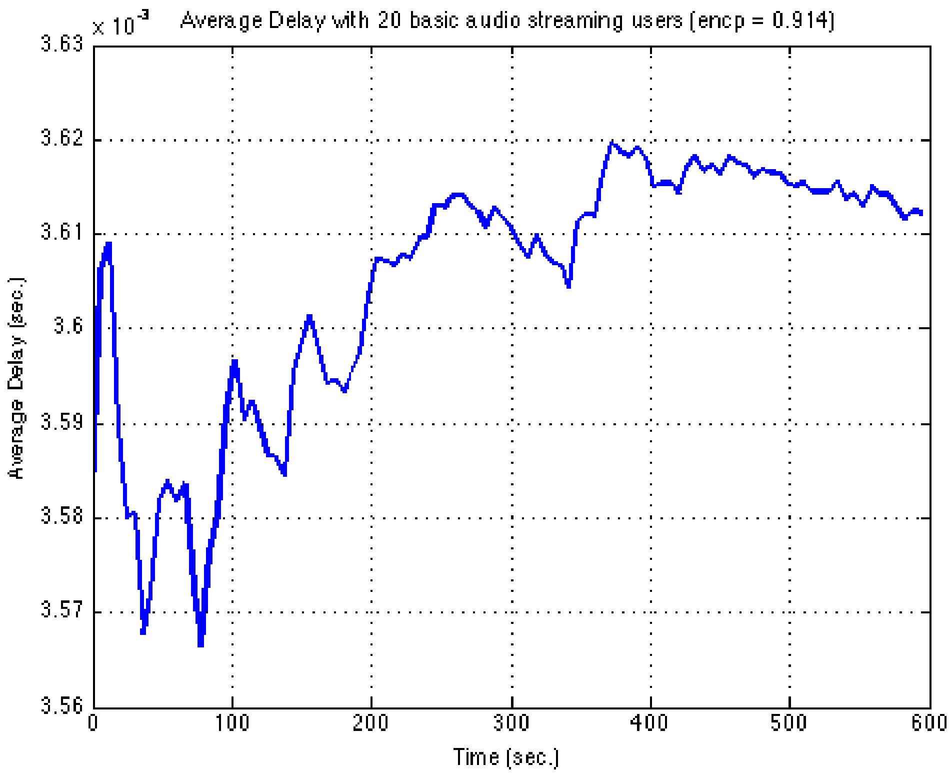


Figure 6-23 Average Delay with 20 basic audio streaming users (e_{ncp} = 0.914)

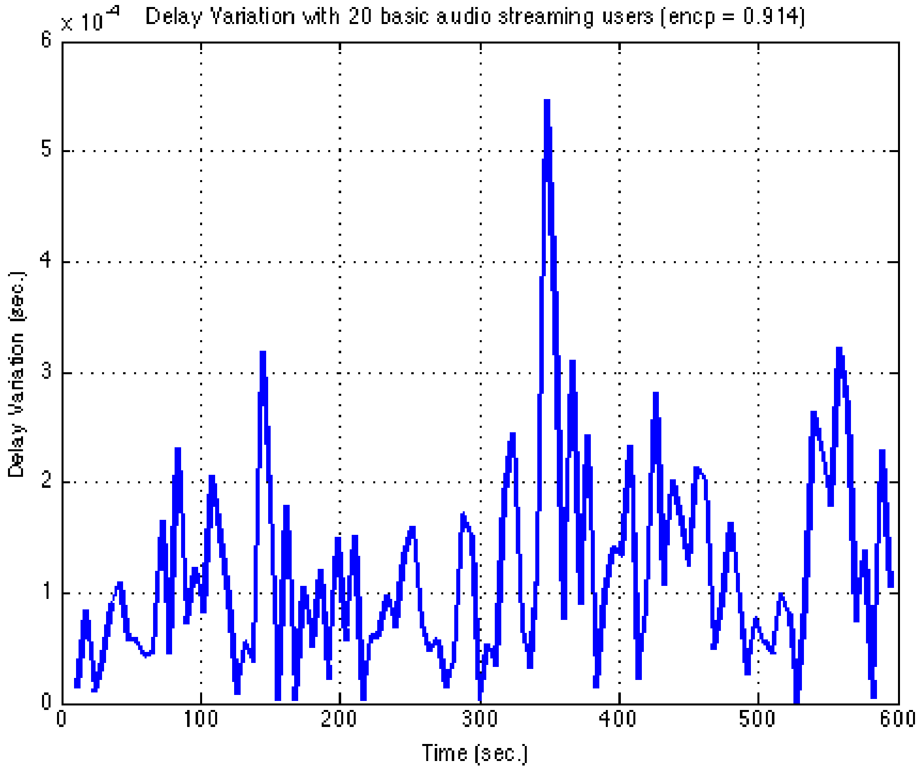


Figure 6-24 Delay Variation with 20 basic audio streaming users (e_{ncp} = 0.914)

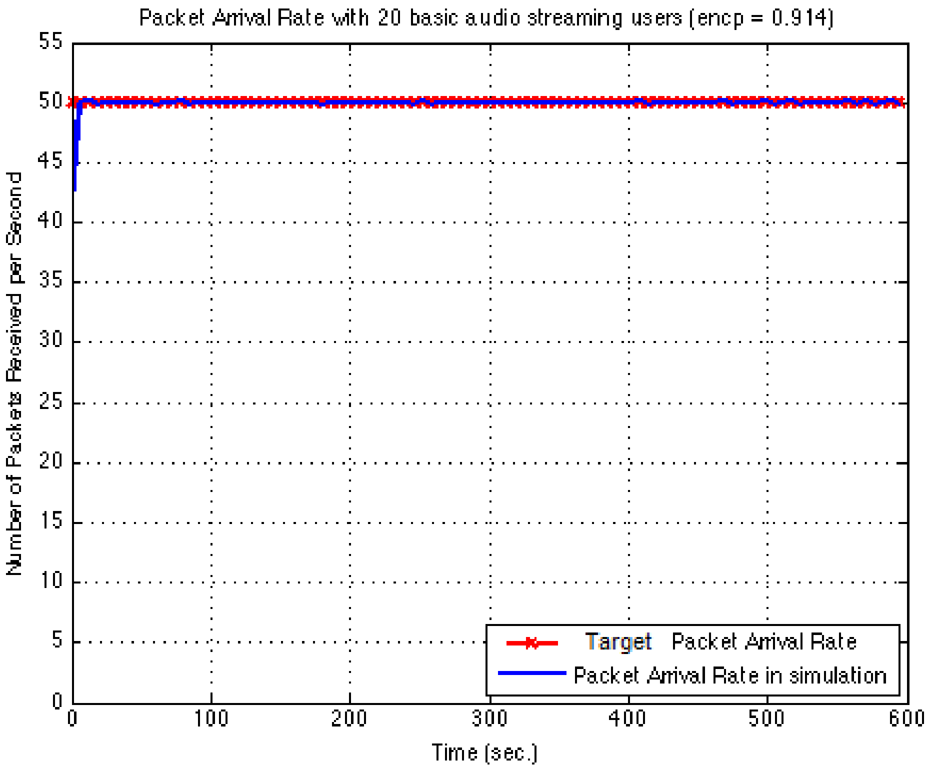


Figure 6-25 Packet Arrival Rate with 20 basic audio streaming users (e_{ncp} = 0.914)

When there are 20 basic audio streaming users in the WLAN, Figure 6-23 shows that, after the fluctuations in the early stage of the simulation, the average packet delay begins to stabilise and achieve about 3.61 ms. Figure 6-24 shows that the delay variation is less than 0.55 ms. The packet delay and delay variation even comply with the constraint for conversational class services. Furthermore, Figure 6-25 depicts that the packet arrival rate is stable and the same as the target value, which is 50 packets per second. Therefore, the WLAN is capable of providing acceptable quality of service when there are 20 basic audio streaming users.

6.2.3.3 Evaluation Results of Resource Availability (Premium Audio Streaming Service)

The evaluation model assumes that the audio streaming service employs G.726 as the codec. The G.726 codec generates 50 packets per second, which means that the packet inter arrival time is 20 ms. For the premium service class, the data rate is assumed as 64 kbps, which means that the packet payload is 160 bytes. Therefore, $e\_t\_p_{audio\_str\_premium}$ can be calculated as:

$$\begin{aligned}
e_t_p_{audio_str_premium} &= DIFS + e_idle + phy_mac_hdr + \\
&t_b \times (payload + ip_udp_rtp_hdr) + \\
&delay + SIFS + delay + t_ack \\
&= 0.00005 + 0.0003 + (0.000192 + 272/11000000) + \\
&(8/11000000) \times (160 + 40) + 0.000001 + 0.00001 + \\
&0.000001 + 0.000248 \\
&= 0.000972
\end{aligned}$$

Then, $p\_on_{audio\_str\_premium}$ can be derived based on equation 5.27 as:

$$\begin{aligned}
p_on_{audio_str_premium} &= e_t_p_{audio_str_premium} / in_p_{audio_str} \\
&= 0.000972/0.02 \\
&= 0.048609
\end{aligned}$$

Based on equation 5.21, assuming only the basic audio streaming service users are in the network the value of e_p can be obtained:

$$\begin{aligned}
e_p &= \sum_s p_on_s \times n_s / N \\
&= p_on_{audio_str_premium} \times n_{audio_str_premium} / N \\
&= p_on_{audio_str_premium} \\
&= 0.0457
\end{aligned}$$

e_{ncp} can be calculated by using equation 5.20. Assuming there are 20 premium audio streaming users, the number of user connections in the WLAN is 20 and e_{ncp} can be derived as:

$$\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1 - e_p)^{(N-i)} \\
&= \sum_{i=0}^{20} i \times \binom{20}{i} \times 0.048609^i \times (1 - 0.048609)^{(20-i)} \\
&= 0.972182
\end{aligned}$$

e_{ncp} is smaller than 1, which means, on average, there is less than one packet in contention to access the network channel. Therefore, when there are 20 users, the contention to access the network channel is still low and the quality requirements for the audio streaming service can be satisfied. Assuming the 21st premium audio streaming user is connected to the WLAN, e_{ncp} can be calculated as:

$$\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1 - e_p)^{(N-i)} \\
&= \sum_{i=0}^{21} i \times \binom{21}{i} \times 0.048609^i \times (1 - 0.048609)^{(21-i)} \\
&= 1.020791
\end{aligned}$$

e_{ncp} slightly exceeds 1. The WLAN is estimated as being capable to support 20 premium audio streaming users

6.2.3.4 OPNET Simulation Configurations and Results (Premium Audio Streaming Service)

The OPNET simulation focuses on the downlink traffic between the wired stations and the wireless stations. The wired stations act as the audio packet transmitters and the wireless stations are the receivers. Considering G.726 is used as the audio codec, each wired station will transmit a packet every 20 ms. The size of each packet is 200 bytes, involving the IP/UDP/RTP headers which are 40 bytes. The traffic generation parameters are shown in Table 6.8.

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (60)
OFF State Time (seconds)	constant (0)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.02)
Packet Size (bytes)	constant(200)(*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.8 Traffic Generation Parameters for the Premium Audio Streaming Service

The OPNET simulation results are shown in Figure 6-26 to Figure 6-31. Figures 6-26, 6-27 and 6-28 present the average delay of the packets received by a wireless station, the delay variation, and the packet arrival rate at a wireless station when there are 20 premium audio streaming users and the value of e_{ncp} is 0.972182.

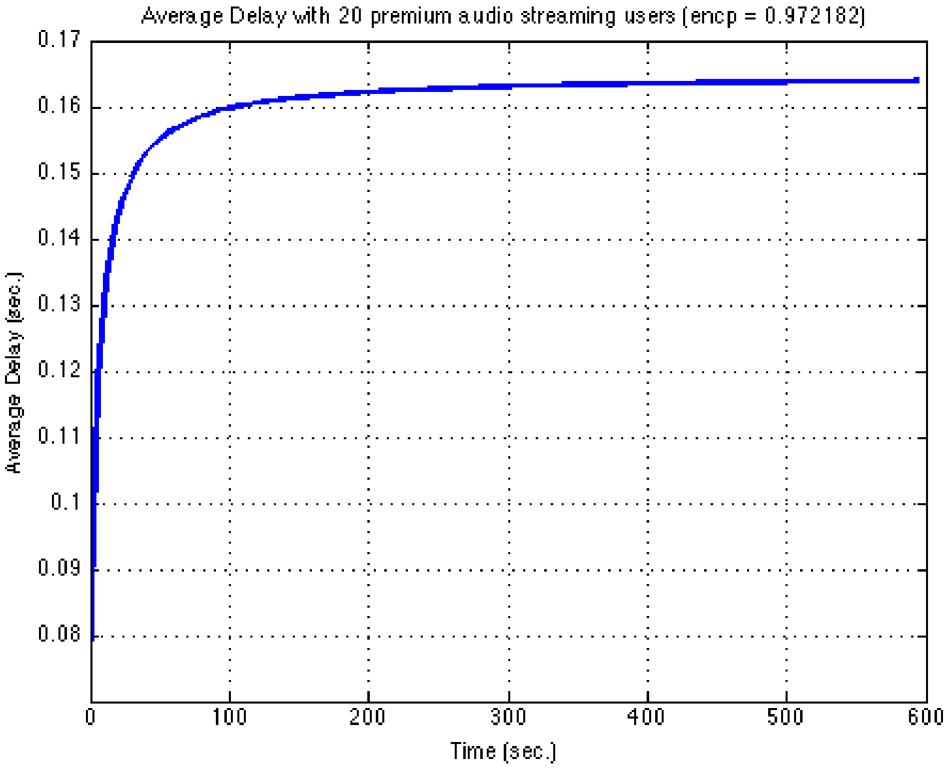


Figure 6-26 Average Delay with 20 premium audio streaming users (e_{ncp} = 0.972182)

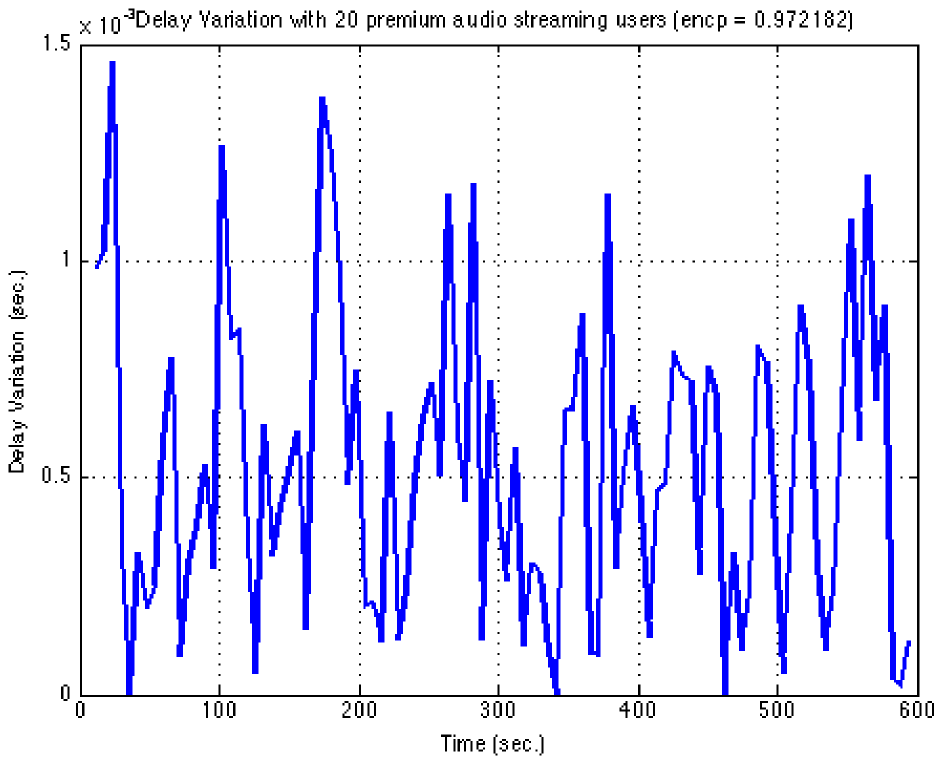


Figure 6-27 Delay Variation with 20 premium audio streaming users (e_{ncp} = 0.972182)

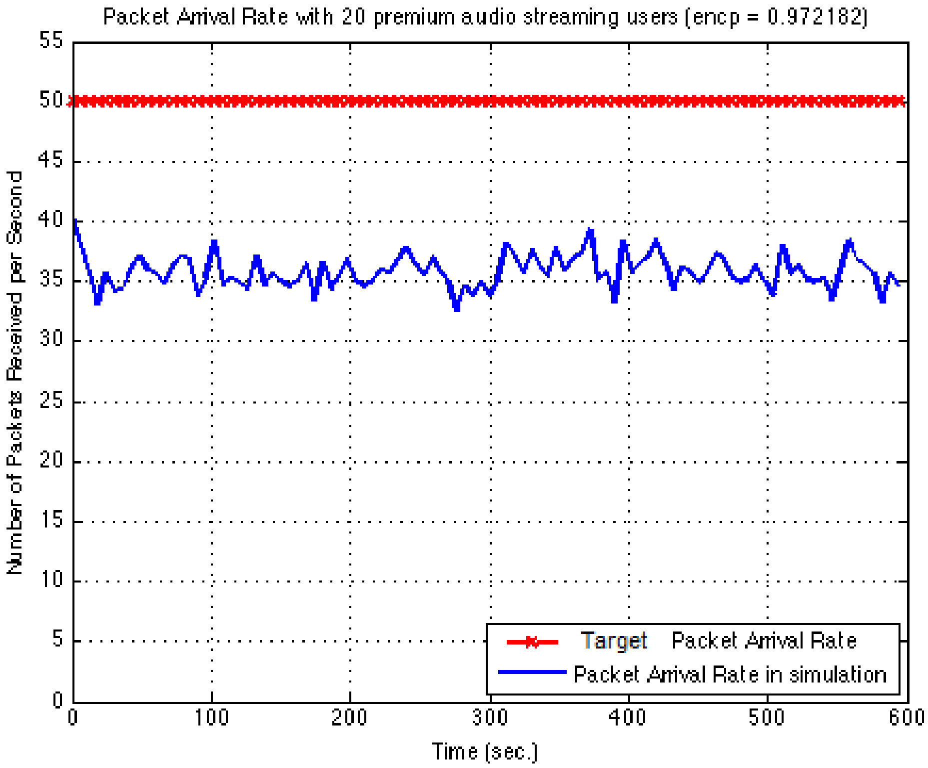


Figure 6-28 Packet Arrival Rate with 20 premium audio streaming users (e_{ncp} = 0.972182)

When there are 20 premium audio streaming users in the WLAN, Figure 6-26 shows that the average packet delay increases and achieves about 165 ms. Figure 6-27 shows that the delay variation is less than 1.5 ms. The simulation results in Figure 6-26 and Figure 6-27 comply with the QoS constraints for streaming class services, which require that the start-up delay should be less than 10 seconds and the delay variation should be less than 2 seconds [Ser08]. However, Figure 6-28 depicts significant fluctuations in the packet arrival rates. Furthermore, the packet arrival rate does not achieve their target values. Based on the codecs implemented for the premium audio streaming service, the required packet arrival rate is 50 packets per second. Because Figure 6-28 shows that the packet arrival rate does not achieve the required values, it indicates that packet losses occur in the network. Figure 6-28 shows that the actual packet arrival rate is about 37 packets per second, which indicates that the packet loss rate is about 26%. [Ser08] requires that the packet loss rate for audio streaming service should be less than 1%. The simulation results demonstrate that the WLAN cannot provide acceptable quality of service when there are 20 premium audio streaming users.

Considering that the WLAN is not able to support 20 premium audio streaming users, the number of users is reduced to be 19 and the corresponding value of e_{ncp} is 0.9235727. The simulation results are presented in Figures 6-29, 6-30 and 6-31.

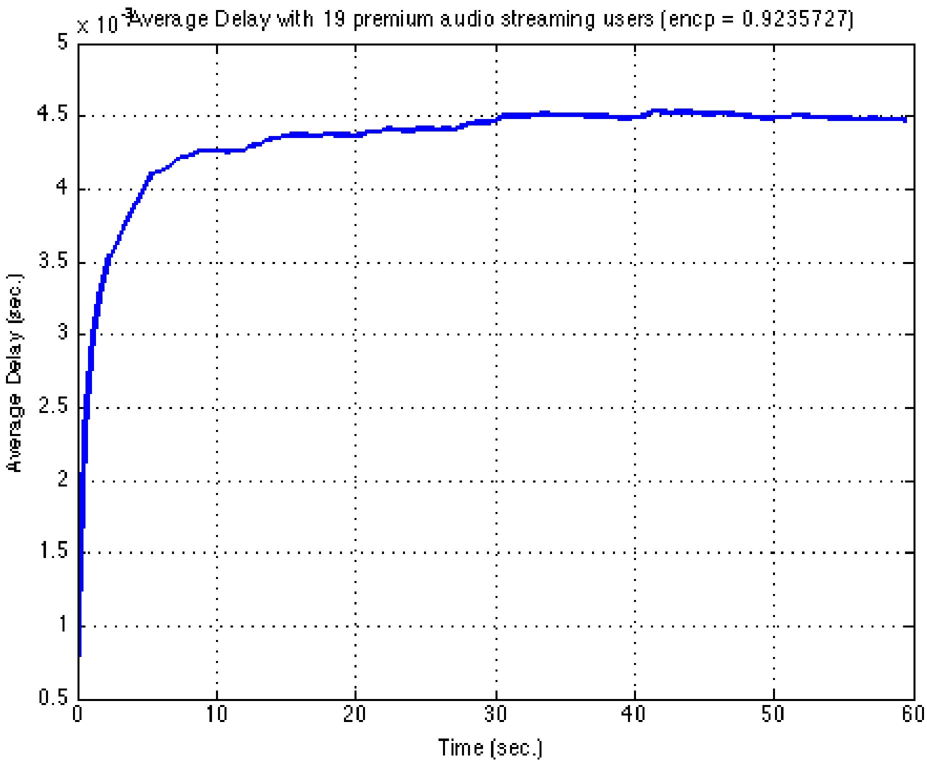


Figure 6-29 Average Delay with 19 premium audio streaming users (e_{ncp} = 0.9235727)

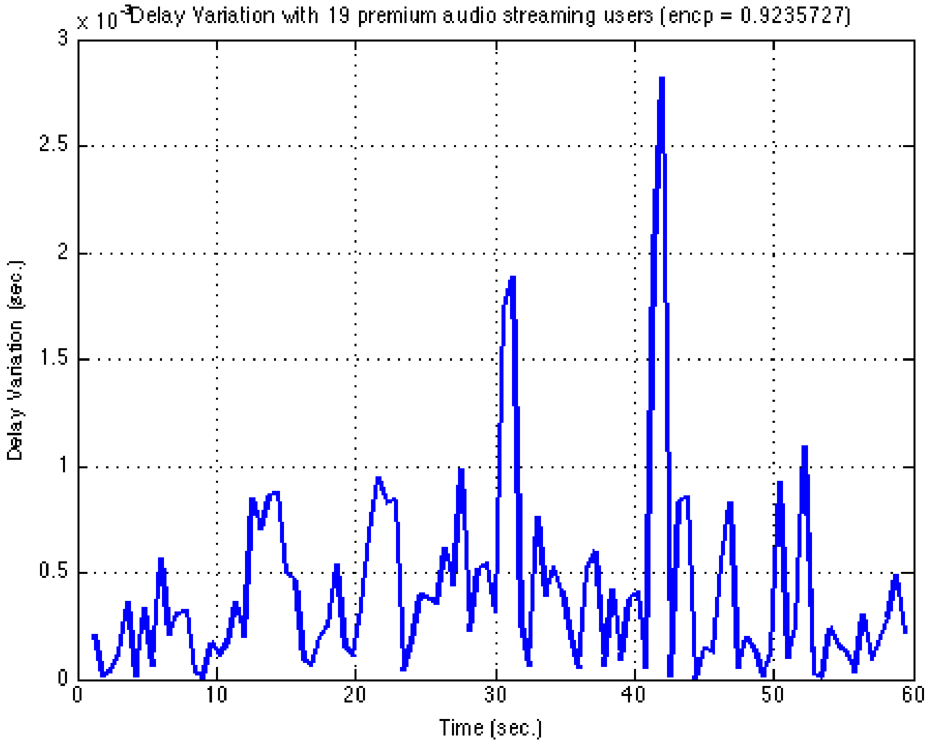


Figure 6-30 Delay Variation with 19 premium audio streaming users (e_{ncp} = 0.9235727)

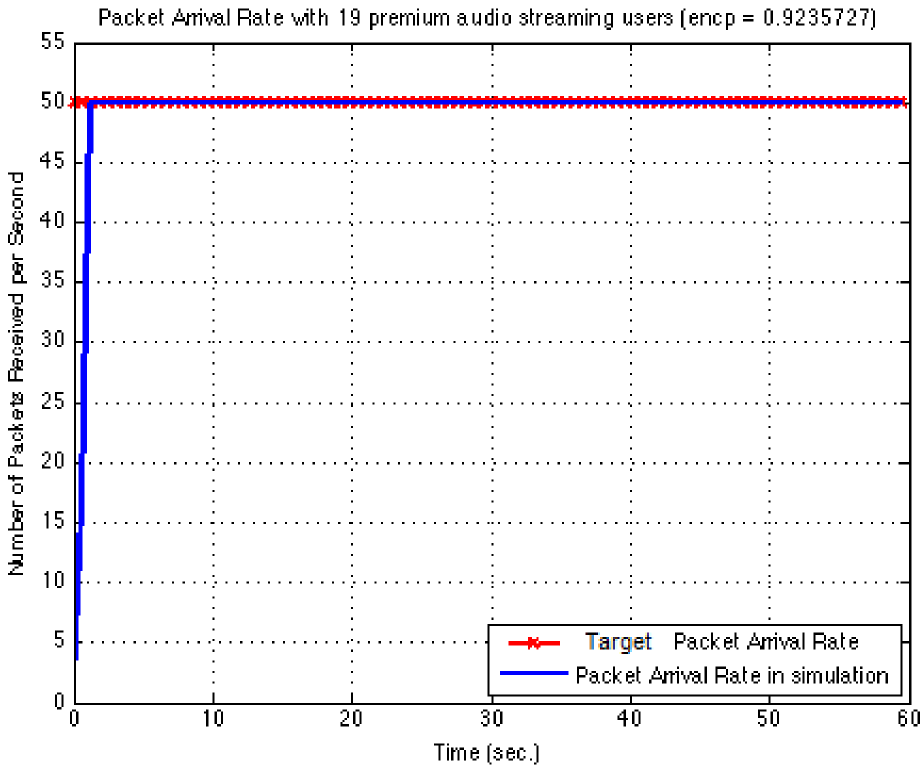


Figure 6-31 Packet Arrival Rate with 19 premium audio streaming users (e_{ncp} = 0.9235727)

When there are 19 premium audio streaming users in the WLAN, Figure 6-29 shows that, as the simulation runs, the average packet delay begins to increase and achieve about 4.5 ms. Figure 6-30 depicts fluctuations in packet delay and the delay variation is less than 3 ms. Both results are under the QoS constraints for streaming class services. Furthermore, Figure 6-31 indicates that the packet arrival rate is stable and the same as the target value, which is 50 packets per second. Therefore, when there are 19 premium audio streaming users, the WLAN is capable of providing acceptable quality of service.

6.2.4 Scenarios of the Video Streaming Service in the IEEE802.11b based WLAN

Similar to the video call service, the video streaming service also comprises two parts: video and audio. Two service classes, basic and premium, are considered in the evaluation model.

6.2.4.1 Evaluation Results of Resource Availability (Basic Video Streaming Service)

For the basic video streaming service, the evaluation model assumes that the video part uses H.264 as the codec. The H.264 codec generates 30 frames per second, which means that the frame inter arrival time is 33 ms. The video data rate is assumed to be 64 kbps, which means that, on average, the size of a frame is 266 bytes. One frame is divided into 4 slices and the size of one slice is 66 bytes. Therefore, $e\_t\_p_{video\_str\_video\_part\_basic}$ can be calculated as:

$$
\begin{aligned}
e\_t\_p_{video\_str\_video\_part\_basic} &= \begin{pmatrix} DIFS + e\_idle + phy\_mac\_hdr + \\ t\_b \times (payload + ip\_udp\_rtp\_hdr) + \\ delay + SIFS + delay + t\_ack \end{pmatrix} \times 4 \\
&= \begin{pmatrix} 0.00005 + 0.0003 + (0.000192 + 272/11000000) + \\ (8/11000000) \times (66 + 40) + 0.000001 + 0.00001 + \\ 0.000001 + 0.000248 \end{pmatrix} \times 4 \\
&= 0.003615
\end{aligned}
$$

Then, $p\_on_{video\_str\_video\_part\_basic}$ can be derived as:

$$\begin{aligned} p\_on_{video\_str\_video\_part\_basic} &= e\_t\_p_{video\_str\_video\_part\_basic} / in\_p_{video\_str\_video\_part} \\ &= 0.003615/0.033 \\ &= 0.109554 \end{aligned}$$

The evaluation model assumes that the audio part of the basic video streaming service employs G.726 as the codec. The G.726 codec generates 50 packets per second and its data rate is 32 kbps. It means that, the packet inter arrival time is 20 ms and the packet payload is 80 bytes. Therefore, $e\_t\_p_{video\_str\_audio\_part}$ can be calculated as:

$$\begin{aligned} e\_t\_p_{video\_str\_audio\_part} &= DIFS + e\_idle + phy\_mac\_hdr + \\ &\quad t\_b \times (payload + ip\_udp\_rtp\_hdr) + \\ &\quad delay + SIFS + delay + t\_ack \\ &= 0.00005 + 0.0003 + (0.000192 + 272/11000000) + \\ &\quad (8/11000000) \times (80 + 40) + 0.000001 + 0.00001 + \\ &\quad 0.000001 + 0.000248 \\ &= 0.000914 \end{aligned}$$

Then, $p\_on_{video\_str\_audio\_part}$ can be derived as:

$$\begin{aligned} p\_on_{video\_str\_audio\_part} &= e\_t\_p_{video\_str\_audio\_part} / in\_p_{video\_str\_audio\_part} \\ &= 0.000914/0.02 \\ &= 0.0457 \end{aligned}$$

Based on equation 5.21, assuming only the basic video streaming service users are in the network, e_p can be calculated as:

$$\begin{aligned} e_p &= \sum_s p\_on_s \times n_s / N \\ &= p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{n_{video\_str\_audio\_part} + n_{video\_str\_video\_part\_basic}} + \\ &\quad p\_on_{video\_str\_video\_part\_basic} \times \frac{n_{video\_str\_video\_part\_basic}}{n_{video\_str\_audio\_part} + n_{video\_str\_video\_part\_basic}} \\ &= \frac{p\_on_{video\_str\_audio\_part} + p\_on_{video\_str\_video\_part\_basic}}{2} \\ &= 0.07762686 \end{aligned}$$

$n_{video\_str\_audio\_part}$ is the number of the connections introduced by the audio part of the video streaming service. $n_{video\_str\_video\_part\_basic}$ is the number of the connections introduced by the video part of the basic video streaming service. Then, e_{ncp} can be calculated by using equation 5.20. Assuming there are 6 basic video streaming users, the number of user connections in the WLAN is 12 and e_{ncp} can be derived as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^{i} \times (1-e_p)^{(N-i)} \\
&= \sum_{i=0}^{12} i \times \binom{12}{i} \times 0.07762686^{i} \times (1-0.07762686)^{(12-i)} \\
&= 0.9315223
\end{aligned}
$$

e_{ncp} is smaller than 1, which means, on average, there is less than one packet in contention to access the network channel. Therefore, when there are 6 users, the contention to access the network channel is still low and the quality requirements for the video streaming service can be satisfied. Assuming the 7th basic video streaming user is connected to the WLAN, e_{ncp} can be derived as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^{i} \times (1-e_p)^{(N-i)} \\
&= \sum_{i=0}^{14} i \times \binom{14}{i} \times 0.07762686^{i} \times (1-0.07762686)^{(14-i)} \\
&= 1.086776
\end{aligned}
$$

e_{ncp} slightly exceeds 1. The WLAN is estimated as being able to support 6 basic video streaming users.

6.2.4.2 OPNET Simulation Configuration and Results (Basic Video Streaming Service)

The OPNET simulation focuses on the downlink traffic between the wired stations and the wireless stations. The video streaming service consists of two parts: audio and video. Therefore, two types of packet transmitters are simulated. The audio packet transmitter uses the G.726 codec and transmits a packet every 20 ms. The size of each packet is 120 bytes, involving the IP/UDP/RTP headers which are 40 bytes. The video packet transmitter uses the H.264 codec and generates a frame every 33 ms. Each frame will be divided into 4 slices. Each slice will be encapsulated into a packet, whose size is 106 bytes, including the IP/UDP/RTP headers. The video packet transmitter has two traffic generation states: ON and OFF. The ON state lasts 13 ms and the OFF state lasts 20 ms. The video packets will only be transmitted during the ON state and the inter transmission time is 3 ms. The traffic generation parameters are shown in Table 6.9 and Table 6.10.

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (60)
OFF State Time (seconds)	constant (0)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.02)
Packet Size (bytes)	constant(120)(*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.9 Audio Traffic Generation Parameters for the Basic Video Streaming Service

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (0.013)
OFF State Time (seconds)	constant (0.02)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.003)
Packet Size (bytes)	constant(106)(*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.10 Video Traffic Generation Parameters for the Basic Video Streaming Service

The OPNET simulation results are shown in Figure 6-32 to Figure 6-37. Figures 6-32 and 6-33 present the average delay of the packets received by a wireless station and the delay variation when there are 6 basic video streaming users and the value of e_{ncp} is 0.9315223.

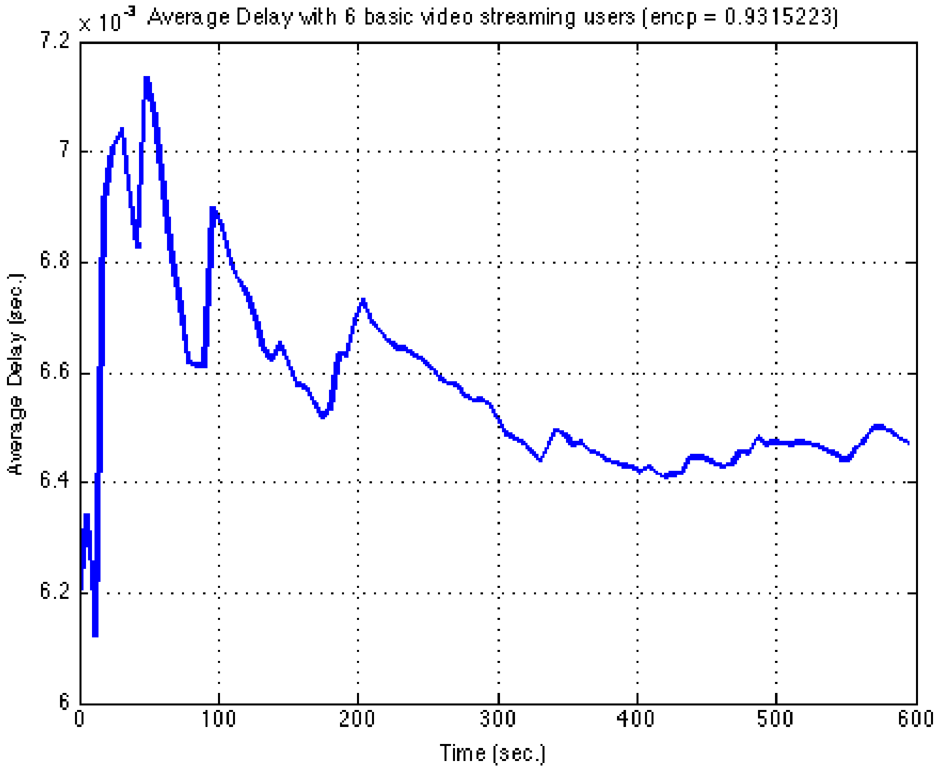


Figure 6-32 Average Delay with 6 basic video streaming users (e_{ncp} = 0.9315223)

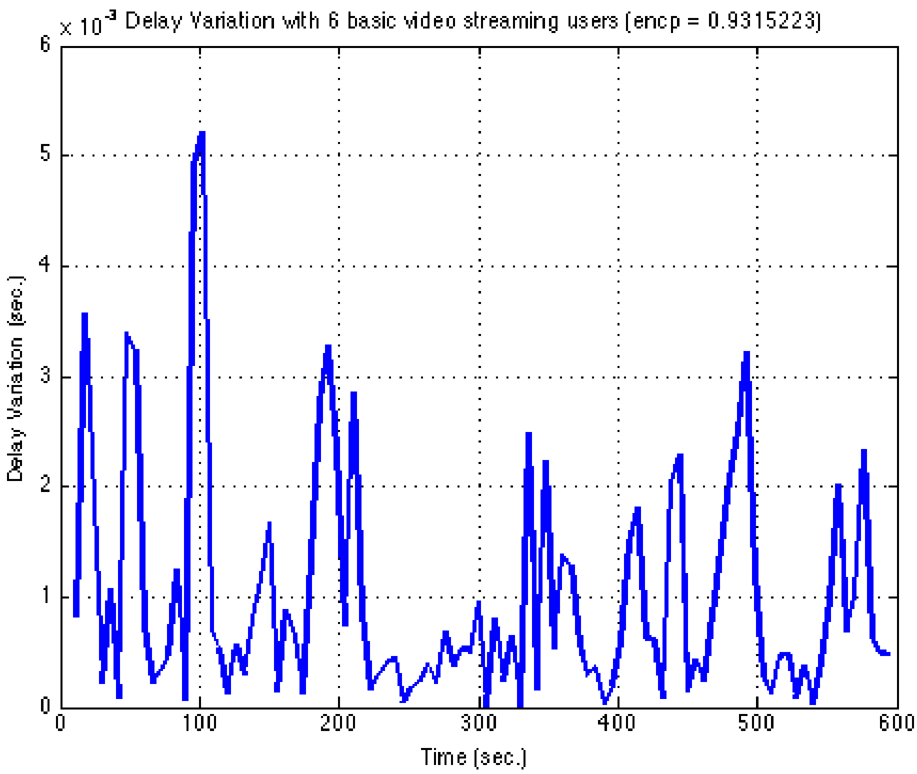


Figure 6-33 Delay Variation with 6 basic video streaming users (e_{ncp} = 0.9315223)

When there are 6 basic video streaming users in the WLAN, Figure 6-32 shows that, after the fluctuations in the early stage of the simulation, the average packet delay begins to stabilise and achieve around 6.5 ms. Figure 6-41 depicts the fluctuations in packet delay and the delay variation is less than 5.3 ms. These values comply with the QoS constraints for streaming class services, which require the start-up delay should be less than 10 seconds and the delay variation should be less than 2 seconds [Ser08].

Figures 6-34, 6-35, 6-36 and 6-37 present the average delay of the packets received by a wireless station, the delay variation, and the audio and video packet arrival rates at a wireless station when there are 7 basic video streaming users and the value of e_{ncp} is 1.086776.

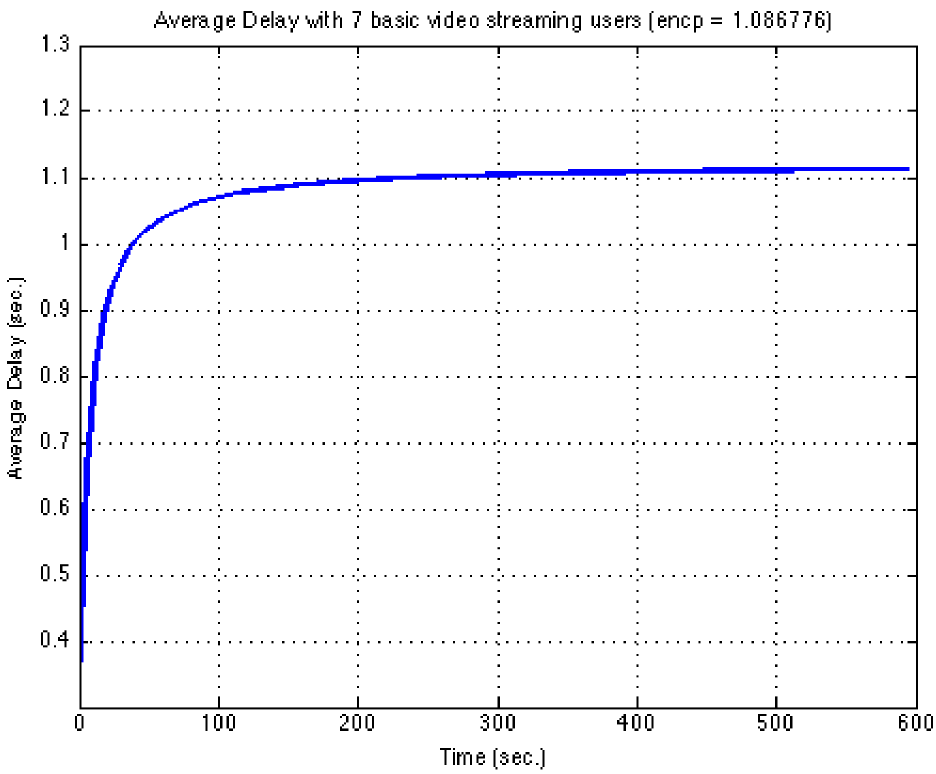


Figure 6-34 Average Delay with 7 basic video streaming users (e_{ncp} = 1.086776)

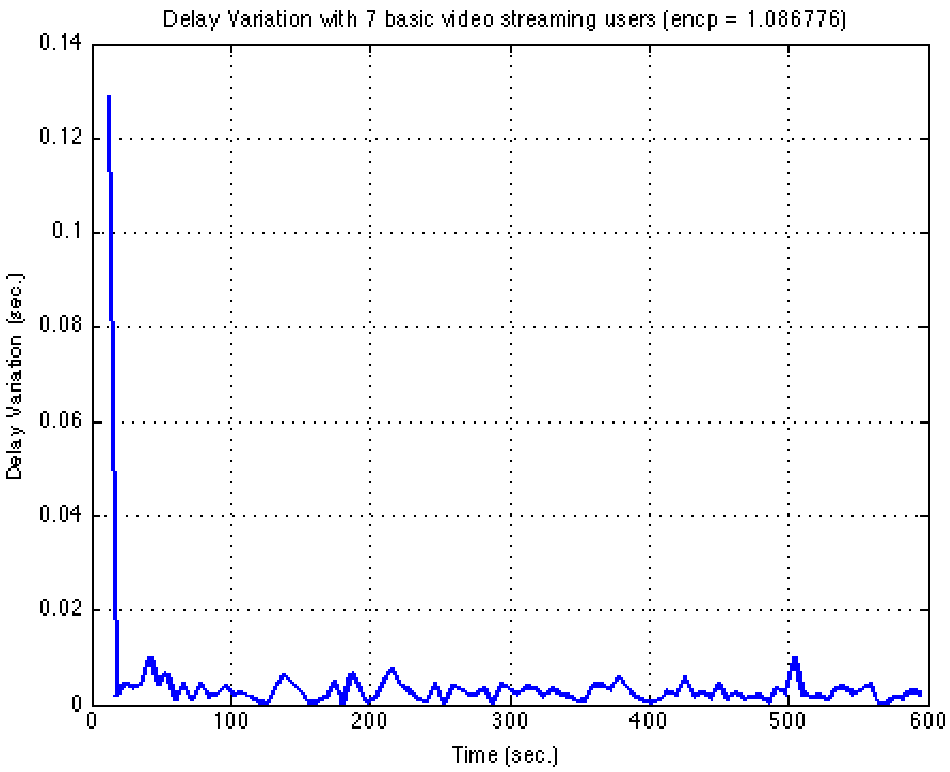


Figure 6-35 Delay Variation with 7 basic video streaming users (e_{ncp} = 1.086776)

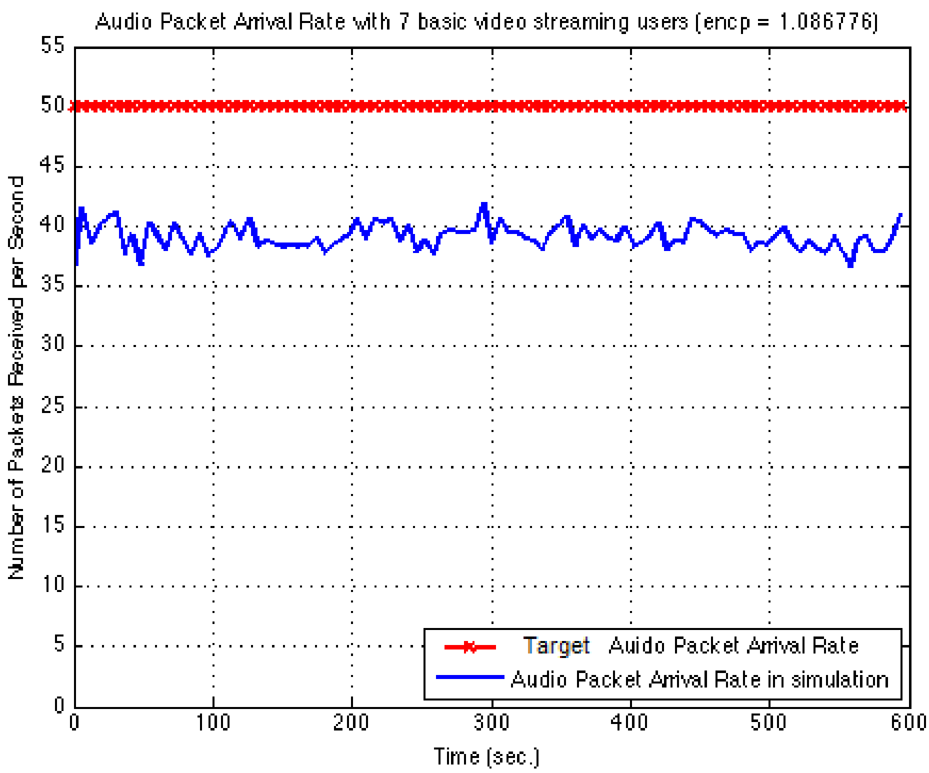


Figure 6-36 Audio Packet Arrival Rate with 7 basic video streaming users (e_{ncp} = 1.086776)

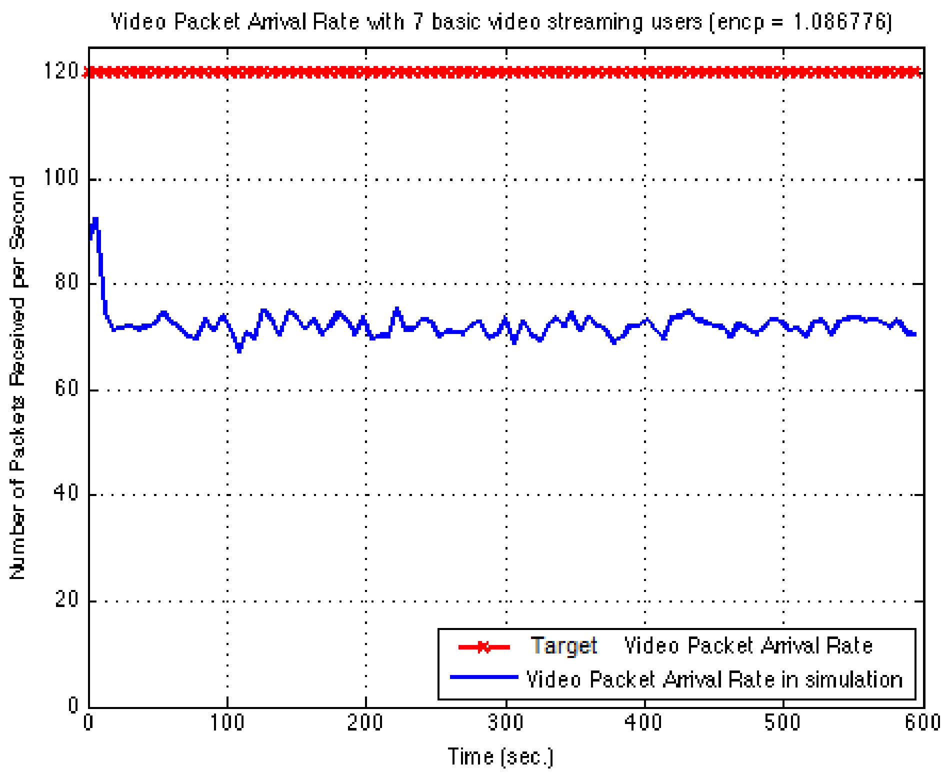


Figure 6-37 Video Packet Arrival Rate with 7 basic video streaming users (e_{ncp} = 1.086776)

When there are 7 basic video streaming users in the WLAN, Figure 6-34 shows that the average packet delay increases dramatically as simulation runs and achieves about 1.1 seconds. This value is acceptable as [Ser08] requires that the start-up delay of the streaming class services should be less than 10 seconds. Figure 6-35 shows that the delay variation is less than 10 ms, which is under the constraint of 2 seconds [Ser08]. However, Figure 6-36 and Figure 6-37 depict significant fluctuations in the audio and video packet arrival rates. Furthermore, both of the audio and video packet arrival rates do not achieve the target values. Based on the codecs implemented for the basic video streaming service, the required audio packet arrival rate is 50 packets per second and the required video frame arrival rate is 30 frames per second. Each frame is divided into 4 slices and each slice is encapsulated into a packet. Therefore, the required video packet arrival rate is 120 packets per second. Because Figures 6-36 and 6-37 show that the packet arrival rates do not achieve the required values, it indicates that packet losses occur in the network. Figure 6-36 shows that the actual audio packet arrival rate is about 40 packets per second, which indicates that the audio packet loss rate is about 20%. Figure 6-37 shows that the actual video packet arrival rate is about 75 packets per second, which indicates that the video packet loss rate is about 37.5%. However, [Ser08] requires that the audio packet loss rate should be less than 1% and the video packet loss rate should be less than 2%. Therefore, when there are 7 basic video streaming users, the WLAN cannot provide acceptable quality of service and should not admit more than 6 basic video streaming users.

6.2.4.3 Evaluation Results of Resource Availability (Premium Video Streaming Service)

For the premium video streaming service, the evaluation model assumes that the video part uses H.264 as the codec. The H.264 codec generates 30 frames per second, which means that the frame inter arrival time is 33 ms. The video data rate is assumed as 128 kbps. On average, the size of a frame is 533 bytes. One frame is divided into 8 slices and the size of one slice is 67 bytes. Therefore, $e\_t\_p_{video\_str\_video\_part\_premium}$ can be calculated as:

$$
\begin{aligned}
e\_t\_p_{video\_str\_video\_part\_premium} &= \begin{pmatrix} DIFS + e\_idle + phy\_mac\_hdr + \\ t\_b \times (payload + ip\_udp\_rtp\_hdr) + \\ delay + SIFS + delay + t\_ack \end{pmatrix} \times 8 \\
&= \begin{pmatrix} 0.00005 + 0.0003 + (0.000192 + 272/11000000) + \\ (8/11000000) \times (67 + 40) + 0.000001 + 0.00001 + \\ 0.000001 + 0.000248 \end{pmatrix} \times 8 \\
&= 0.007236
\end{aligned}
$$

Then, $p\_on_{video\_str\_video\_part\_premium}$ can be derived as:

$$
\begin{aligned}
p\_on_{video\_str\_video\_part\_premium} &= e\_t\_p_{video\_str\_video\_part\_premium} / in\_p_{video\_str\_video\_part} \\
&= 0.007236/0.033 \\
&= 0.219284
\end{aligned}
$$

The evaluation model assumes that the audio part employs G.726 as the codec. The G.726 codec generates 50 packets per second and its data rate is 32 kbps. Based on the value of $p\_on_{video\_str\_audio\_part}$ calculated in subsection 6.2.4.1 and equation 5.21, assuming only the premium video streaming service users are in the network, e_p can be calculated as:

$$
\begin{aligned}
e_p &= \sum_s p\_on_s \times n_s / N \\
&= p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{n_{video\_str\_audio\_part} + n_{video\_str\_video\_part}} + \\
&\qquad p\_on_{video\_str\_video\_part\_premium} \times \frac{n_{video\_str\_video\_part\_premium}}{n_{video\_str\_audio\_part} + n_{video\_str\_video\_part\_premium}} \\
&= \frac{p\_on_{video\_str\_audio\_part} + p\_on_{video\_str\_video\_part\_premium}}{2} \\
&= 0.132492
\end{aligned}
$$

$n_{video\_str\_audio\_part}$ is the number of the connections introduced by the audio part of the premium video streaming service. $n_{video\_str\_video\_part\_premium}$ is the number of the connections introduced by the video part of the premium video streaming service. Assuming there are 3 premium video streaming users, the number of user connections in the WLAN is 6 and e_{ncp} can be derived as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1 - e_p)^{(N-i)} \\
&= \sum_{i=0}^{6} i \times \binom{6}{i} \times 0.132492^i \times (1 - 0.132492)^{(6-i)} \\
&= 0.794951
\end{aligned}
$$

e_{ncp} is smaller than 1, which means, on average, there is less than one packet in contention to access the network channel. Therefore, when there are 3 users, the contention to access the network channel is still low and the quality requirements for the video streaming service can be satisfied. Assuming the 4th premium video streaming user is connected to the WLAN, e_{ncp} can be derived as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times (1 - e_p)^{(N-i)} \\
&= \sum_{i=0}^{8} i \times \binom{8}{i} \times 0.132492^i \times (1 - 0.132492)^{(8-i)} \\
&= 1.059935
\end{aligned}
$$

The value of e_{ncp} exceeds 1. The WLAN is estimated as being able to support 3 premium video streaming users.

6.2.4.4 OPNET Simulation Configurations and Results (Premium Video Streaming Service)

The OPNET simulation focuses on the downlink traffic between the wired stations and the wireless stations. The video streaming service consists of two parts: audio and video. Therefore, two types of packet transmitters are simulated. The audio packet transmitter uses the G.726 codec and transmits a packet every 20 ms. The size of each packet is 120 bytes, involving the IP/UDP/RTP headers which are 40 bytes. The video packet transmitter uses the H.264 codec and generates a frame every 33 ms. Each frame will be divided into 8 slices. Each slice will be encapsulated into a packet, whose size is 107 bytes, including the IP/UDP/RTP headers. The video packet transmitter has two traffic generation states: ON and OFF. The ON state lasts 25 ms and the OFF state lasts 8 ms. The video packets will only be transmitted during the ON state and the inter transmission time is 3 ms. The traffic generation parameters are shown in Table 6.11 and Table 6.12.

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (60)
OFF State Time (seconds)	constant (0)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.02)
Packet Size (bytes)	constant(120)(*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.11 Audio Traffic Generation Parameters for the Premium Video Streaming Service

Traffic Generation Parameters	
Start Time (seconds)	uniform (0.1,1.1)
ON State Time (seconds)	constant (0.025)
OFF State Time (seconds)	constant (0.008)
Packet Generation Arguments	
Interarrival Time (seconds)	constant (0.003)
Packet Size (bytes)	constant(107)(*Including IP/UDP/RTP header*)
Segmentation Size (bytes)	No Segmentation
Stop Time (seconds)	Never

Table 6.12 Video Traffic Generation Parameters for the Premium Video Streaming Service

The OPNET simulation results are shown in Figure 6-38 to Figure 6-43. Figures 6-38 and 6-39 present the average delay of the packets received by a wireless station and the delay variation when there are 3 premium video streaming users and the value of e_{ncp} is 0.794951.

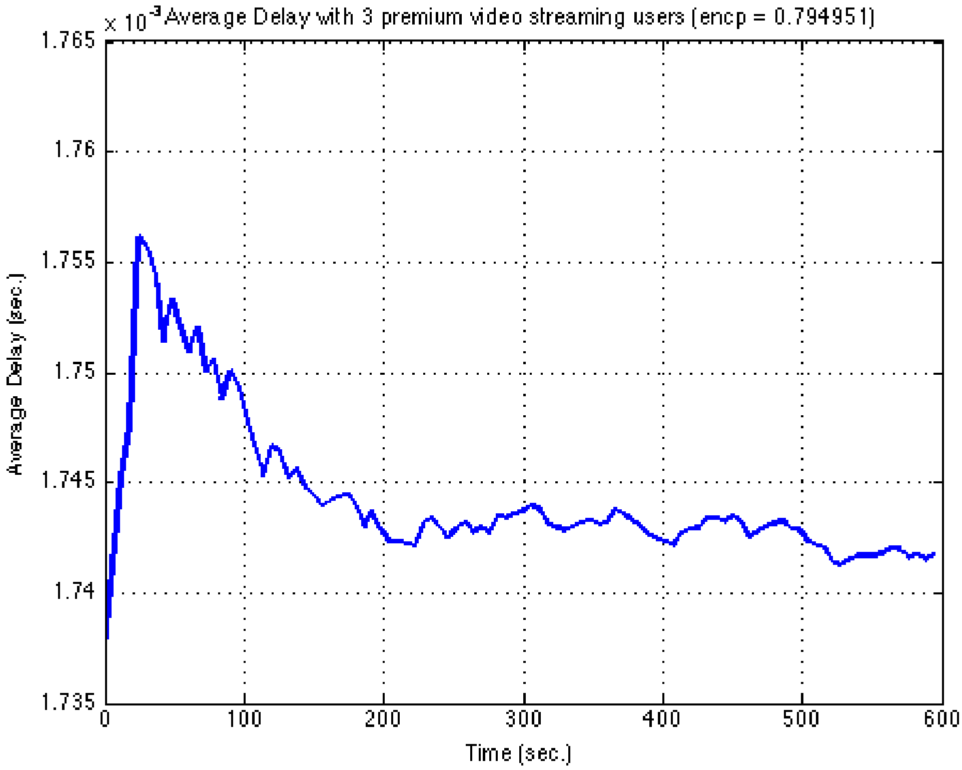


Figure 6-38 Average Delay with 3 premium video streaming users (e_{ncp} = 0.794951)

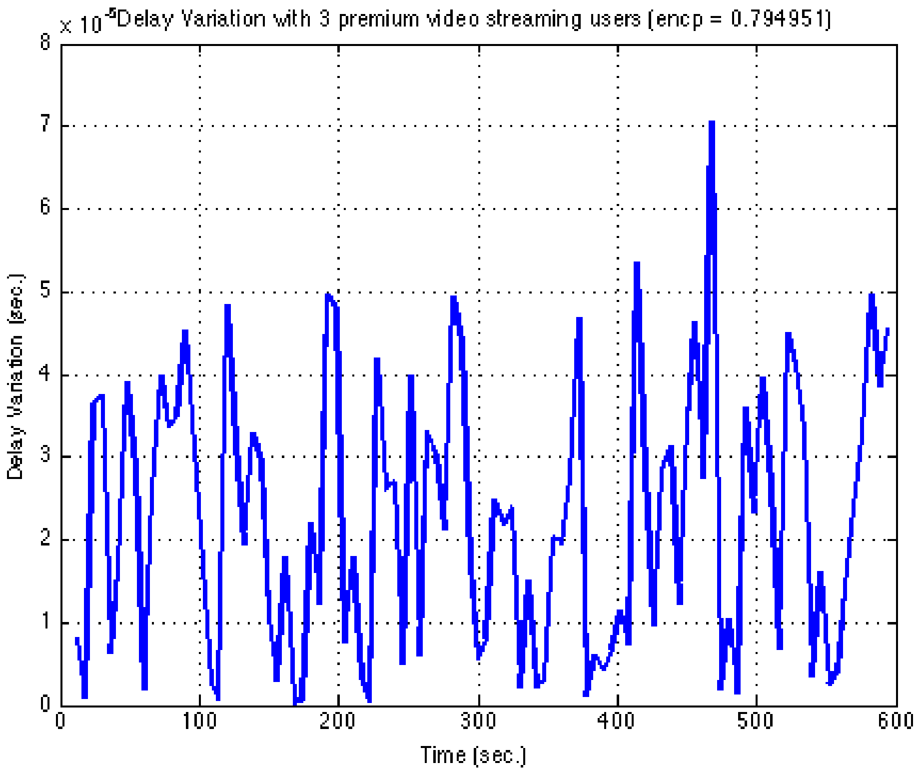


Figure 6-39 Delay Variation with 3 premium video streaming users (e_{ncp} = 0.794951)

When there are 3 premium video streaming users in the WLAN, Figure 6-38 shows that, after the fluctuations in the early stage of the simulation, the average packet delay begins to stabilise and achieve around 1.742 ms. Figure 6-39 depicts the fluctuations in packet delay and the delay variation is less than 0.07 ms. The above simulation results comply with the QoS constraint for streaming class services (start-up delay should be less than 10 seconds and delay variation should be less than 2 seconds [Ser08]).

Figures 6-40, 6-41, 6-42 and 6-43 present the average delay of the packets received by a wireless station, the delay variation, and the audio and video packet arrival rates at a wireless station when there are 4 premium video streaming users and the value of e_{ncp} is 1.059935.

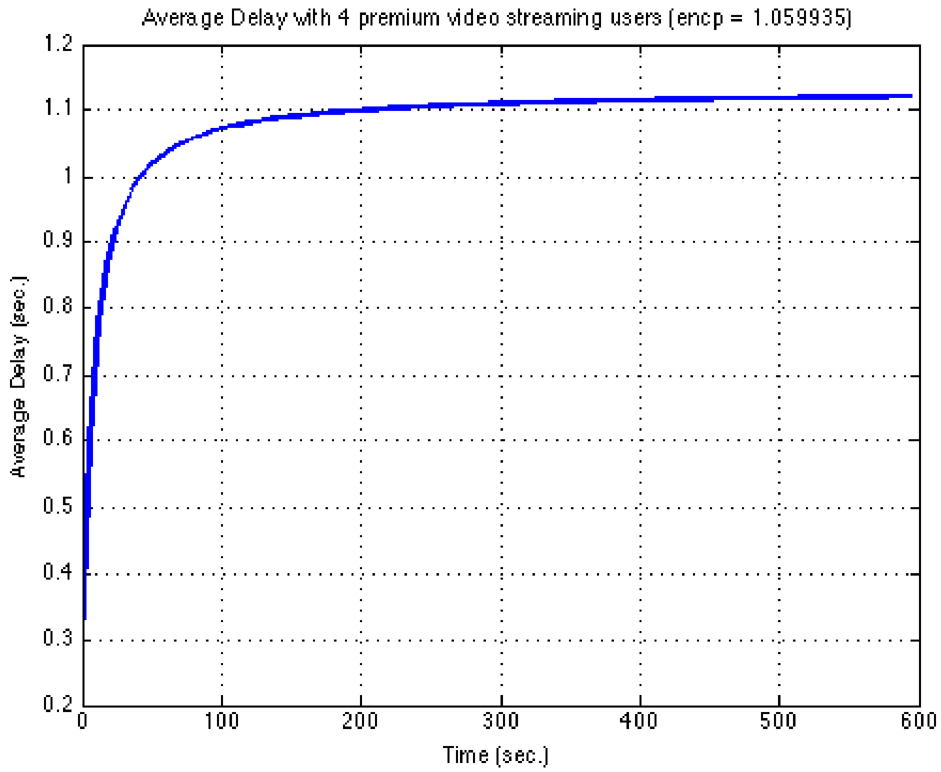


Figure 6-40 Average Delay with 4 premium video streaming users (e_{ncp} = 1.059935)

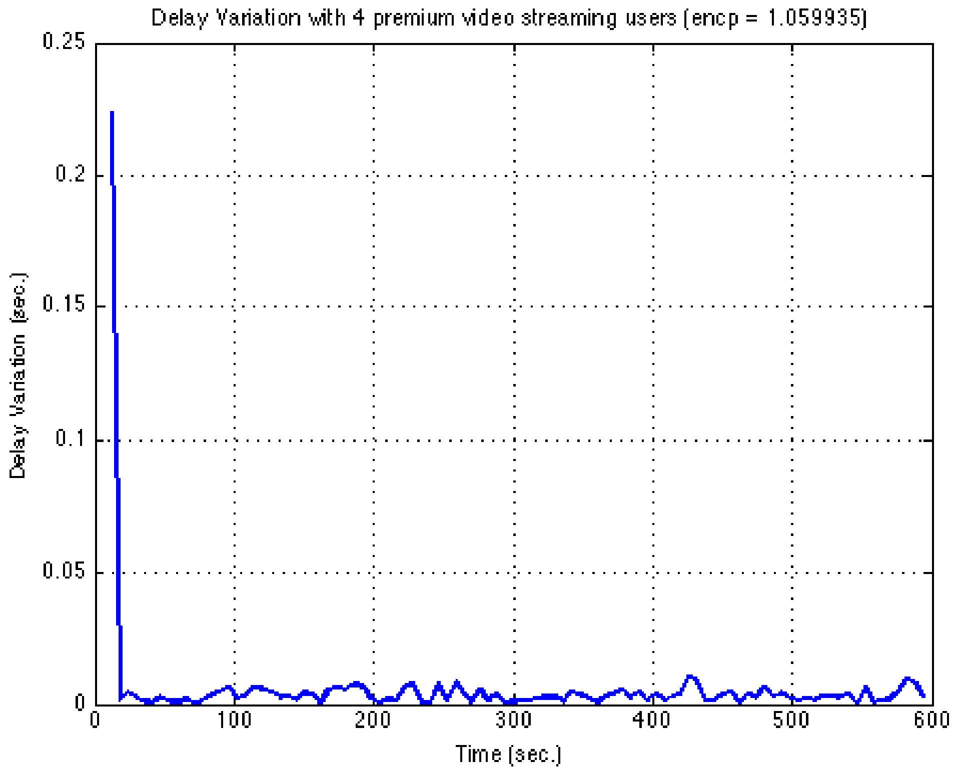


Figure 6-41 Delay Variation with 4 premium video streaming users (e_{ncp} = 1.059935)

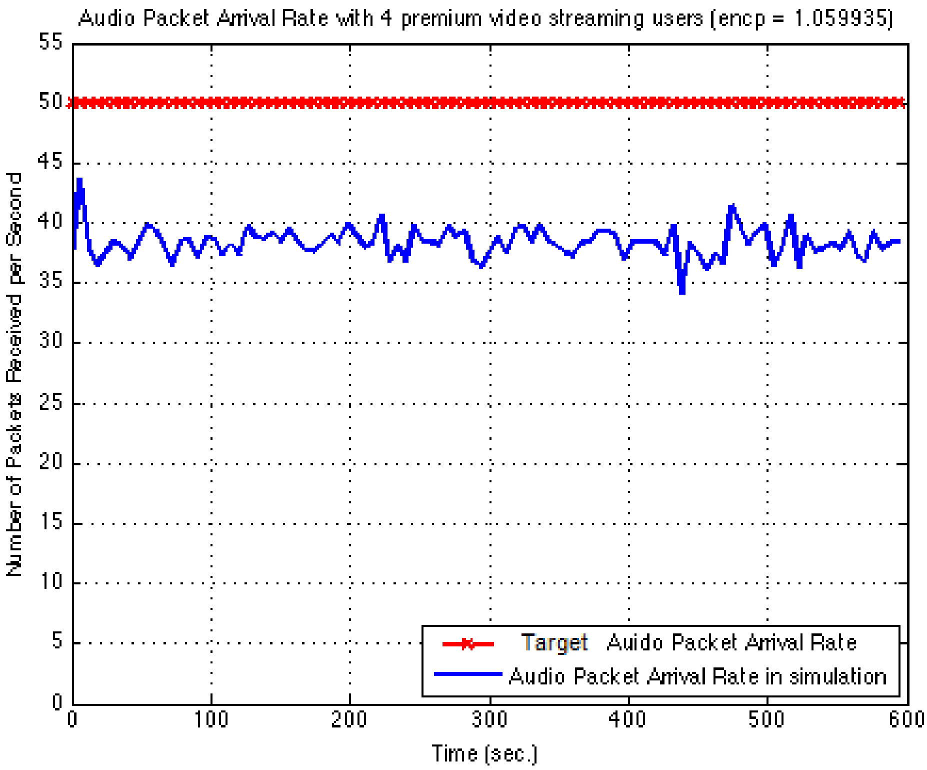


Figure 6-42 Audio Packet Arrival Rate with 4 premium video streaming users (e_{ncp} = 1.059935)

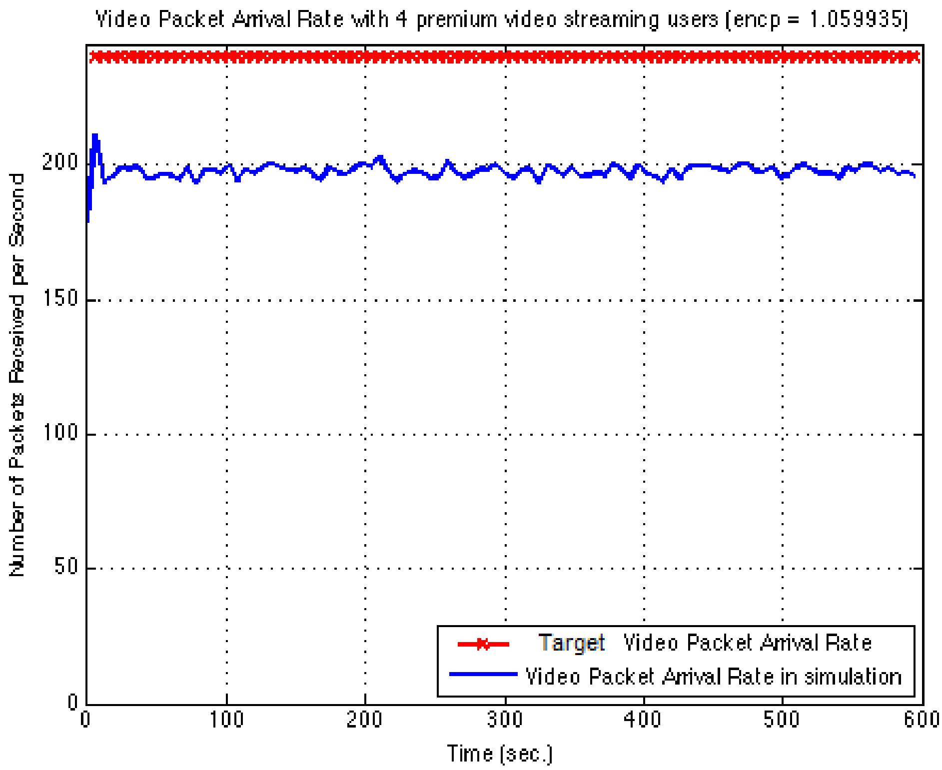


Figure 6-43 Video Packet Arrival Rate with 4 premium video streaming users (e_{ncp} = 1.059935)

When there are 4 premium video streaming users in the WLAN, Figure 6-40 shows that the average packet delay increases dramatically as simulation runs and achieves about 1.1 seconds. Figure 6-41 shows that the delay variation is less than 10 ms. Both values are under the QoS constraints for streaming class services, whose start-up delay should be less than 10 seconds and the delay variation should be less than 2 seconds [Ser08]. However, Figure 6-42 and Figure 6-43 depict significant fluctuations in the audio and video packet arrival rates. Furthermore, both of the audio and video packet arrival rates do not achieve the target values. Based on the codecs implemented for the basic video streaming service, the required audio packet arrival rate is 50 packets per second and the required video frame arrival rate is 30 frames per second. Each frame is divided into 8 slices and each slice is encapsulated into a packet. Therefore, the required video packet arrival rate is 240 packets per second. Because Figures 6-42 and 6-43 show that the packet arrival rates do not achieve the required values, it indicates that packet losses occur in the network. Figure 6-42 shows that the actual audio packet arrival rate is about 38 packets per second, which indicates that the audio packet loss rate is about 24%. Figure 6-43 shows that the actual video packet arrival rate is about 200 packets per second, which indicates that the video packet loss rate is about 17%. However, [Ser08] requires that the audio packet loss rate should be less than 1% and the video packet loss rate should be less than 2%. Therefore, when there are 4 premium video streaming users, the WLAN cannot provide acceptable quality of service. The WLAN should not admit more than 3 premium video streaming users.

6.2.5 Scenarios of Hybrid Service Types in the IEEE802.11b based WLAN

In this subsection, a scenario considering a variety of real-time services is studied. The services include Video Call with premium service class, VoIP with basic and premium service classes, Audio Streaming with basic and premium service classes, and Video Streaming with basic and premium service classes. Based on the calculation presented in the above subsection, the probability that the channel is occupied by the transmission of the packets belonging to the each type of service is listed in Table 6.13.

Service Type	*p_on*
VoIP with basic service class	$p\_on_{VoIP\_basic} = 0.029$
VoIP with premium service class	$p\_on_{VoIP\_premium} = 0.044$
Video Call with premium service class: video part	$p\_on_{video\_call\_video\_part\_premium} = 0.0801$
Video Call: audio part	$p\_on_{video\_call\_audio\_part} = 0.0291$
Audio Streaming with basic service class	$p\_on_{audio\_str\_basic} = 0.0457$
Audio Streaming with premium service class	$p\_on_{audio\_str\_premium} = 0.048609$
Video Streaming with basic service class: video part	$p\_on_{video\_str\_video\_part\_basic} = 0.109554$
Video Streaming with premium service class: video part	$p\_on_{video\_str\_video\_part\_premium} = 0.219284$
Video Streaming: audio part	$p\_on_{video\_str\_audio\_part} = 0.0457$

Table 6.13 Channel Occupation Probability of Each Type of Service

6.2.5.1 Scenario 1

Assuming that the number of each service type presented above is 1, these services will introduce 14 connections to the WLAN, including: 2 connections from the basic VoIP service, 2 connections from the premium VoIP service, 4 connections from the premium video call service, 1 connection from the basic audio streaming service, 1 connection from the premium audio streaming service, 2 connections from the basic video streaming service, and 2 connections from the premium video streaming service. Therefore, based on equation 5.21 and the values of p_on_{st} of each type of service, the expected probability that the channel is occupied by a packet transmission (e_p) can be calculated as:

$$
\begin{aligned}
e_p &= \sum_s \left(p\_on_s \times n_s / N \right) \\
&= p\_on_{VoIP\_basic} \times \frac{n_{VoIP\_basic}}{N} + p\_on_{VoIP\_premium} \times \frac{n_{VoIP\_premium}}{N} + \\
&\quad p\_on_{video\_call\_video\_part\_premium} \times \frac{n_{video\_call\_video\_part\_premium}}{N} + p\_on_{video\_call\_audio\_part} \times \frac{n_{video\_call\_audio\_p}}{N} \\
&\quad p\_on_{audio\_str\_premium} \times \frac{n_{audio\_str\_premium}}{N} + p\_on_{audio\_str\_basic} \times \frac{n_{audio\_str\_basic}}{N} + \\
&\quad p\_on_{video\_str\_video\_part\_premium} \times \frac{n_{video\_str\_video\_part\_premium}}{N} + p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{N} \\
&\quad p\_on_{video\_str\_video\_part\_basic} \times \frac{n_{video\_str\_video\_part\_basic}}{N} + p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{N} + \\
&= 0.044 \times \frac{2}{14} + 0.029 \times \frac{2}{14} + 0.0801 \times \frac{2}{14} + 0.0291 \times \frac{2}{14} + 0.048609 \times \frac{1}{14} + 0.0457 \times \frac{1}{14} + \\
&\quad 0.219284 \times \frac{1}{14} + 0.0457 \times \frac{1}{14} + 0.109554 \times \frac{1}{14} + 0.0457 \times \frac{1}{14} \\
&= 0.06279774
\end{aligned}
$$

The value of e_{ncp} can be calculated by using equation 5.20 as follows:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times \left(1 - e_p\right)^{(N-i)} \\
&= \sum_{i=0}^{14} i \times \binom{14}{i} \times 0.06279774^i \times \left(1 - 0.06279774\right)^{(12-i)} = 0.8791684
\end{aligned}
$$

The value of e_{ncp} is smaller than 1. That means, on average, there is less than one packet in contention to access the network channel. The WLAN is estimated to be able to support these services.

6.2.5.2 OPNET Simulation Configurations and Results (Scenario 1)

In the simulations, the traffic generation parameters are based on Table 6.1 to Table 6.12. The OPNET simulation results are shown in Figure 6-44 to Figure 6-46, which present the average delay, the delay variation, and the packet arrival rate of the premium VoIP service in scenario 1.

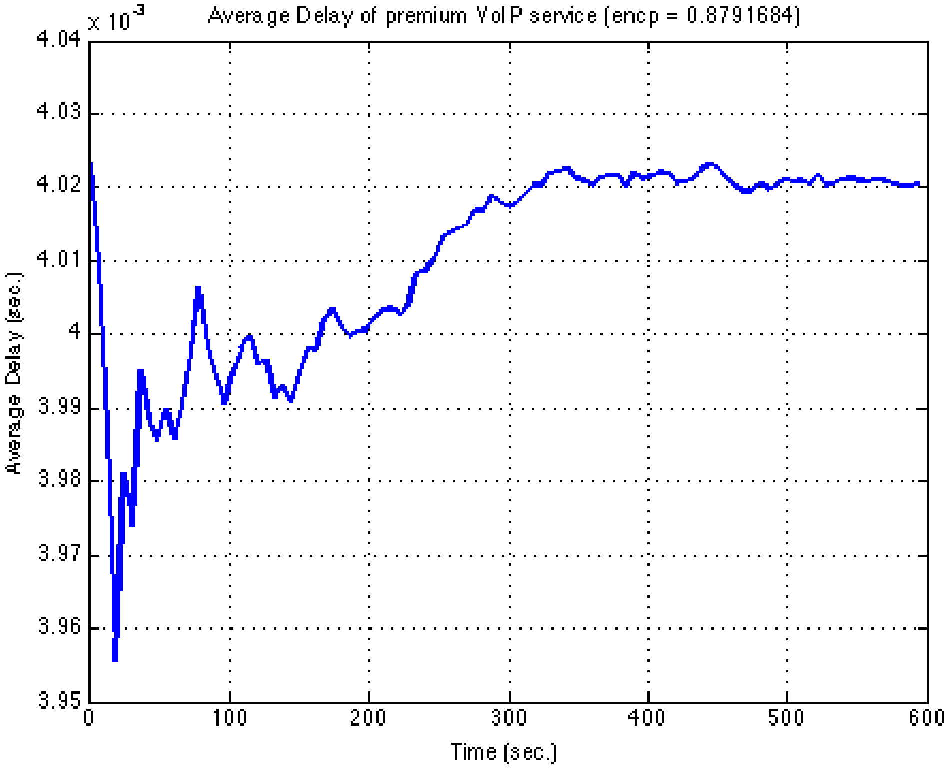


Figure 6-44 Average Delay of premium VoIP service (e_{ncp} = 0.8791684)

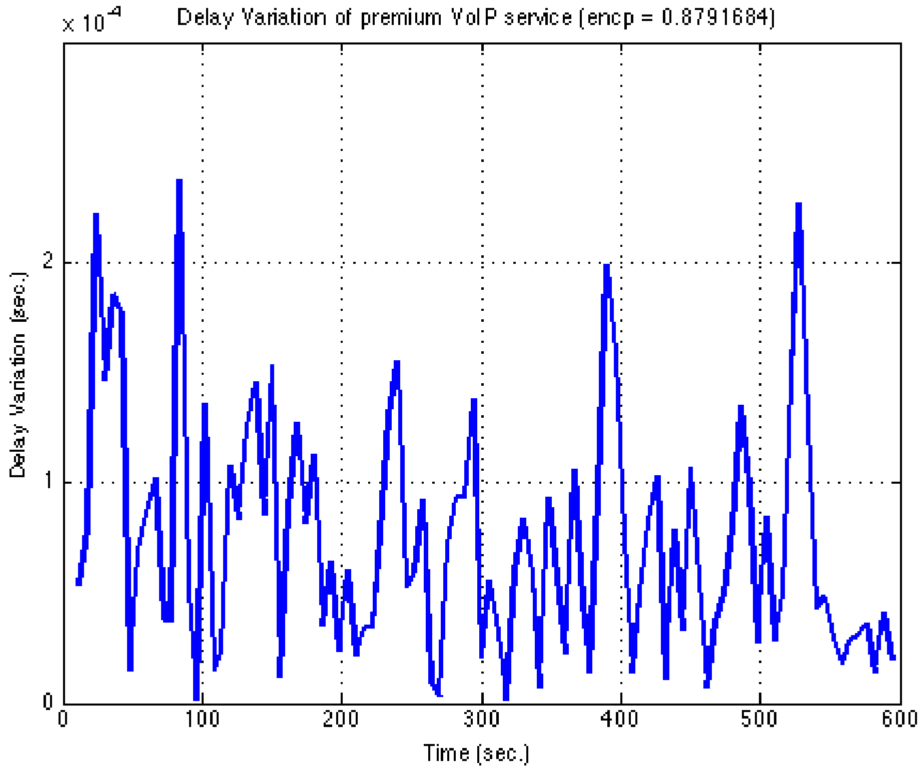


Figure 6-45 Delay Variation of premium VoIP service (e_{ncp} = 0.8791684)

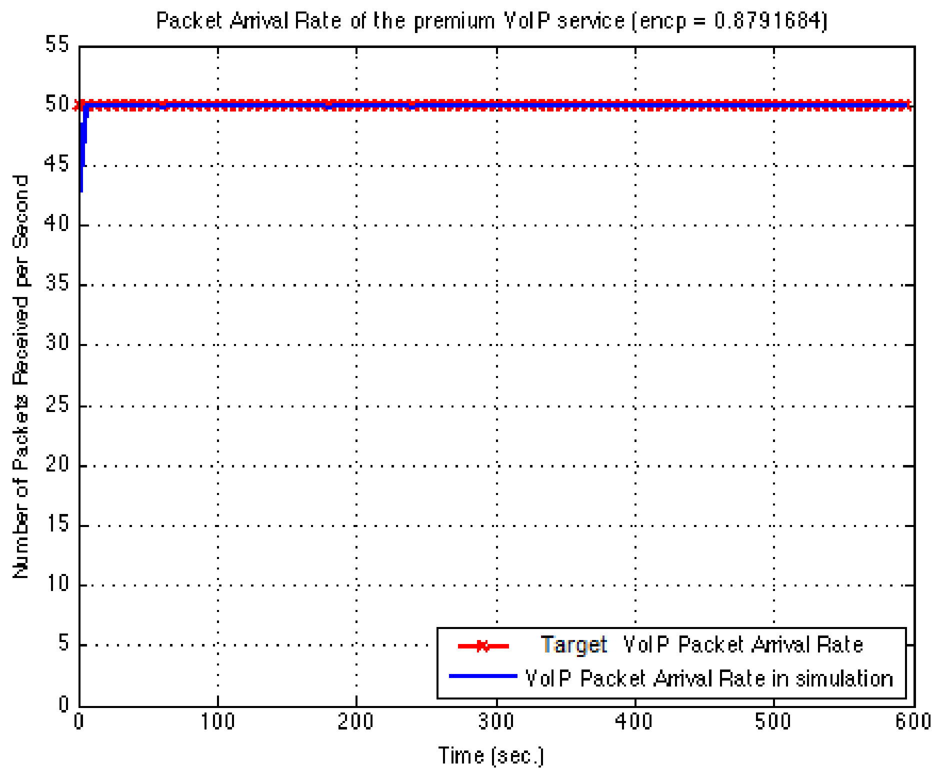


Figure 6-46 Packet Arrival Rate of premium VoIP service (e_{ncp} = 0.8791684)

Figure 6-44 shows that, after the fluctuations in the early stage of the simulation, the average packet delay of the premium VoIP service stabilises at around 4.02 ms. Figure 6-45 depicts fluctuations in packet delay and the delay variation is less than 0.25 ms. Both of these values comply with the QoS requirements for conversational class services [Ser08]. Furthermore, Figure 6-46 indicates that the packet arrival rate in the simulation is the same as the target value, which is 50 packets per second. These results show that, with the value of $\boldsymbol{e_{ncp}}$ less than 1, the WLAN network is capable of providing acceptable quality of service.

6.2.5.3 Scenario 2

Assuming that one more premium video call service request arrives at the WLAN, the number of connections would increase to 18 and the value of e_p can be calculated as:

$$
\begin{aligned}
e_p &= \sum_s \left(p\_on_s \times n_s / N\right) \\
&= p\_on_{VoIP\_basic} \times \frac{n_{VoIP\_basic}}{N} + p\_on_{VoIP\_premium} \times \frac{n_{VoIP\_premium}}{N} + \\
&\quad p\_on_{video\_call\_video\_part} \times \frac{n_{video\_call\_video\_part}}{N} + p\_on_{video\_call\_audio\_part} \times \frac{n_{video\_call\_audio\_part}}{N} + \\
&\quad p\_on_{audio\_str\_premium} \times \frac{n_{audio\_str\_premium}}{N} + p\_on_{audio\_str\_basic} \times \frac{n_{audio\_str\_basic}}{N} + \\
&\quad p\_on_{video\_str\_video\_part\_premium} \times \frac{n_{video\_str\_video\_part\_premium}}{N} + p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{N} \\
&\quad p\_on_{video\_str\_video\_part\_basic} \times \frac{n_{video\_str\_video\_part\_basic}}{N} + p\_on_{video\_str\_audio\_part} \times \frac{n_{video\_str\_audio\_part}}{N} + \\
&= 0.044 \times \frac{2}{18} + 0.029 \times \frac{2}{18} + 0.0801 \times \frac{4}{18} + 0.0291 \times \frac{4}{18} + 0.048609 \times \frac{1}{18} + 0.0457 \times \frac{1}{18} + \\
&\quad 0.219284 \times \frac{1}{18} + 0.0457 \times \frac{1}{18} + 0.109554 \times \frac{1}{18} + 0.0457 \times \frac{1}{18} \\
&= 0.06097723
\end{aligned}
$$

Therefore, the value of e_{ncp} can be obtained as:

$$
\begin{aligned}
e_{ncp} &= \sum_{i=0}^{N} i \times \binom{N}{i} \times e_p{}^i \times \left(1 - e_p\right)^{(N-i)} \\
&= \sum_{i=0}^{18} i \times \binom{18}{i} \times 0.06097723^i \times \left(1 - 0.06097723\right)^{(18-i)} = 1.09759
\end{aligned}
$$

e_{ncp} exceeds 1 to be 1.09759 and the increased channel contention will affect the QoS of the real-time UDP based service. In order to maintain the service quality, the new video call service request should be rejected.

6.2.5.4 OPNET Simulation Configurations and Results (Scenario 2)

In the simulations, the traffic generation parameters are based on Table 6.1 to Table 6.12. The OPNET simulation results are shown in Figure 6-47 to Figure 6-49, which present the average delay and the delay variation, and the audio packet arrival rate of the premium video streaming service in scenario 2.

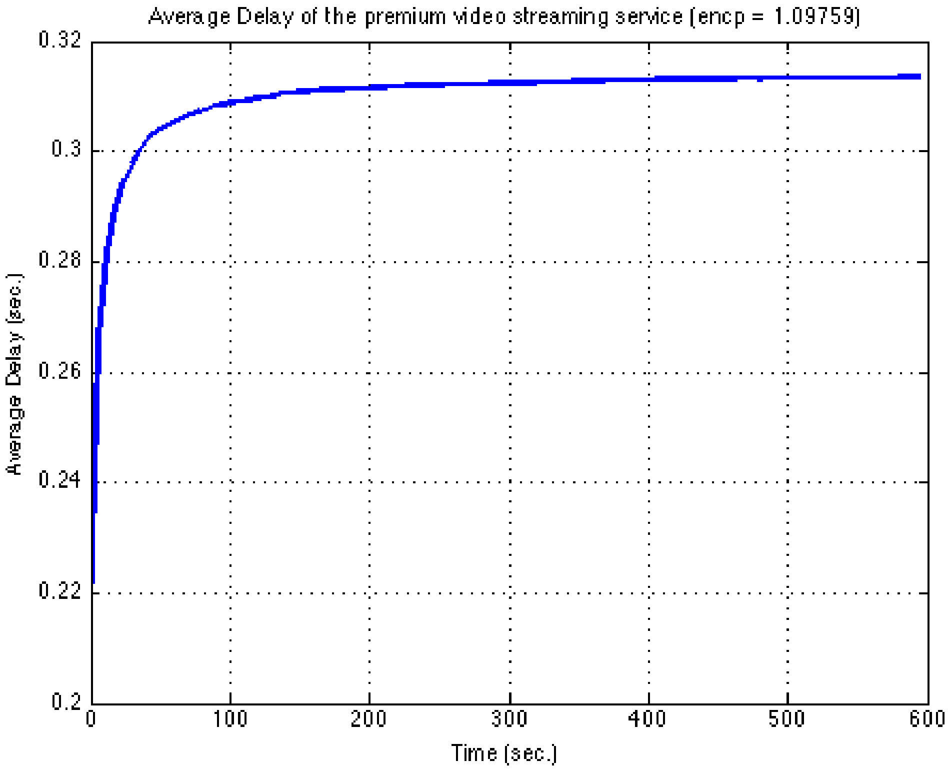


Figure 6-47 Average Delay of premium video streaming service (e_{ncp} = 1.09759)

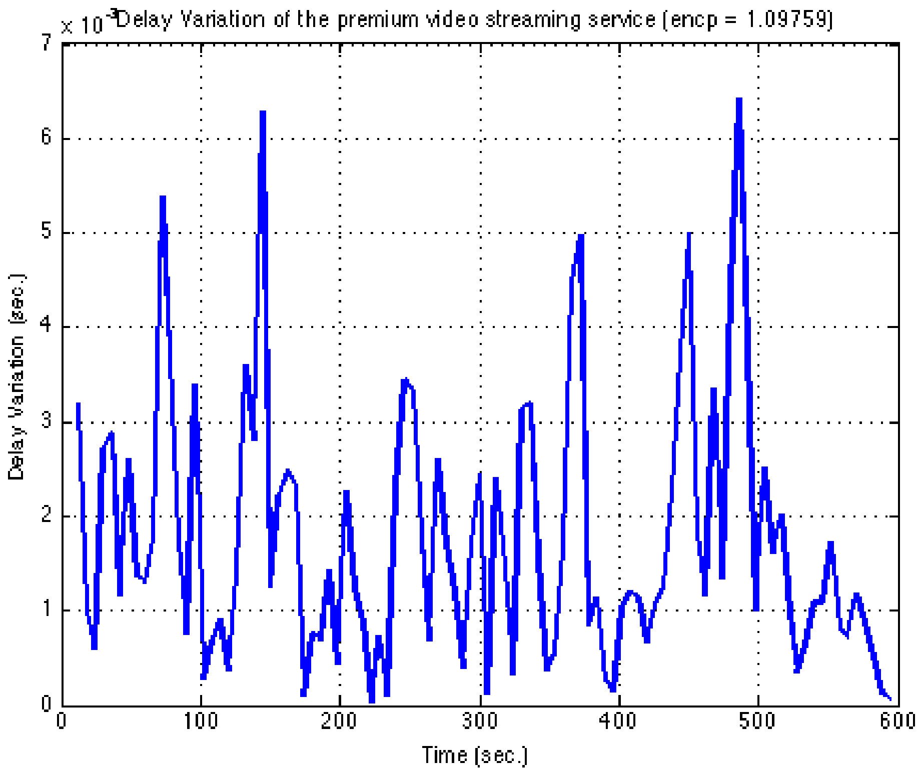


Figure 6-48 Delay Variation of premium video streaming service (e_{ncp} = 1.09759)

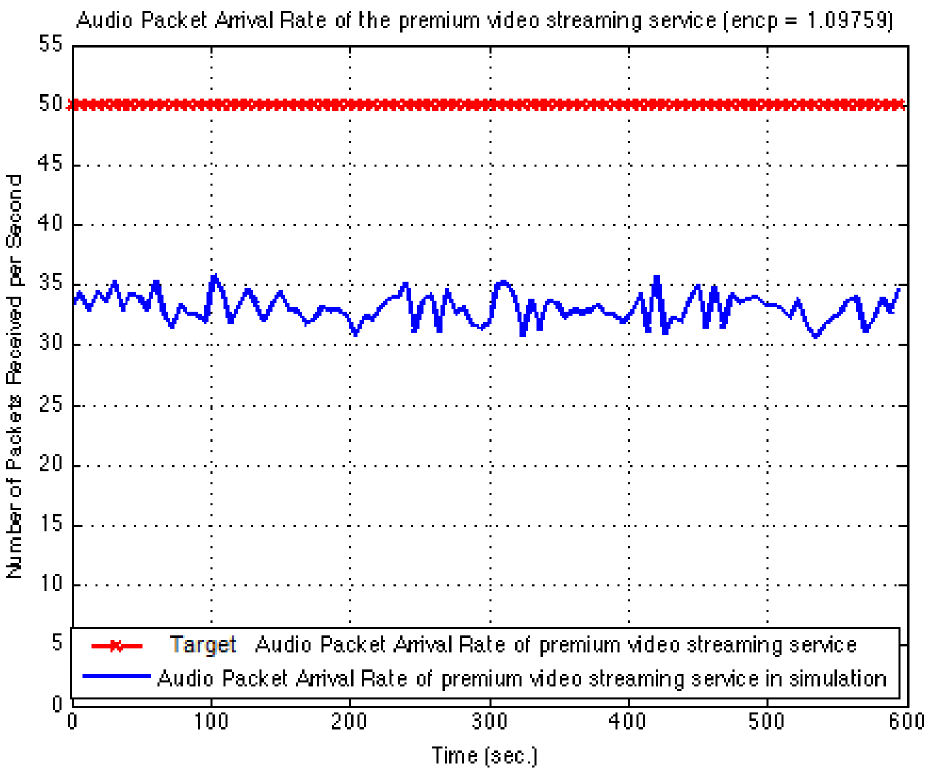


Figure 6-49 Audio Packet Arrival Rate of premium video streaming service (e_{ncp} = 1.09759)

Figure 6-47 shows that the average packet delay increases dramatically as simulation runs and achieves about 0.313 seconds. Figure 6-48 shows the fluctuations in packet delay and the delay variation is less than 6.5 ms. Both of these values comply with the QoS constraints for streaming class services, whose start-up delay should be less than 10 seconds and the delay variation should be less than 2 seconds [Ser08]. However, Figure 6-49 depicts severe fluctuations in the audio packet arrival rate. In addition, the audio packet arrival rate does not achieve the target value. Based on the codecs implemented for the basic video streaming service, the required audio packet arrival rate is 50 packets per second. Because Figures 6-49 shows that the audio packet arrival rate does not achieve the required value, it indicates that packet losses occur in the network. Figure 6-49 shows that the actual audio packet arrival rate is about 32 packets per second, which indicates that the audio packet loss rate is about 36%. However, [Ser08] requires that the audio packet loss rate should be less than 1%. Therefore, in scenario 2, the WLAN cannot provide acceptable quality of service and the new premium video streaming service request should be rejected.

6.3 Concluding Remarks

This chapter validates the network resource availability evaluation models. It demonstrates the implementation of the evaluation models and compares the evaluation results with call level and packet level simulation results for validation.

The call level simulation results validate the proposed UTRAN network resource availability evaluation model, which provides a solution to evaluate total received interference and base station transmission power without overestimating the inter-cell interference.

The OPNET simulation results show that, for most of the services, the WLAN resource availability evaluation model produces expected and consistent results. For certain services, when the value of e_{ncp} is less than but very close to 1, although the end-to-end delay and delay variation are under the QoS constraints, there are unacceptable packet losses in the network. One solution to cope with such cases is to use the value of 0.95, rather than 1, to determine the availability of network resources. For example, if the value of e_{ncp} is less than or equal to 0.95, the network resources will be evaluated as sufficient. Otherwise, the network resources are considerred as insufficient to serve the requested services. Therefore, although the evaluation model is not precise for all services, it provides a very good approximation in evaluating resource availability. Currently, the validations concentrate on the UDP based real-time services. The validation can be extended to take into account TCP based non-real-time services in the future.

The next chapter introduces the RAN selection and optimisation algorithms and present the simulation results.

Chapter 7 Radio Access Network Selection and Optimisation Algorithms

The context information, system architecture and network resource availability evaluation models presented in previous chapters enable the development of context-aware RAN selection and optimisation algorithms. In this chapter, several algorithms are introduced to facilitate an adaptive and efficient RAN selection and optimisation. The performances of these algorithms are evaluated in a series of call level simulations where two RAN, a UTRAN cell and an IEEE802.11b based WLAN, are simulated to form a heterogeneous communication environment. This chapter also presents typical values of various services and the configurations of different simulated scenarios.

7.1 Multiple Interface Selection Algorithm and Multiple Interface Selection and Service Adaptation Algorithm

An adaptive and efficient RAN selection process requires an algorithm which is able to consider several aspects of context information, including the performance and conditions of the networks, the user terminal conditions, the service types, the user preference, etc. However, the different components of context information introduce complexity to the algorithm design. The algorithm must handle the varying dynamics of context information and make the RAN selection straightforward. In this section, two algorithms, the *Multiple Interface Selection* (MIS) algorithm and the *Multiple Interface Selection and Service Adaptation* (MISSA) algorithm are presented. They constitute a simple and preliminary solution for a RAN selection algorithm.

The MIS and MISSA algorithms consider five types of services, whose abbreviations are shown below:

- VC: Video Call
- VS: Video Streaming
- AS: Audio Streaming
- WB: Web Browsing (HTTP)
- FT: File Transfer (FTP)

The first three services are real-time and the rest are non-real-time. The underline transport protocols are UDP for the real-time services and TCP for the non-real-time services. Comparing with the non-real-time services, the real-time services have higher requirements for network performance, low delay, and low packet loss. The above services can belong to two different service classes, *basic* and *premium*, which differ from each other by their QoS requirements. The *basic* service class has lower QoS requirements and provides the minimum acceptable quality the service should present. The *premium* service class has higher QoS requirements and provides a better quality of service and it is chosen when there are enough resources in the network.

When a service request arrives, the MIS algorithm considers the type of the requested service and determines the preferred network. Then, the MIS algorithm firstly investigates whether the network resources are sufficient to admit the service request with its *premium* service class. If the network resources are insufficient, the MIS algorithm will degrade the service class to the *basic* class and investigate the network resources again. If the resources are still not sufficient, a second preferred network will be selected and investigated. The operation flowchart is shown in Figure 7-1.

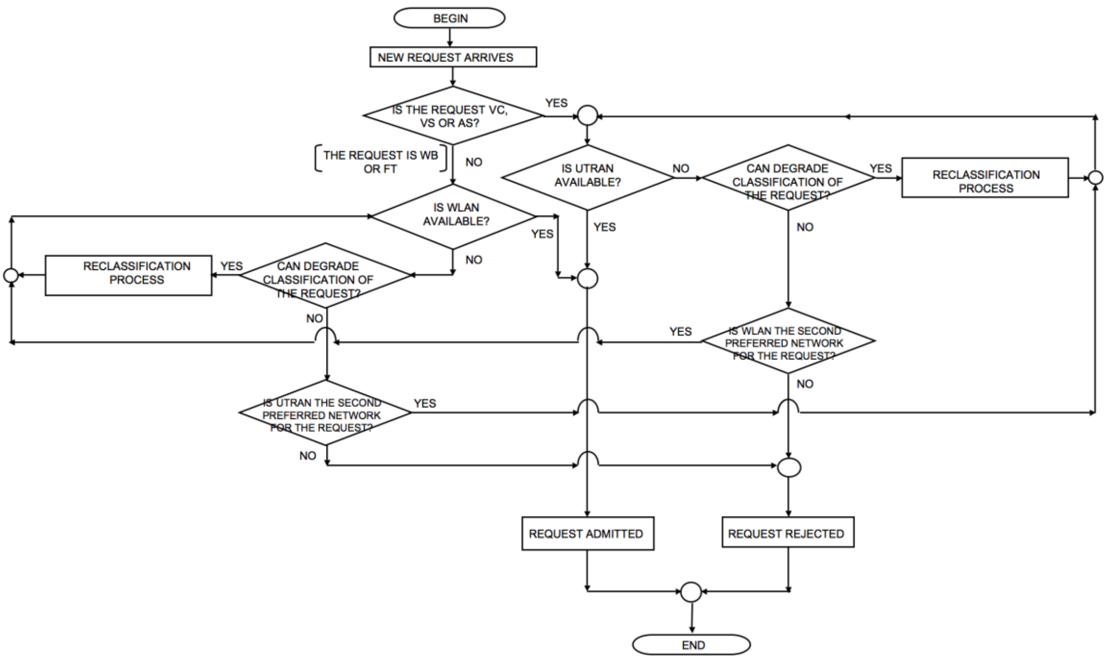


Figure 7-1 Flowchart of the MIS algorithm

The flowchart can be described as follows:

1. A new service request arrives.
2. According to the type of the requested service, the MIS algorithm selects a preferred network for it. Considering that UMTS network is able to support QoS, the preferred network for real-time services (e.g. VC, VS and AS) is the UTRAN cell. While the IEEE802.11a/b based WLAN primarily provides users with non-real-time data services, the preferred network for the non-real-time services (e.g. WB and FT) is the IEEE802.11b based WLAN.
3. Investigate whether the network resources are sufficient or not to support this new request with *premium* service class. If there are sufficient network resources to support this request, go to step 6. Otherwise, go to step 4.
4. Investigate whether the service class of the request is degradable. If it is degradable, adjust the service class from *premium* class to *basic* class and go to step 3 but to check the resource availability for *basic* class. Otherwise, select the second preferred network and then go to step 5.
5. Investigate whether the resources of the second preferred network are sufficient to support the request with *premium* or *basic* service class. If the resources are sufficient, go to step 6. Otherwise, go to step 7.
6. A new connection is made to the selected network and the status of network resources is updated. Go to step 8.
7. The service request is rejected and the status of network resources remains. Go to step 8.
8. End the algorithm.

The MISSA algorithm takes into account the existing services admitted by all networks. When the network resources are insufficient to admit a new service request, the MISSA algorithm is not only able to degrade the service class of the requested service but also able to adjust the usage of network resources by degrading the service class of some existing services admitted by the target RAN. Such operation can save extra resources possibly allowing more requests to be admitted. The flowchart of the above operation is shown in Figure 7-2.

The flowchart can be described as follows:

1. A new service request arrives.

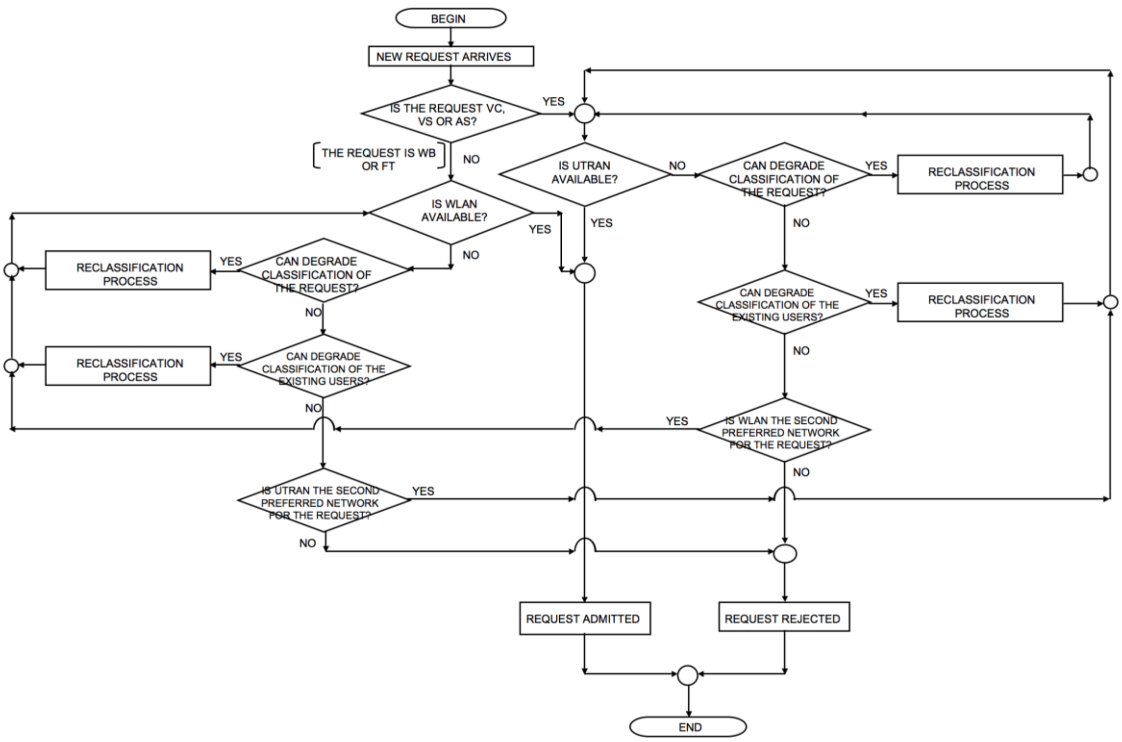


Figure 7-2 Flowchart of the MISSA Algorithm

2. According to the type of the requested service, the MISSA algorithm selects a preferred network for it. The preferred network for real-time services (e.g. VC, VS and AS) is the UTRAN cell, while for non-real-time services (e.g. WB and FT) it is the IEEE802.11b based WLAN.
3. Investigate whether the network resources are sufficient to support the service request. If there are enough network resources to support this request, go to step 8. Otherwise, go to step 4.
4. Investigate whether the service class of the request is degradable. If it is degradable, adjust the service class from *premium* class to *basic* class and go to step 3 but to check the resource availability for the *basic* class. Otherwise, go to step 5.
5. Investigate whether the service classes of the existing users are degradable. If the investigation indicates no service classes are degradable, select a second preferred network and go to step 6. Otherwise, a service class adaptation process is conducted. The adaptation is priority-based. The services are classified from low to high priorities as FT, WB, AS, VS and VC. Degradation attempt begins from the services with lowest priority. The

adaptation process only manipulates one existing service which is degradable. After finishing the adaptation process, go to step 3.

6. Investigate whether the resources of the second preferred network are sufficient to support the service request with *premium* or *basic* service class. If the resources are sufficient, go to step 7. Otherwise, perform the service class adaptation process. If the service class adaptation process is successful to save sufficient network resources for admitting the request, go to step 7. Otherwise, go to step 8.
7. A new connection is made to the selected network and the status of the network resources is updated. Go to step 9.
8. The service request is rejected and the status of the network resources remains. Go to step 9.
9. End the algorithm.

7.1.1 Simulations and Results

In order to compare and evaluate the performances of the MIS and MISSA algorithms, call level simulations are carried out. The simulations take into two types of RANs, a UTRAN cell and an IEEE802.11b based WLAN, and no background traffic is assumed [KAT07]. All the users are covered by both the UTRAN cell and the WLAN with acceptable radio channel conditions.

The radius of the UTRAN cell is 100 meters and an omnidirectional antenna is used. The maximum base station transmission power is 1 W and Twenty percent of this power is used for signalling and the remaining, 0.8 W, is used for traffic [HT07]. The average orthogonal factor in the cell is assumed to be 0.6. The interference margin is defined as 1 dB [HT07]. The average thermal noise at the base station and the user terminal are -103.2 dBm and -169.0 dBm [HT07], respectively. The radius of the WLAN is 100 meters. The FTP and HTTP servers are collocated with the AP. Furthermore, the size of the congestion window of each TCP flow is assumed to be the same and, after an initial phase, the congestion window of each TCP flow grows to its maximum value.

The simulations implement five types of services as previously described in section 7.1. The service users are randomly distributed in the coverage area. The general configuration and parameters of the services are listed in Table 7.1, which is derived based on [KAT07]. For the UDP based real-time services, the data rate values are fixed. For the TCP based non-real-time services, the data rate values are the minimum requirements for each service class.

Service Type	VC	VS	AS	WB	FT
Required Data Rate	64/128	64/128	32/64	128/256	64/128
Transport Protocol	UDP	UDP	UDP	TCP	TCP
Number of Users	7	13	2	13	9

Table 7.1 Service Parameters and Typical Values [KAT07]

The specific configuration and parameters of the services in the UTRAN cell are presented in Table 7.2. The parameters include the activity factor and the target value of the E_b/N_o in the uplink and the downlink for each service, which are derived based on [Che03], [HT07] and [LWN05].

Service Type	VC	VS	AS	WB	FT
Required Data Rate	64/128	64/128	32/64	128/256	64/128
Activity Factor	1	1	1	0.2	1
Uplink E_b/N_o (dB)	4.0/3.5	n/a	n/a	n/a	n/a
Downlink E_b/N_o (dB)	7.1/6.1	7.1/6.1	8.2/7.1	6.1/5.6	7.1/6.1

Table 7.2 Service Parameters and Typical Values for the Services in the UTRAN Cell [Che03][HT07][LWN05]

The specific configuration and parameters of the services in the WLAN include the size of the packet payload, the number of packet per second, and the packet inter arrival time. The values of the above parameters are not shown here but have been presented in section 5.2.

In the simulations, users start the services gradually. When a service request is received, the RAN selection algorithm assigns the service to an appropriate network and adjusts the usage of network resources if necessary. The performances of the algorithms are evaluated and compared in two aspects: the data rates perceived by the users and the ratio of blocked requests to the total number of requests.

The simulation results are presented in Figures 7-3, 7-4, 7-5, and 7-6, respectively. Figure 7-3 compares the data rate perceived by all users in the UTRAN cell and the WLAN when different RAN selection algorithms are used. Figure 7-4 compares the data rate perceived by the users served by the UTRAN cell and Figure 7-5 compares the data rate perceived by the users served by the WLAN when different RAN selection algorithms are implemented.

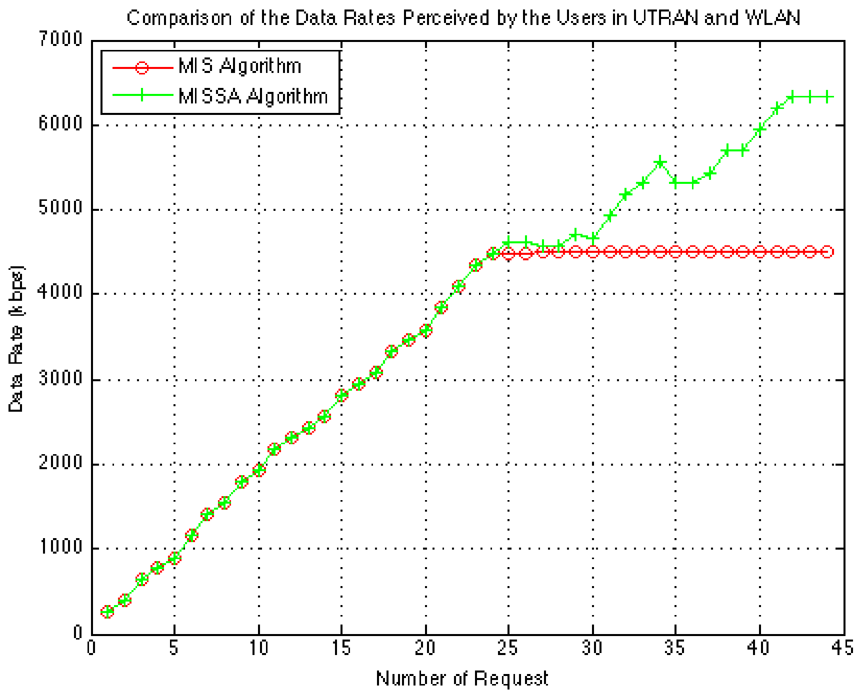


Figure 7-3 Comparison of the Data Rates Perceived by All Users in the UTRAN Cell and WLAN

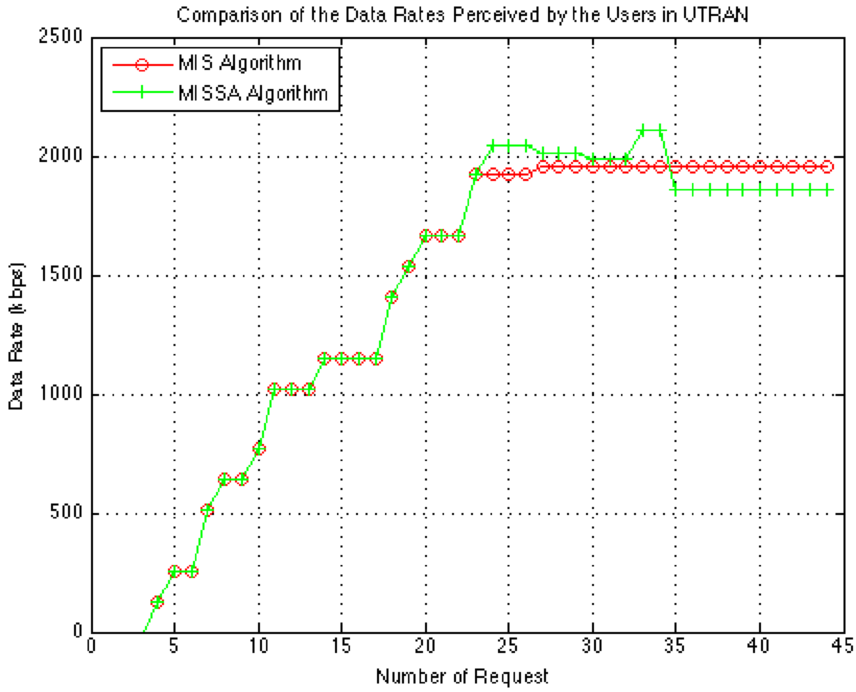


Figure 7-4 Comparison of the Data Rates Perceived by the Users Served by the UTRAN Cell

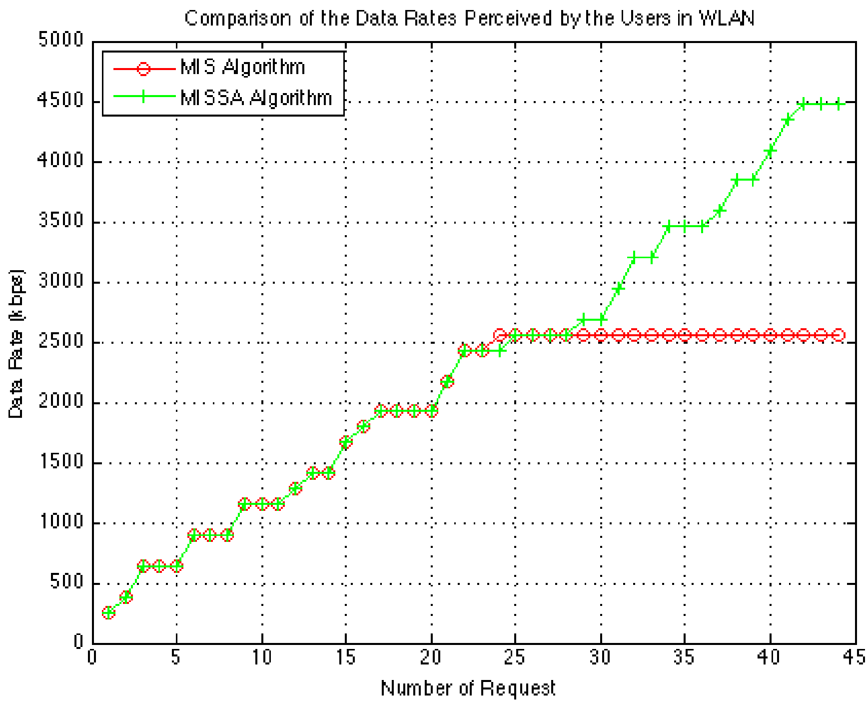


Figure 7-5 Comparison of the Data Rate Perceived by the Users Served by the WLAN

By observing the simulation results, one can note that as the number of service requests increases, the data rate perceived by the user also increases up to the point where the network resources become insufficient to admit more service requests. The simulation results presented in Figure 7-3 show that, comparing with the MIS algorithm, collectively the MISSA algorithm can allow higher data rates to be perceived by all the users. However, Figure 7-4 indicates that the users in the UTRAN cell can perceive higher data rate by implementing the MIS algorithm. In contrast, Figure 7-5 reveals that the MISSA algorithm is able to provide the users in the WLAN with better data rate.

The reason for such performance lies in the fact that, different from MISSA, the MIS algorithm only attempts to degrade the service class of the requested service. If this attempt is not successful, it will select the second preferred network for the requested service and perform the corresponding evaluation process to determine the network resource availability. For example, if the MIS algorithm determines that the UTRAN cell is unable to admit the new real-time service request, the WLAN will be selected. Because the service classes of the existing services in UTRAN cell are not adjusted, the data rates perceived by the users may be higher comparing with using the MISSA algorithm. However, once the WLAN admits a UDP based real-time service, its capacity will be dramatically impaired. This is because, if the types of the existing services in a WLAN are hybrid, the value of e_{ncp} must remain less than or equal to 1 so as to ensure the quality requirements from the existing UDP based services. Nevertheless, WLAN primarily provides users with TCP based non-real-time services. For the TCP based services, e_{ncp} can be greater than 1. But the existence of UDP based services limits the value of e_{ncp} and prevents the WLAN from admitting further TCP based services and the perceived data rates from being increased.

Figure 7-6 compares the ratios of the number of rejected requests with the number of total requests.

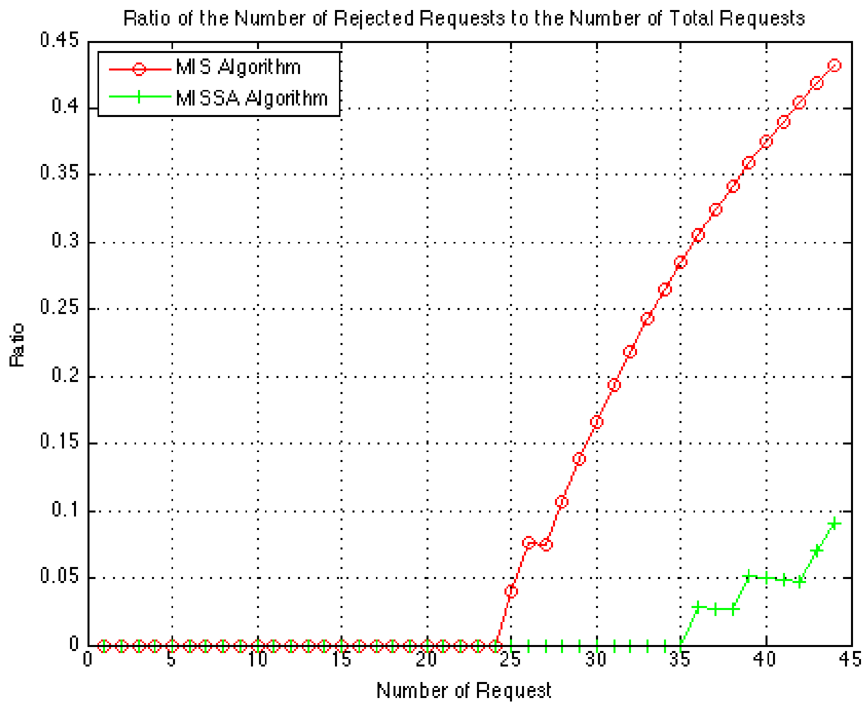


Figure 7-6 Ratio of the Number of Rejected Requests to the Number of Total Requests

Figure 7-6 shows that, before the 25th service request comes forth, the ratios generated by both algorithms are zero. Then, when MIS is used, the ratio begins to increase and reaches 43% at the end. In contrast, when MISSA is used, the ratio remains zero until the 36th service request arrives and finally reaches about 9%.

The above simulation results demonstrate that the MISSA algorithm performs better than the MIS algorithm. This is because the MISSA algorithm is able to adjust the service classes of the existing users and save extra network resources for admitting more service requests.

7.2 Radio Access Optimisation Algorithm and Objective Function based Data Rate Adaptation Scheme

The MISSA algorithm demonstrates that service adaptation is a promising approach for RAN selection. Although the service adaptation process performed by MISSA is simple and preliminary, it is inspirational and provides a direction for adjusting and optimising the utilisation of network resources. Based on the previous effort, this section introduces an efficient *Radio Access Optimisation* (RAO) algorithm and an *Objective Function* based data rate adaptation scheme.

The RAO algorithm is based on an objective function which measures the benefits obtained by a requesting user from a candidate network. Each network covering the requesting user is evaluated. The goal of the RAO algorithm is to maximise the objective function while considering two conflicting factors: the increased carried traffic and the user satisfaction. The RAO algorithm also performs data rate adaptation so as to optimise network resource utilisation. The network that provides the greatest value for the objective function by implementing RAO is selected.

7.2.1 RAO Objective Function

Assuming network A is evaluated for a requested service, the RAO objective function (OF) is calculated as:

$$OF(RS, X_A) = US(RS, X_A) + \sum_{i=1}^{N_{X_A}} US(S_i) + \sum_{X_j, j \neq A} \sum_{k=1}^{N_{X_j}} US(S_k) \qquad (6.1)$$

RS represents the requested service, X_A represents network A, N_{X_A} is the number of the existing service sessions in network A, S_i represents the *i*th existing service session in network A, X_j represents the *j*th network that is not being evaluated, N_{X_j} is the number of the existing service sessions in network X_j, and S_k represents the *k*th existing service session in network X_j. The objective function includes $US(RS, X_A)$, which is the user satisfaction experienced by the requesting user when selecting network A for the requested service *RS*, $\sum_{i=1}^{N_{X_A}} US(S_i)$ is the impact upon the user satisfaction experienced by the existing users in the network A if *RS* is admitted, and $\sum_{X_j, j \neq A} \sum_{k=1}^{N_{X_j}} US(S_k)$ is the collective user satisfaction experienced by the existing users in the networks which are not being evaluated.

The user satisfaction *US* is calculated as:

$$US(S, X) = SNCL(S, X) \times \sum_i W_{S,i} \times NORM\left(Attr_{S,i}^{X}\right) \qquad (6.2)$$

SNCL(S, X) is the *Service-Network Compatibility Level of service S in the network X*. This parameter measures the level of support a network provides for a specific service. For example, a UTRAN cell provides a better support for real-time services, such as speech, than an IEEE802.11b based WLAN. An IEEE802.11b based WLAN is better fitted to support non-real-time services, such as file transfer. $NORM\left(Attr_{S,i}^{X}\right)$ is the normalised value of an attribute $Attr_{S,i}^{X}$ belonging to service *S* and it represents the service quality provided by network *X* if this network accepts service *S*. An example of $Attr_{S,i}^{X}$ can be available data rate, data transmission delay, etc. The values of $NORM\left(Attr_{S,i}^{X}\right)$ range from 0 to 1. The greater the value is, the better the service is provided. $W_{S.i}$ is the weight representing the importance of attribute $Attr_{S,i}^{X}$ to service *S*. For example, real-time services, such as speech, are specified with high requirements for delay and jitter. Therefore, attributes representing delay or jitter will have greater weights. The value of $W_{S.i}$ also ranges from 0 to 1. Examples of the definition of the objective function are presented in Table 7.3.

Service Type	Service-Network Compatibility Level (*SNCL*)	Considered Attributes	Normalisation Function	Weights
Speech	*SNCL(VC, UTRAN)* = 0.717 *SNCL(VC, WLAN)* =0.091	Supplied date rate ***dr*** (kbps); delay ***d*** (ms); jitter ***j*** (ms); mobility support ***m***	$NORM(dr)=\begin{cases}0, & dr<6.7\\ 0.5+\frac{1}{2+2\times e^{-2\times(dr-9.45)}}, & dr\geq 6.7\end{cases}$ $NORM(d)=\frac{1}{1+e^{0.04\times(d-275)}}$ $NORM(j)=\begin{cases}0.5+\frac{0.5}{1+e^{0.55\times(j-40)}}, & j\leq 60\\ \frac{0.5}{1+e^{0.55\times(j-75)}}, & j>60\end{cases}$ *NORM(m, WLAN)* = 0.091 *NORM(m, UTRAN)* = 1	$Weight_{dr}=0.717$ $Weight_{d}=0.717$ $Weight_{j}=0.717$ $Weight_{m}=0.5$
Video Call	*SNCL(VC, UTRAN)* = 0.717 *SNCL(VC, WLAN)* =0.091	Supplied date rate ***dr***; delay ***d***; jitter ***j***; mobility support ***m***	$NORM(dr)=\begin{cases}0, & dr<64\\ 0.5+\frac{1}{2+2\times e^{-2\times(dr-9.6)}}, & dr\geq 64\end{cases}$ *NORM*(*d*): the same as Speech *NORM*(*j*): the same as Speech *NORM(m, WLAN)* = 0.091 *NORM(m, UTRAN)* = 1	$Weight_{dr}=0.717$ $Weight_{d}=0.717$ $Weight_{j}=0.717$ $Weight_{m}=0.5$
Video Streaming	*SNCL(VS, UTRAN)* = 0.717 *SNCL(VS, WLAN)* =0.091	Supplied date rate ***dr***; mobility support ***m***	$NORM(dr)=\begin{cases}0, & dr<16\\ \frac{1}{1+e^{-0.06\times(dr-48.5)}}, & dr\geq 16\end{cases}$ *NORM(m, WLAN)* = 0.091 *NORM(m, UTRAN)* = 1	$Weight_{dr}=0.909$ $Weight_{m}=0.283$
Audio Streaming	*SNCL(AS, UTRAN)* = 0.717 *SNCL(AS, WLAN)* = 0.283	Supplied date rate ***dr***; mobility support ***m***	$NORM(dr)=\begin{cases}0, & dr<32\\ \frac{1}{1+e^{-0.1743\times(dr-24.25)}}, & dr\geq 32\end{cases}$ *NORM(m, WLAN)* = 0.091 *NORM(m, UTRAN)* = 1	$Weight_{dr}=0.909$ $Weight_{m}=0.283$
Web Browsing	*SNCL(WB, UTRAN)*=0.5 *SNCL(WB, WLAN)* =1	Supplied date rate ***dr***; mobility support ***m***	$NORM(dr)=\begin{cases}0, & dr<32\\ 1-e^{-5.85\times dr/384}, & dr\geq 32\end{cases}$ *NORM(m, WLAN)* = 0.091 *NORM(m, UTRAN)* = 1	$Weight_{dr}=0.717$ $Weight_{m}=0.091$
File Transfer	*SNCL(FT, UTRAN)* =0.283 *SNCL(FT, WLAN)* =0.909	Supplied date rate ***dr***; mobility support ***m***	$NORM(dr)=\begin{cases}0, & dr<32\\ 1-e^{-11.4\times dr/384}, & dr\geq 32\end{cases}$ *NORM(m, WLAN)* = 0.091 *NORM(m, UTRAN)* = 1	$Weight_{dr}=0.717$ $Weight_{m}=0.091$

Table 7.3 Parameters of the Objective Function

The value of *SNCL(S, X)* ranges from 0 to 1, where 0 means minimum compatibility and 1 means maximum compatibility. These crisp numbers are converted from linguistics terms, including *extremely high*, *very high*, *high*, *medium*, *low*, *very low*, and *extremely low*. For example, the *SNCL* of a video call service in a UTRAN cell is defined as *high*, and the *SNCL* of a video call service in an IEEE802.11b based WLAN is defined as *very low*. These linguistic terms are converted to crisp numbers based on a fuzzy scoring method [CHH92]. For the linguistic term '*high*', the corresponding crisp number is *0.717*. For the linguistic term '*very low*', the crisp number is *0.091*.

For speech, video call, video and audio streaming services, the *data rate* attribute provided by the network is normalised based on a *Sigmoid Function*, when the minimum required data rate is achieved. The Sigmoid function has been used before to estimate user satisfaction (perceived QoS) [XSC01]. As the data rate increases, the user satisfaction also increases. As shown in Figure 7-7, a sigmoid curve has a convex and a concave characteristic. For services like video call, when the data rate is quite low or very high, an increase of data rate will not significantly improve the user satisfaction. This is because, at low data rates, the increase in data rate needs to be significant in order to change the perceived QoS by the user. At high data rates, the perceived QoS is good or excellent, an increase in data rate will hardly affect further the user's perception. Furthermore, the sigmoid function is also used to normalise the *delay* and *jitter* attributes. However, as shown in Figure 7-8, in contrast to the data rate, as the delay or the jitter increases, the user satisfaction deteriorates.

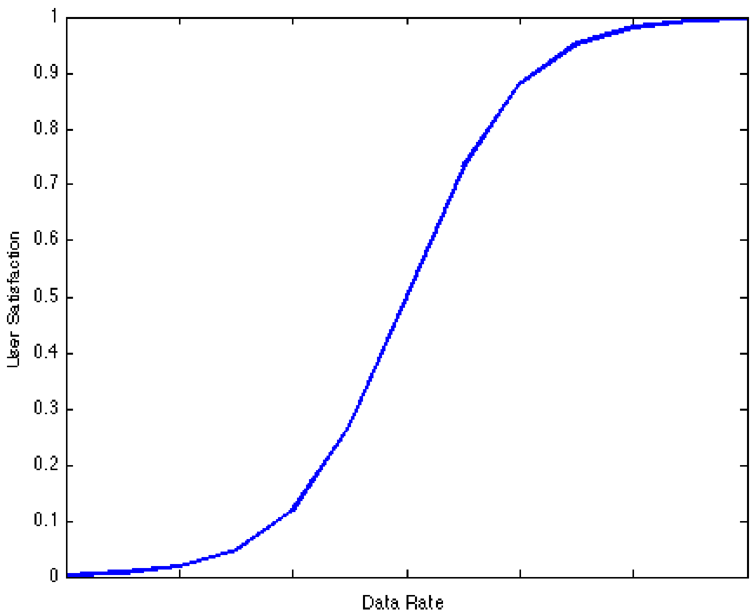


Figure 7-7 Curve of the Normalisation Function for the Delay and Jitter of Speech, Video Call, Video Streaming and Audio Streaming Services

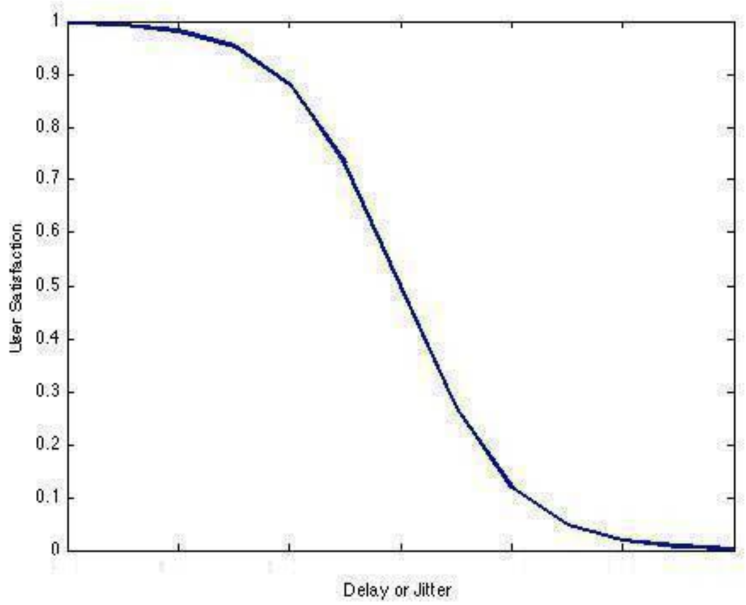


Figure 7-8 Curve of the Normalisation Function for the Delay and Jitter of Speech, Video Call, Video Streaming and Audio Streaming Services

Web browsing and file transfer services possess a bursty pattern. The increase in data rate has a significantly positive effect on the user satisfaction up to the high data rate values. Therefore, for these services, the *data rate* attribute is normalised by an exponential function [RGC01], whose curve is presented in Figure 7-9.

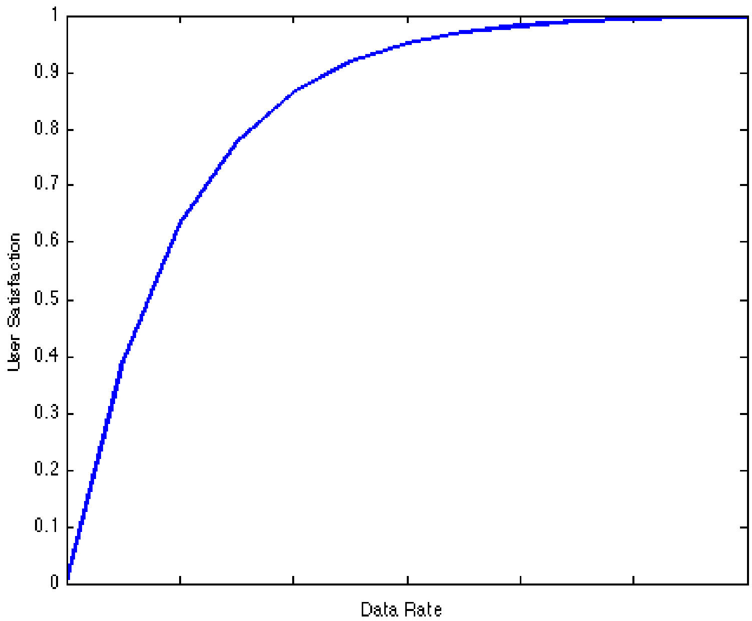


Figure 7-9 Curve of the Normalisation Function for the Data Rate of Web Browsing and File Transfer Services

The mobility support provided by the WLAN is defined as *very low*. The mobility support provided by the UTRAN cell is defined as *extremely high*. Based on a fuzzy scoring method [CHH92], for the linguistic terms '*very low*', the corresponding crisp number is 0.091. For the linguistic term '*extremely high*', the crisp number is *1*.

The values shown in Table 7.3 are based on the analysis of the characteristics and capabilities of the RANs, as well as the characteristics and requirements of the service types. These values present examples for defining objective functions.

7.2.2 Data Rate Adaptation Scheme

When the resources in network A are insufficient to accept a user request, the RAO performs a data rate adaptation scheme. This scheme considers the requirements from the requested service for data rate and the data rates allocated to the existing users. The scheme will attempt to decrease the data rate of selected services in order to obtain sufficient network resources for the new requested service. The aim of the rate adaptation is to find the best balance between the increased carried traffic and the user satisfaction.

The data rate adaptation scheme is an iterative process. Firstly it groups the required data rate from the new user request and the data rates of the existing users served by the same network into a vector V_X . In each iteration, the scheme selectively decreases the data rate of one service, and the data rate vector V_X is updated with the decreased data rate. This process stops when one of the following conditions are reached:

1. The required data rates for supporting all users in the vector V_X are reached within the network constraints.
2. The adaptation is found unfeasible.

7.2.2.1 Adaptation in a UTRAN Network — Downlink

Given a data rate vector V_X , the required base station power P_{BS} can be obtained according to Equation 5.19. When there are not sufficient resources in the UTRAN cell, the value of P_{BS} will depend on the following two overload phases:

- *Negative overload phase*: In this case, the users are limited by intra-cell interference, which cannot be overcome by increasing the base station power. In Equation 5.19,

$$\sum_{m=1}^{N+1}\left[\frac{(E_b/N_0)_m \times v_m \times R_m}{W}\times\left[(1-\alpha)+\sum_{\substack{k=1\\k\neq i}}^{M}\frac{L_{BS\_in\_cell\_k,mth\_user\_in\_central\_cell}}{L_{BS\_in\_central\_cell,mth\_user\_in\_central\_cell}}\right]\right]$$

(denoted as D_{BS} for simplicity) will be greater than 1; P_{BS} becomes negative because the data rates required by all the users in the cell are too high and/or the number of users is too large. Infinitely increasing the base station power still will not satisfy the demands.
- *Positive overload phase*: In this case, the users are limited by inter-cell interference and thermal noise. The base station has to increase its power P_{BS} to overcome the inter-cell interference and thermal noise received by the users. When there are not sufficient resources, the required P_{BS} becomes greater than the maximum power the base station is allowed to transmit.

If V_X leads to an overload phase, the data rate adaptation scheme will decrease the data rates of certain users so as to obtain a feasible value of P_{BS}. The pseudo code of the scheme is presented in Fig. 7-10.

When the UTRAN cell is in the negative overload phase, the adaptation scheme extracts a subset DS_X from V_X . The subset DS_X includes the data rates of the services which are capable of suffering degradation but still comply with minimum QoS requirements. They are candidates for adaptation. In each round, the scheme will hypothetically decrease the data rate of each candidate service belonging to DS_x by one level and calculate the ratio of $D_{BS}-D_{BS}^{'}$ to the outcome difference of the objective function, $OF-OF^{'}$ (difference of the values before and after decreasing the data rate). The ratio for the *i*th service is denoted as $RT(R_i)$, where R_i is the data rate of the *i*th service. The service whose hypothetical data rate degradation results in the greatest ratio will be selected for an actual degradation. The selected data rate is denoted as R_m. After degradation, a new value, $R_m^{'}$, is obtained. Then, R_m will be deleted from DS_x and the data rate vector V_x will be updated with $R_m^{'}$. The next round of adaptation will proceed until the UTRAN cell moves to the positive overload phase, or a feasible P_{BS} is reached, or no more data rates can be decreased.

Given a data rate vector $V_X = [R_1, R_2, ..., R_{N_X}, R_{N_X+1}]$, calculate the Objective Function *OF*

Negative overload phase:

Take V_X and calculate D_{BS} according to Equation 5.19
Extract the subset $DS_X = [R_1, R_2, ...R_i, ...R_m, ...]$
While the UTRAN is in the negative overload phase
  For each user *i* in DS_X
    Decrease the data rate R_i to a lower level as $R_i^{'}$
    Form a vector $V_X^{'} = [R_1, R_2, ..., R_i^{'}, ..., R_{N_X+1}]$, calculate $D_{BS}^{'}$ and *OF'*
    Calculate the ratio $RT(R_i)$:

$$RT(R_i) = \frac{D_{BS} - D_{BS}^{'}}{OF - OF'}$$

  End For
  Among the calculated *RT*s, select the service whose degraded data rate (R_m) supplies the greatest *RT* value: $R_m = \arg\max RT(R_i)$
  Delete R_m from DS_X
  Form a new $V_X = [R_1, R_2, ..., R_m^{'}, ..., R_{N_X+1}]$
  Take the new V_X and calculate P_{BS} according to Equation 5.19.
End While

Positive overload phase:

Take V_X and calculate P_{BS} according to Equation 5.19.
Extract the subset $DS_X = [R_1, R_2, ...R_i, ...R_m, ...]$
While the UTRAN is in the positive overload phase
  For each service *i* in DS_X
    Decrease the data rate R_i to a lower level as $R_i^{'}$
    Form a vector $V_X^{'} = [R_1, R_2, ..., R_i^{'}, ..., R_{N_X+1}]$, calculate $P_{BS}^{'}$ and *OF'*
    Calculate the ratio $RT(R_i)$:

$$RT(R_i) = \frac{P_{BS} - P_{BS}^{'}}{OF - OF'}$$

  End For
  Among the calculated *RT*s, select the service whose degraded data rate (R_m) supplies the greatest RT value: $R_m = \arg\max RT(R_i)$
  Delete R_m from DS_X
  Form a new $V_X = [R_1, R_2, ..., R_m^{'}, ..., R_{N_X+1}]$
  Take the new V_X and calculate P_{BS} according to Equation 5.19
End While

Figure 7-10 Pseudo Code of the Data Rate Adaptation Scheme for UTRAN Downlink

When the UTRAN cell is in the positive overload phase, similarly to the process in the negative overload phase, a subset DS_x will be extracted from V_x. In each round, the scheme will hypothetically decrease the data rate of each candidate service belonging to DS_x by one level and calculate the ratio of the difference of the base station power, $P_{BS} - P_{BS}^{'}$, to the difference of the objective function, $OF - OF^{'}$. The candidate service whose hypothetical data rate (R_m) degradation results in the greatest ratio will be selected for an actual rate degradation. Then, R_m will be deleted from DS_x and the data rate vector V_x is updated with $R_m^{'}$ and the next round of adaptation will proceed until a feasible P_{total} is reached or no more data rates can be decreased.

In both overload phases, the adaptation scheme aims to maximise the reduction of power consumption and minimise the loss of user satisfaction.

7.2.2.2 Adaptation in a UTRAN Network — Uplink

Given a data rate vector V_X, the total interference received by the base station, I_{BS}, can be obtained by implementing the iterative algorithm presented in subsection 5.1.1. If the noise rise introduced by the total interference exceeds the interference margin, the UTRAN cell is overloaded and the data rate adaptation scheme will be carried out to decrease the data rates of certain users so as to obtain a feasible value of I_{BS}. The pseudo code of the scheme is presented in Figure 7-11.

When the UTRAN cell is in the overload phase, the adaptation scheme extracts a subset DS_X from V_X. Similar to the scheme for the downlink, the subset DS_X includes the data rates of the degradable services through their minimum QoS requirements. In each round, the scheme will hypothetically decrease the data rate of each candidate service belonging to DS_X by one level and calculate the ratio of the difference of total interference received at the base station, $I_{BS} - I_{BS}^{'}$, to the difference of the objective function, $OF - OF^{'}$. The ratio for the *i*th service is denoted as $RT(R_i)$, where R_i is the data rate of the *i*th service. The service whose hypothetical data rate degradation results in the greatest ratio will be selected for an actual degradation. The selected data rate is denoted as R_m. After degradation, a new value, $R_m^{'}$, is obtained. Then, R_m will be deleted from DS_X and the data rate vector V_X will be updated with $R_m^{'}$. The next round of adaptation will proceed until a feasible I_{BS} is reached, or no more data rates can be decreased.

Given a data rate vector $V_X = [R_1, R_2, ..., R_{N_X}, R_{N_X+1}]$, calculate the Objective Function *OF*

Take V_X and calculate I_{BS}
Extract the subset $DS_X = [R_1, R_2, ...R_i, ...R_m, ...]$
While the UTRAN is in the overload phase
  For each user *i* in DS_X
    Decrease the data rate R_i to a lower level as $R_i^{'}$
    Form a vector $V_X^{'} = [R_1, R_2, ..., R_i^{'}, ..., R_{N_X+1}]$, calculate $I_{BS}^{'}$ and *OF'*
    Calculate the ratio $RT(R_i)$:

$$RT(R_i) = \frac{I_{BS} - I_{BS}^{'}}{OF - OF'}$$

  End For
  Among the calculated *RT*s, select the service whose degraded data rate (R_m) supplies the greatest *RT* value: $R_m = \arg\max RT(R_i)$
  Delete R_m from DS_X
  Form a new $V_X = [R_1, R_2, ..., R_m^{'}, ..., R_{N_X+1}]$
  Take the new V_X and calculate I_{BS}
End While

Figure 7-11 Pseudo Code of the Data Rate Adaptation Scheme for UTRAN Uplink

The adaptation scheme aims to maximise the reduction of received interference at the base station and minimise the loss of user satisfaction.

7.2.2.3 Adaptation in an IEEE802.11b based WLAN

Given a data rate vector V_X and service characteristics, the expected number of contending packets e_{ncp} can be obtained. When there are insufficient resources in the WLAN, the value of e_{ncp} will depend on the following two overload phases:

- *Hybrid overload phase*: In this case, the service types are hybrid. The users are limited by packet collisions. The packet collisions will cause delays and packet loss for the real-time UDP based services and there are no more guarantees that the delay and packet loss will be acceptable according to the requirements of the services. e_{ncp} is greater than 1.

- *Non-real-time service overload phase*: In this case, the service types are all TCP based. The users are limited by packets contention to access the network channel. The effective packet transmission rate calculated according to e_{ncp} cannot satisfy the lowest service level.

If the WLAN enters into the hybrid overload phase, similar to the UTRAN cell, the data rate adaptation scheme will selectively decrease the data rates of certain real-time services in order to obtain a value of e_{ncp} less than 1. The data rate of some real-time services can be lowered by adjusting their encoders. The pseudo code of the scheme is presented in Figure 7-12.

However, the adaptation scheme will not be applied to the non-real-time service overload phase. This is because the data rates of non-real-time services depend on the channel contention in the WLAN, but not on the encoder. Also, the congestion window of each TCP based service session is assumed to be the same. Consequently, no further adjustment is available.

Given a data rate vector $V_X = [R_1, R_2, ..., R_{N_X}, R_{N_X+1}]$, calculate the Objective Function *OF*

Hybrid overload phase:

Take V_X and service characteristics, calculate e_{ncp}
Extract the subset $DS_X = [R_1, R_2, ...R_i, ...R_m, ...]$
While the WLAN is in the hybrid overload phase
  For each service *i* in DS_X
    Decrease the data rate R_i to a lower level as R_i'
    Form a vector $V_X' = [R_1, R_2, ..., R_i', ..., R_{N_X+1}]$, calculate e_{ncp}' and *OF'*
    Calculate the ratio $RT(R_i)$:

$$Rt_n = \frac{e_{ncp} - e_{ncp}'}{OF - OF'}$$

  End For
  Among the calculated *RT*s, select the service whose degraded data rate (R_m) supplies the greatest RT value: $R_m = \arg\max RT(R_i)$
  Delete R_m from DS_X
  Form a new $V_X = [R_1, R_2, ..., R_m', ..., R_{N_X+1}]$:
  Take V_X and service charateristics, calculate e_{ncp}
End While

Figure 7-12 Pseudo Code of the Data Rate Adaptation Scheme in WLAN

When the WLAN is in the hybrid overload phase, the adaptation scheme extracts a subset DS_X from V_X. The subset DS_X includes the data rates of the services which are capable of suffering degradation, but they still would comply with minimum QoS requirements. In each iteration, the scheme will hypothetically decrease the data rate of each candidate service belonging to DS_X by one level and calculate the ratio of $e_{ncp} - e'_{ncp}$ to the outcome difference of the objective function, $OF - OF'$ (difference of the values before and after decreasing the data rate). The ratio for the *i*th service is denoted as $RT(R_i)$, where R_i is the data rate of the *i*th service. The service whose hypothetical data rate degradation results in the greatest ratio will be selected for an actual degradation. The selected data rate is denoted as R_m. After degradation, a new value, R'_m, is obtained. Then, R_m will be deleted from DS_x and the data rate vector V_x will be updated with R'_m. The next round of adaptation will proceed until a feasible e_{ncp} is reached, or no more data rates can be decreased.

The adaptation scheme aims to minimise channel contention and maximise user satisfaction.

7.2.3 Radio Access Network Selection Simulation and Results

In order to evaluate the performance of the RAO algorithm, call level simulations are carried out. The simulation configuration and scenario definition are identical to the simulations carried out in section 7.1. The simulation results compare the RAO, MIS, and MISSA algorithms in three aspects: the data rate perceived by the users, the value of objective function, and the ratio of blocked requests to the total number of requests.

The simulation results are presented in Figure 7-13, 7-14, 7-15, 7-16, 7-17, 7-18, and 7-19. Figure 7-13 compares the collective data rate perceived by all users in the UTRAN cell and the WLAN when different RAN selection algorithms are used. Figure 7-14 compares the data rate perceived by the users in the UTRAN cell and Figure 7-15 compares the data rate perceived by the users in WLAN when different RAN selection algorithms are implemented.

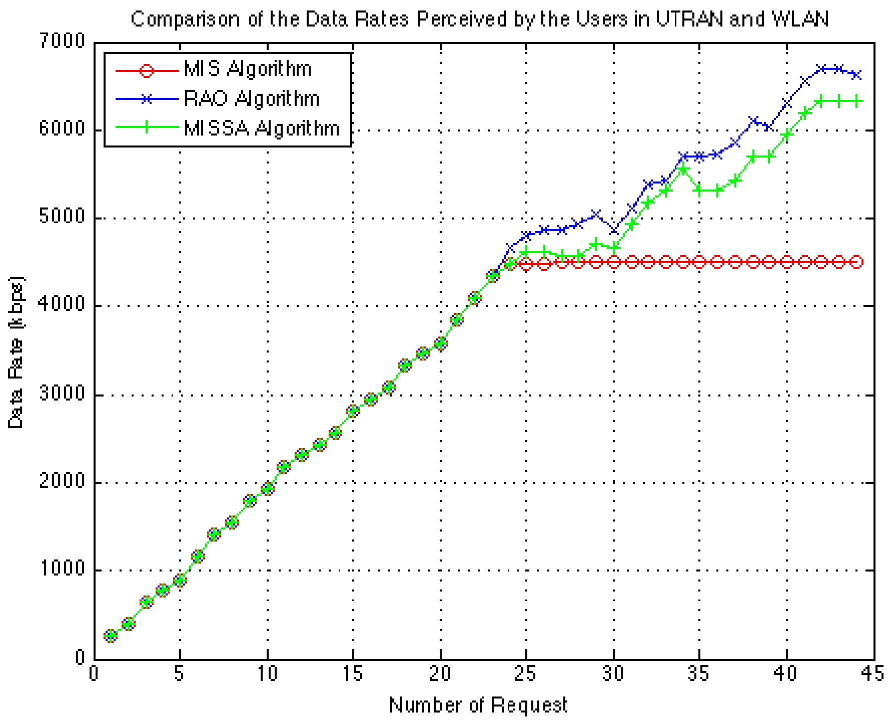


Figure 7-13 Comparison of the Data Rates Perceived by all Users in the UTRAN Cell and the WLAN

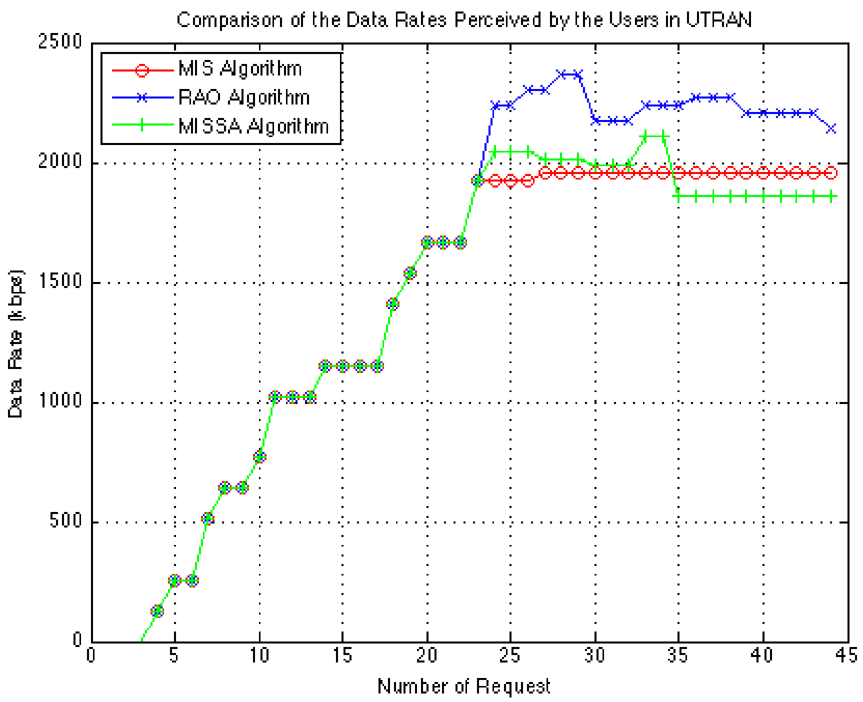


Figure 7-14 Comparison of Data Rate Perceived by the Users Served by the UTRAN Cell

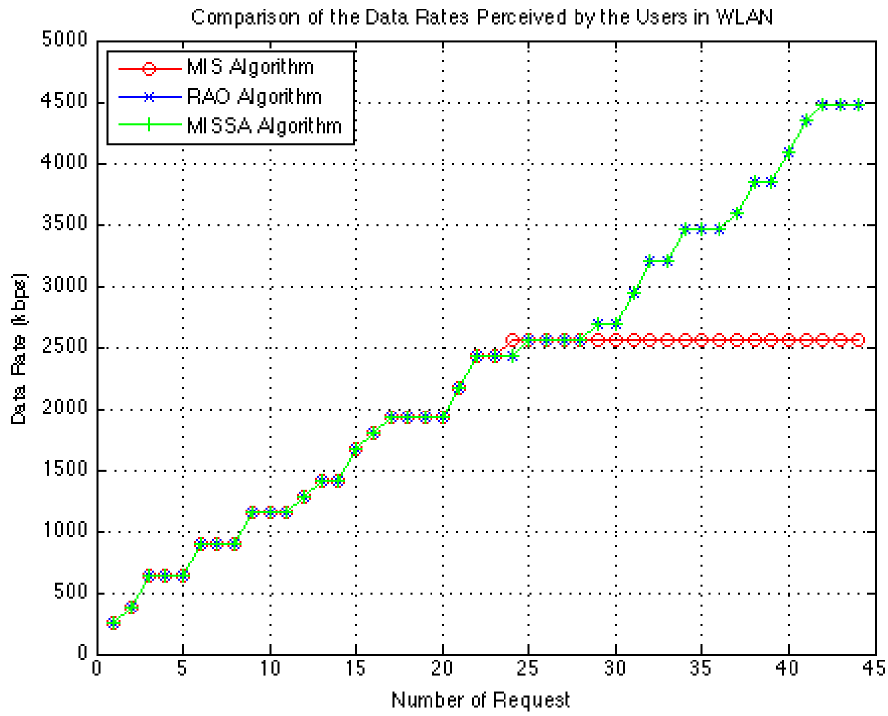


Figure 7-15 Comparison of Data Rate Perceived by the Users Served by the WLAN

By observing the simulation results shown in Figures 7-13 and 7-14, they indicate that, comparing with MISSA and MIS, collectively the RAO algorithm can allow higher data rates to be perceived by all the users. The MISSA algorithm can achieve the same performance as RAO only in the WLAN.

Figure 7-16 compares the collective objective function value provided by the UTRAN cell and the WLAN when different RAN selection algorithms are used. Figure 7-17 compares the objective function value provided by the UTRAN cell and Figure 7-18 compares the objective function value provided by WLAN when different RAN selection algorithms are implemented. The greater is the value of the objective function, the higher is the level of user satisfaction.

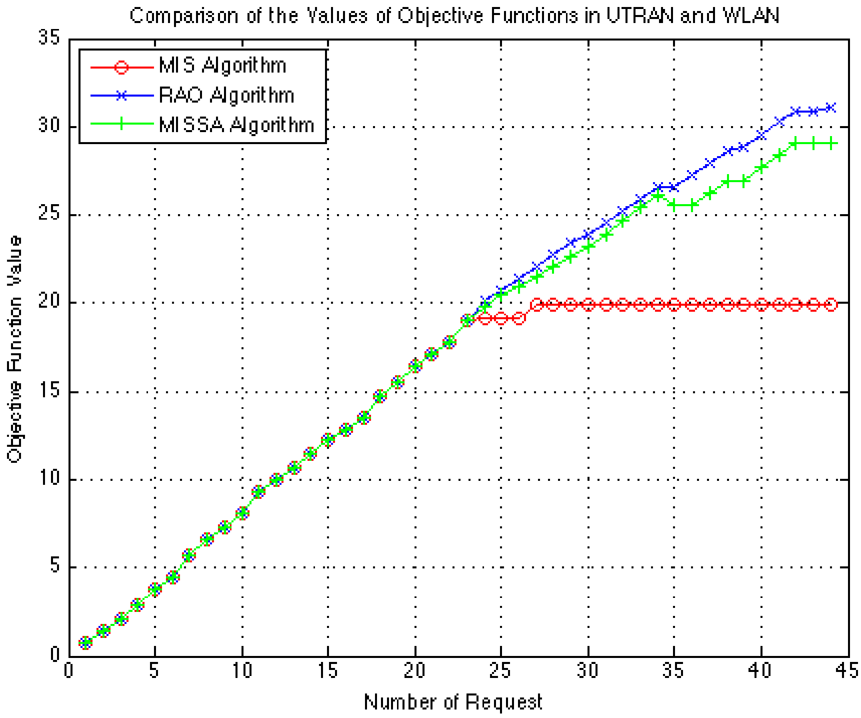


Figure 7-16 Comparison of the Collective Objective Function Values Provided by the UTRAN Cell and the WLAN

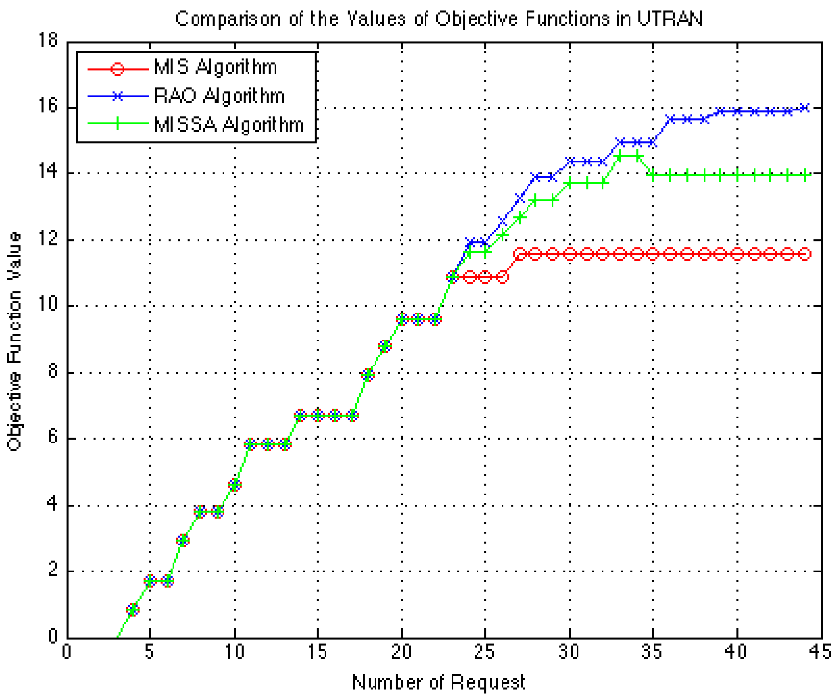


Figure 7-17 Comparison of the Value of Objective Function Provided by the UTRAN Cell

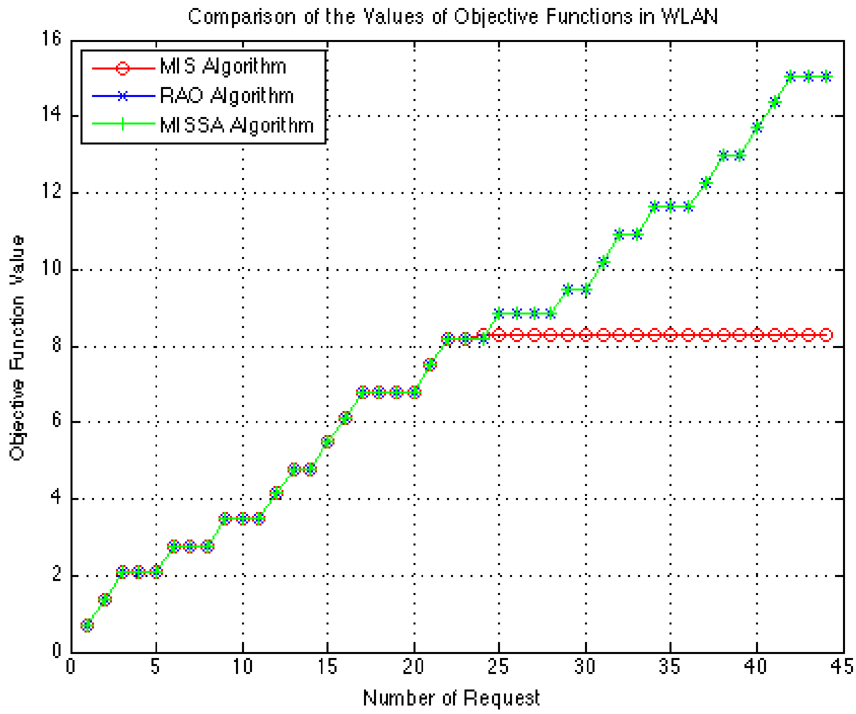


Figure 7-18 Comparison of the Value of Objective Function Provided by the WLAN

The simulation results in Figures 7-16 and 7-17 demonstrate that, collectively the RAO algorithm can enable the RANs to provide higher values for the objective functions when comparing with MISSA and MIS. As presented in Figure 7-18, the MISSA algorithm can only enable the WLAN to achieve the same performance as RAO.

Figure 7-19 compares the ratio of the number of rejected requests to the number of total requests. Figure 7-19 shows that, before the 25th service request arrives, the rejection ratio generated by all algorithms is zero. When the MIS algorithm is used, the rejection ratio increases after the arrival of the 25th service request and reaches 43% in the end of the simulation. In contrast, when the MISSA and the RAO algorithms are used, the rejection ratio remains zero, until the arrival of 36th and 35th service requests, respectively. In the end, the rejection ratio generated by the MISSA algorithm reaches 9%. For the RAO algorithm, the rejection ratio reaches about 5%.

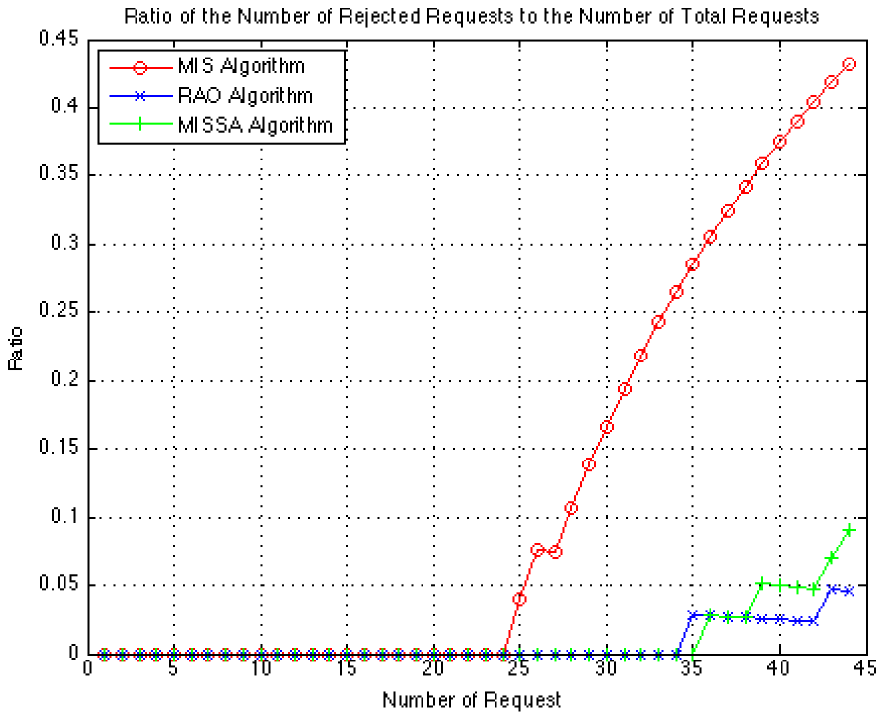


Figure 7-19 Ratio of the Number of Rejected Requests to the Number of Total Requests

The above simulation results demonstrate that the use of RAO algorithm can improve network performance and promote higher level of user satisfaction when comparing with MISSA. This is because the RAO algorithm not only performs an objective function based RAN selection process, but it also employs an efficient and appropriate data rate adaptation scheme which facilitates network resource usage optimisation.

7.3 Policy-based Radio Access Selection and Optimisation Algorithm

This section presents a *Policy-based Radio Access Selection and Optimisation* (P-RASO) algorithm. The P-RASO algorithm is adaptive and works with multiple policies. It can cope with different scenarios, network conditions and operator's aims by implementing different policies in a dynamic way. The policies measure the gains obtained from every candidate network covering a requesting user and select the network which can provide the greatest gains. The policies evaluate the obtained gains in different ways. Some may consider the gains as having a larger group of users. Some may regard the gains as having a higher level of overall user satisfaction. Some may view the gains as obtaining higher revenue. The gains can be represented by the values obtained from objective functions. Similar to the RAO algorithm, the aim of each policy is to maximise the corresponding objective function. The adaptivity of the P-RASO algorithm lies not only in its ability of implementing different policies according to different scenarios, network conditions and aims of operators, but also in the way the policies maximise the objective function when facing varying network resource availability.

Due to user arrival and departure, network resource availability may vary frequently during the lifetime of an active service session. Such variability may affect QoS guarantees and user admission where adaptations to the service sessions and network resource utilisation would be required. It makes the selection and optimisation process even more dynamic and complex. As network resource availability changes, inside each policy, a data rate adaptation may be performed in order to optimise the network resource utilisation while considering two factors: the varying carried traffic and the objective function value. For example, when the resources in network A are not sufficient to accept a user request, the data rate adaptation considers the new user's data rate requirement and takes into account the data rates allocated to the existing users. Then, it extracts a subset of the existing users who are capable of suffering data rate degradation without violating their QoS guarantees. The data rate adaptation decreases the data rates of certain users in the subset to obtain sufficient network resources for admitting the new user. The above degradation may result in a lower objective function value. The guideline of this adaptation is to obtain as many network resources as possible from each degraded user, meanwhile maintaining the objective function value as great as possible. The above adaptation process has been explained and demonstrated in section 7.2. On the other hand, when an existing user leaves network A and releases the resources formerly being consumed, the data rate adaptation just needs to consider the data rates allocated to the existing users. It extracts a subset of the existing users who can have their data rates upgraded. The data rate adaptation increases the data rates of certain users in the subset to consume the extra network resources. The above upgrade may result in a greater objective function value. The aim of this adaptation is to increase the user satisfaction but at the same time consume as little extra network resources as possible, meanwhile maximising the objective function value as much as possible. The aims of both data rate adaptation schemes are to effectively consume the network resources and to maximise the gains. The data rate adaptation scheme performed in the case of having extra resources will be explained later in this section.

7.3.1 Implementation of Multiple Policies

In order to better understand the use of multiple policies, the following scenario can be taken as an example. Assuming that the P-RASO algorithm is supplied with two policies: *Overall User Satisfaction Improvement* (OUSI) policy and *User Number Increase* (UNI) policy. The OUSI policy considers the gains of having a higher level of overall user satisfaction, which is represented by the value of a specific objective function. The UNI policy regards the gains of having a larger group of users. Supposing that a new user arrives, if the OUSI policy is used, firstly it will find out what network can provide the greatest objective function value for admitting the new user. Then, the objective function value is compared with the previous objective function value generated before the new user is admitted. If the new value is greater, the network found by the OUSI policy will be selected to admit the user request. Otherwise, the user request will be rejected. The OUSI policy considers the overall user satisfaction and only admits user requests which can improve the overall user satisfaction, despite of possibly having higher blocking rate. In the contrary, if the UNI policy is used, it will directly select the network *currently* providing the greatest objective function value. For the UNI policy, when users' QoS guarantees are not violated, it will try to accept as many service requests as possible despite of possibly having lower overall user satisfaction. The implementation and dynamic change of the OUSI and the UNI policies in P-RASO is presented in Figure 7-20.

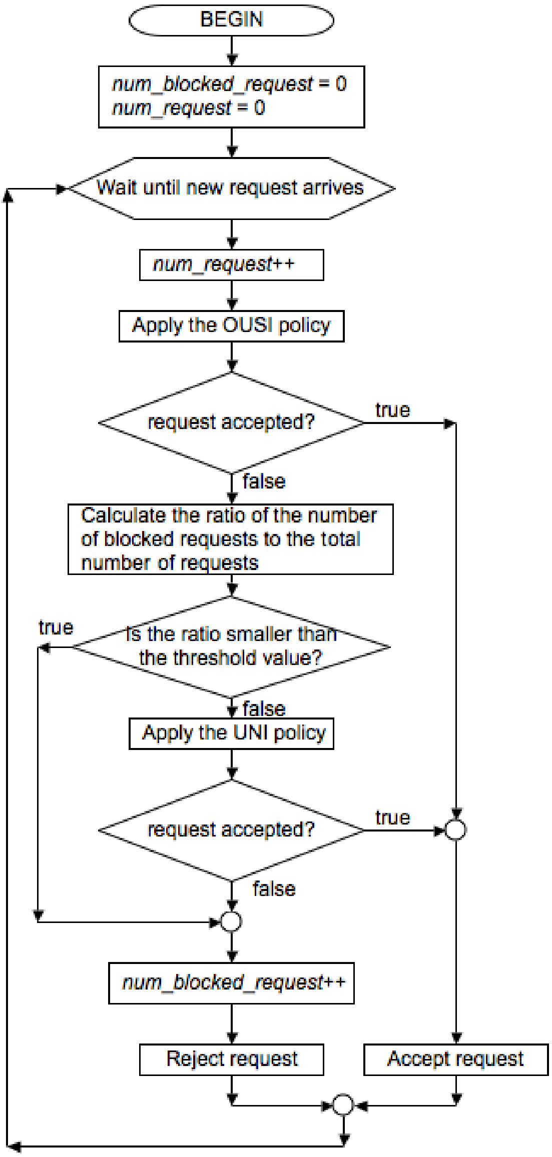


Figure 7-20 Implementation of the OUSI and UNI Policies in the P-RASO Algorithm

When a user request arrives, the OUSI policy is firstly applied for evaluation. If the OUSI policy determines that the user request cannot be admitted, the ratio of blocked requests to the total number of requests, will be calculated and compared with a predefined threshold value, which can be the greatest blocking rate that the operator can accept. If the current rejection ratio is smaller than or equal to the threshold value, the decision made by the OUSI policy will be carried out and the request will be rejected. This is because current rejection ratio is not greater than the threshold value and, in order to maintain a high overall user satisfaction, such rejection is acceptable by the operator. However, if the current rejection ratio is greater than the threshold value, in order to keep the rejection value smaller than or equal to the threshold value and admit more users, the P-RASO algorithm will change to use the UNI policy. Then, the user request will be evaluated by the UNI policy. Finally the radio access network selection and optimisation decision made by the UNI policy will be carried out.

7.3.2 P-RASO Objective Function

Assuming network A is being evaluated for a service request made by a new user, the P-RASO objective function (*OF*) is calculated as:

$$OF(RS, X_A) = G(RS, X_A) + \sum_{i=1}^{N_{X_A}} G(S_i) + \sum_{X_j, j \neq A} \sum_{k=1}^{N_{X_j}} G(S_k) \qquad (6.3)$$

RS represents the requested service, X_A represents the network A, N_{X_A} is the number of the existing service sessions in network A, S_i represents the *i*th existing service session in network A, X_j represents the *j*th network that is not being evaluated, N_{X_j} is the number of the existing service sessions in network X_j, and S_k represents the *k*th existing service session in network X_j. The objective function includes $G(RS, X_A)$, which is the measure of the gains that can be obtained when selecting network A for the new requested service *RS*, $\sum_{i=1}^{N_{X_A}} G(S_i)$ is the impact upon the gains of serving the existing users in the network A if *RS* is accepted, and $\sum_{X_j, j \neq A} \sum_{k=1}^{N_{X_j}} G(S_k)$ is the collective gains obtained from serving the rest of the existing users in the networks which are not being evaluated.

The gain function *G* is calculated as:

$$G(S, X) = SNCL(S, X) \times \sum_{i} W_{S,i} \times NORM\left(Attr_{S,i}^{X}\right) \qquad (6.4)$$

SNCL(S,X) is the *Service-Network Compatibility Level of service S in the network X*. This parameter measures the level of support a network provides for a specific service and its definition is presented in section 7.2. In heterogeneous environments, the RAN selection and optimisation are context aware and consider various types of context information such as data rate, service cost, etc. The types of context information may be different between policies considering their configurations or aims. This can be shown in $NORM\left(Attr_{S,i}^{X}\right)$, which is a normalised value presenting the gains obtained from network *X* if it accepts service *S*. $Attr_{S,i}^{X}$ represents the attribute or characteristic of the network *X*, such as available data rate, service price, etc., corresponding to a specific type of context information *i* which is being considered in the provision of service *S*. The value of $NORM\left(Attr_{S,i}^{X}\right)$ is supplied by a normalisation function and ranges from 0 to 1. The greater is the value of $NORM\left(Attr_{S,i}^{X}\right)$, the higher are the gains. Different policies may have different forms for $NORM\left(Attr_{S,i}^{X}\right)$. For the policies which focuses on user satisfaction, the definition of $NORM\left(Attr_{S,i}^{X}\right)$ presented in section 7.2 can be used. $W_{S,i}$ is the weight representing the importance of attribute $Attr_{S,i}^{X}$ to service *S*. The value of $W_{S,i}$ ranges from 0 to 1 and the definition presented in section 7.2 can be adopted.

7.3.3 Data Rate Adaptation Scheme for Extra Resources

A data rate adaptation scheme for extra resources is performed when an existing user leaves the network. In the case of having extra resources (caused by the departure of users), the data rates of the existing users served by the same network are grouped together into a vector V_Y. In each iteration, the scheme selectively increases the data rate of one service, and the data rate vector V_Y is updated with the increased data rate. This adaptation scheme stops when one of the following conditions are met:

1. No more existing services can be upgraded.
2. No more network resources are available to consume.

The data rate adaptation scheme aims to effectively consume the network resources and maximise the gains obtained by each policy. In [CLD07], Chatterjee et al. demonstrate the rate allocation problem in CDMA network is NP-Complete. The proposed data rate adaptation scheme is a heuristic method that relaxes the computation complexity and develops an efficient and effective RAN selection and optimisation algorithm.

7.3.3.1 Adaptation in a UTRAN — Downlink

Given a data rate vector V_Y, the required base station power P_{BS} can be obtained according to Equation 5.19. If V_Y leads to a P_{BS} value which is smaller than the maximum power the base station can transmit, the data rate adaptation scheme will increase the data rates of certain users so as to effectively consume the extra network resources and improve user satisfaction. The pseudo code of the scheme is presented in Figure 7-21.

```
Given data rate vector V_Y = [R_1, R_2, ..., R_{N_Y}], calculate the Objective Function OF and P_BS
Extract the subset DS_Y = [R_1, R_2, ..., R_{i,...}, R_{m,...}]
While DS_Y is not empty
  For each user i in DS_Y
     Increase the data rate R_i to a upper level as R_i'
      Form a vector V_Y' = [R_1, R_2, ..., R_i', ..., R_{N_Y}], calculate P_BS'
      If P_BS' is smaller than or equal to the maximum power that the base station can use
         Calculate OF' and the ratio RT(R_i):
                     RT(R_i) = (P_BS' - P_BS) / (OF' - OF)
      Else
         Delete R_i from DS_Y
      End If
  End For
   Among the calculated RTs, select the service whose data rate (R_min) supplies the smallest
   RT value: R_min = arg min RT(R_i)
  Increase the data rate R_min to a upper level as R_min'
  Delete R_min from DS_Y
   Form a new V_Y = [R_1, R_2, ..., R_min', ... R_{N_Y}]
   Take the new V_Y and calculate P_BS according to Equation 5.19
End While
```

Figure 7-21 Pseudo Code of the Data Rate Adaptation Scheme for UTRAN Downlink

The scheme extracts a subset DS_Y from V_Y. The subset DS_Y includes the data rates of the services which are candidate for an increase in data rate. In each round, the scheme will hypothetically increase the data rate of each candidate service belonging to DS_Y by one level and calculate the base station power $P_{BS}^{'}$. If $P_{BS}^{'}$ is unfeasible, the former data rate increase process will be cancelled and this candidate service will be deleted from DS_Y. If $P_{BS}^{'}$ is feasible, the ratio of the difference of the base station power, $P_{BS}^{'} - P_{BS}$, to the difference of the objective function, *OF'–OF*, is calculated and denoted as $RT(R_i)$, where R_i is the data rate of the *i*th service. The candidate service whose hypothetical data rate upgrade results in the smallest value of *RT* will be selected for an actual rate upgrade. The selected data rate is denoted as $R_{\min}$. Then, $R_{\min}$ will be deleted from DS_Y and the data rate vector V_Y is updated with $R'_{\min}$ and the next round of adaptation will proceed until no more services can have their data rate increased without producing an unfeasible P_{BS}.

The adaptation scheme aims to minimise the increase of power consumption and maximise the improvement of user satisfaction.

7.3.3.2 Adaptation in a UTRAN — Uplink

Given a data rate vector V_Y, the total interference received by the base station, I_{BS}, can be obtained by implementing the iterative algorithm presented in subsection 5.1.1. If the noise rise introduced by the total interference is smaller than the interference margin, the UTRAN cell has extra resources and the data rate adaptation scheme will be carried out to increase the data rate of certain users so as to improve the user satisfaction. The pseudo code of the scheme is presented in Figure 7-22.

When the UTRAN cell has extra resources, the adaptation scheme extracts a subset DS_Y from V_Y. Similar to the scheme for the downlink, the subset DS_Y includes the data rates of the services which are the candidates to have their data rates increased. In each iteration, the scheme will hypothetically increase the data rate of each candidate service belonging to DS_Y by one level and calculate the total interference received by the base station I'_{BS}. If I'_{BS} is unfeasible, the former data rate increase process will be cancelled and this candidate service will be deleted from DS_Y. If I'_{BS} is feasible, the ratio of the difference of the total interference received at the base station, $I'_{BS} - I_{BS}$, to the difference of the objective function, $OF' - OF$, is calculated and denoted as $RT(R_i)$. The candidate service whose hypothetical data rate ($R_{\min}$) upgrade results in the smallest value of *RT* will be selected for an actual rate upgrade. Then, $R_{\min}$ will be deleted from DS_Y and the data rate vector V_Y is updated with $R'_{\min}$

and the next round of adaptation will proceed until no more services can be improved without producing an unfeasible I_{BS}.

```
Given data rate vector V_Y = [R_1, R_2, ..., R_{N_Y}], calculate the Objective Function OF and I_BS
Extract the subset DS_Y = [R_1, R_2, ..., R_{i,...}, R_{m,...}]
While DS_Y is not empty
  For each user i in DS_Y
     Increase the data rate R_i to a upper level as R_i'
      Form a vector V_Y' = [R_1, R_2, ..., R_i', ..., R_{N_Y}], calculate I_BS'
      If the noise rise introduced by I_BS' is smaller than or equal to the interference margin
        Calculate OF' and the ratio RT(R_i):
                 RT(R_i) = (I_BS' - I_BS) / (OF' - OF)
      Else
         Delete R_i from DS_Y
      End If
  End For
   Among the calculated RTs, select the service whose data rate (R_min) supplies the smallest
   RT value: R_min = arg min RT(R_i)
  Increase the data rate R_min to a upper level as R_min'
  Delete R_min from DS_Y
   Form a new V_Y = [R_1, R_2, ..., R_min', ... R_{N_Y}]
   Take the new V_Y and calculate I_BS
End While
```

Figure 7-22 Pseudo Code of the Data Rate Adaptation Scheme for UTRAN Uplink

The adaptation scheme aims to minimise the increase of interference and maximise the improvement of user satisfaction.

7.3.3.3 Adaptation in an IEEE802.11b based WLAN

Given a data rate vector V_Y and service characteristics, the expected number of contending packets e_{ncp} can be obtained. When the service types are hybrid, if V_Y leads to a e_{ncp} value which is smaller than 1, the data rate adaptation scheme will selectively increase the data rates of certain users so as to effectively consume the extra network resources and improve user satisfaction. The pseudo code of the scheme is presented in Figure 7-23.

```
Given data rate vector V_Y = [R_1, R_2, ..., R_{N_Y}], calculate the Objective Function OF and e_ncp
Extract subset DS_Y = [R_1, R_2, ..., R_{i,...}, R_{m,...}]
While DS_Y is not empty
  For each user i in DS_Y
     Increase the data rate R_i to a upper level as R_i'
      Form a vector V_Y' = [R_1, R_2, ..., R_i', ..., R_{N_Y}], calculate e_ncp'
      If e_ncp' ≤ 1
         Calculate OF' and the ratio RT(R_i):
                    RT(R_i) = (e_ncp' - e_ncp) / (OF' - OF)
      Else
         Delete R_i from DS_Y
      End If
  End For
   Among the calculated RTs, select the service whose data rate
  (R_min) supplies the smallest RT value: R_min = arg min RT(R_i)
  Increase the data rate R_min to a upper level as R_min'
  Delete R_min from DS_Y
   Form a new V_Y = [R_1, R_2, ..., R_min', ...R_{N_Y}]
   Take V_Y and service characteristics, calculate e_ncp
End While
```

Figure 7-23 Pseudo Code of the Data Rate Adaptation Scheme for WLAN

When the WLAN has extra resources, the adaptation scheme extracts a subset DS_Y from V_Y. The subset DS_Y includes the data rates of the candidate services for an increase in the data rate. In each iteration, the scheme will hypothetically increase the data rate of each candidate service belonging to DS_Y by one level and calculate the value of e'_{ncp}. If the value of e'_{ncp} is greater than 1, the former data rate increase process will be cancelled and this candidate service will be deleted from DS_Y. If the value of the ratio of e'_{ncp} is smaller than or equal to 1, the ratio of the difference of expected number of contending packets, $e'_{ncp} - e_{ncp}$, to the difference of the objective function, $OF' - OF$ is calculated and denoted as $RT(R_i)$, where R_i is the data rate of service *i*. The service whose hypothetical data rate degradation results in the smallest value of RT will be selected for an actual degradation. The selected data rate is denoted as $R_{\min}$. After upgrade, a new value, $R'_{\min}$, is obtained. Then, $R_{\min}$ will be deleted from DS_Y and the data rate vector V_Y will be updated with $R'_{\min}$. The next round of adaptation will proceed until no more services can be improved without producing an unfeasible e_{ncp}.

The adaptation scheme aims to minimise the increase of channel contention and maximise user satisfaction.

7.3.4 Simulations and Results

In order to compare the performance obtained from the different RAN selection and optimisation algorithms, call level simulations are carried out. A heterogeneous communication environment is simulated with two RANs, a UTRAN cell and an IEEE802.11b based WLAN, and no background traffic is assumed. The simulation environment is identical to the simulation environment presented in section 7.1. All the users are covered by both the UTRAN cell and the WLAN with acceptable radio channel conditions. The objective is to gauge the performance of the P-RASO algorithm and compare with the algorithm MUSE-VDA presented in [ZM06]. In the simulation, the P-RASO algorithm implements two policies, the OUSI policy and the UNI policy. The P-RASO algorithm changes to different policies according to the example as shown in Figure 7-20.

In the simulations, six types of services are considered and their abbreviations are shown below:

- SP: Speech
- VC: Video Call
- VS: Video Streaming
- AS: Audio Streaming
- WB: Web Browsing

- FT: File Transfer

The first four services are real-time and UDP based. The others are non-real-time and TCP based. The services can belong to two service classes: *basic* and *premium*. The proportions of users Speech, Video Call, Video Streaming, Audio Streaming, Web Browsing and File Transfer are 12.5%, 12.5%, 12.5%, 12.5%, 25%, and 25%, respectively. The general service parameters and their typical values are listed in Table 7.4. For the UDP based real-time services, the data rate values are fixed. For the TCP based non-real-time services, the data rate values are the minimum requirements for each service class.

Service Type	SP	VC	VS	AS	WB	FT
Required Data Rate	6.7/12.2	64/128	64/128	32/64	128/256	64/128
Transport Protocol	UDP	UDP	UDP	UDP	TCP	TCP
Proportions of Users	12.5%	12.5%	12.5%	12.5%	25%	25%

Table 7.4 Service Parameters and Typical Values

Different from the simulations carried out in sections 7.1 and 7.2, one more real-time service is considered: the speech service. The specific configuration and parameters of the speech service in the UTRAN cell are presented in Table 7.5. The parameters include the activity factor and the target value of the E_b/N_o in the uplink and the downlink, which are derived based on [LWN05] and [HT07].

Service Type	**Required Data Rate**	**Activity Factor**	**Uplink E_b/N_o**	**Downlink E_b/N_o**
Speech	6.7/12.2	0.67	6 dB/5 dB	11 dB/10 dB

Table 7.5 Service Parameters and Values for the Speech Service in the UTRAN Cell [LWN05][HT07]

The specific configuration and parameters of the speech service in the WLAN include the size of the packet payload, the number of packets per second, and the packet inter arrival time. The values of the above parameters are not shown here but they have been presented in section 5.2.

In the simulations, the user arrival process is modelled based on a *Poisson* process and the user arrival rate is denoted as λ. The service duration is modelled based on an *Exponential* distribution. The mean value of the service duration is assumed to be three minutes and denoted as $1/\mu$, where μ is the user departure rate. Therefore, the network traffic load can be defined as the ratio of the user arrival rate to the user departure rate, λ/μ, which is denoted as ρ. The performance of the MUSE-VDA and P-RASO algorithms are evaluated and compared in three aspects: the data rate perceived by the users, the value of objective function, and the request blocking rate, with the increasing network traffic load.

The simulation results are presented in Figures 7-24, 7-25, and 7-26. Figure 7-24 compares the collective data rates perceived by all users when different RAN selection algorithms are used.

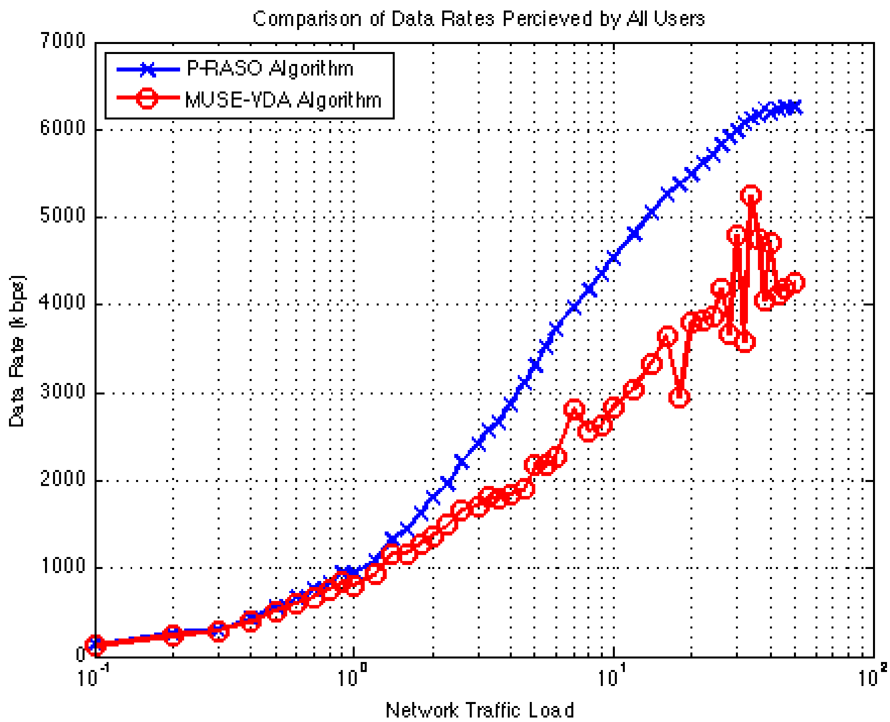


Figure 7-24 Comparison of the Data Rates Perceived by All Users

The simulation results plotted in Figure 7-24 demonstrate that, as the network traffic load (ρ) increases, the data rate introduced by both RAN selection algorithms also increases. Before ρ increases to 1.6, the data rates allocated by both algorithms are very close. When ρ becomes greater than 1.6, the data rate allocated based on the P-

RASO algorithm apparently and stably begins to exceed the data rate allocated by the MUSE-VDA algorithm. Finally P-RASO allocates around 6300 kbps when the value of ρ achieves 50. When MUSE-VDA is used, the data rate perceived by the users fluctuates acutely several times as the value of ρ is around 7.5, 18 and 30, and reaches at the end of the simulation around 4200 kbps.

Figure 7-25 compares the collective objective function values provided by both of the UTRAN cell and the WLAN with the increasing network traffic load when different RAN selection algorithms are used.

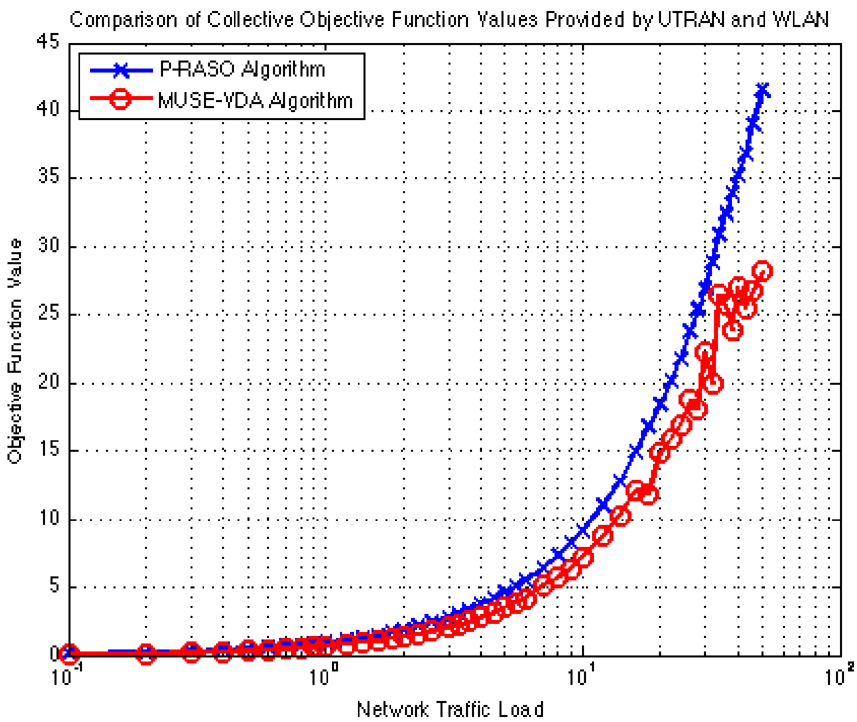


Figure 7-25 Comparison of Collective Objective Function Values provided by the UTRAN Cell and the WLAN

The simulation results presented in Figure 7-25 illustrate that, as the network traffic load builds up, the objective function values provided by the UTRAN cell and the WLAN based on both RAN selection algorithms also increase. The greater is the value of the objective function, the higher is the level of user satisfaction. When ρ increases beyond 6, the objective function value generated by the P-RASO algorithm apparently and stably begins to exceed the objective function value of MUSE-VDA and smoothly reaches about 42 when the value of ρ is 50. When MUSE-VDA is used, the objective function value fluctuates acutely two times as the value of ρ is around 30 and 40, and finally reaches about 28.

Figure 7-26 compares the request blocking rate with the increasing network traffic load when different RAN selection algorithms are used.

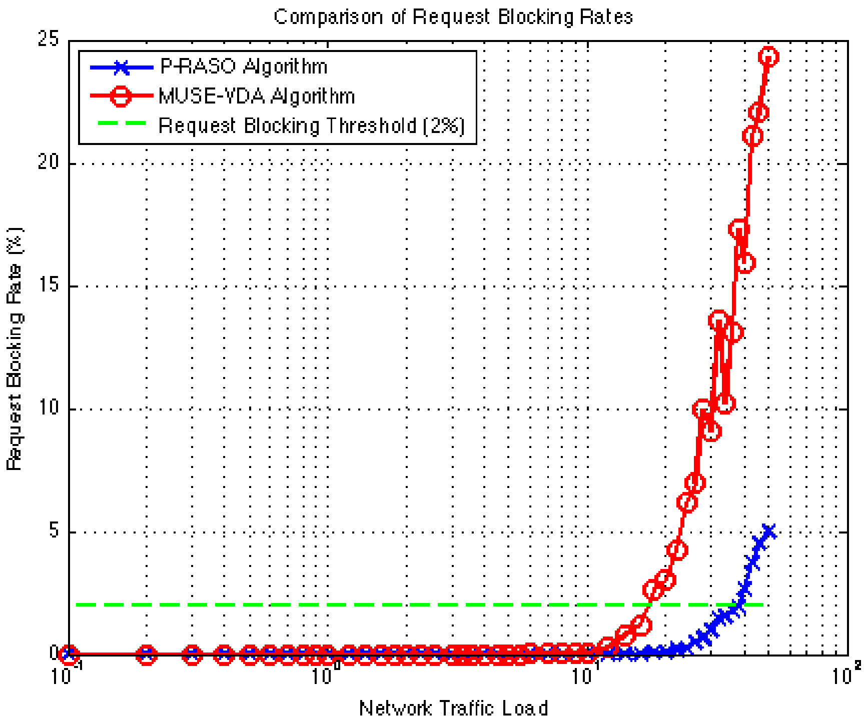


Figure 7-26 Comparison of Request Blocking Rates

In the simulation, the value of the request blocking threshold is assumed to be 2%. When MUSE-VDA is used, the request blocking rate starts to surge as ρ builds up to 12 and gets over 2% as ρ achieves 20. When P-RASO is implemented, the request blocking rate starts to grow as ρ is greater than 26 and exceeds 2% when ρ is beyond 40. Finally, the request blocking rate of MUSE-VDA rapidly increases to 24% and the request blocking rate of P-RASO increases to 5% when ρ is 50.

The above simulation results show that the P-RASO algorithm can provide better performance than the MUSE-VDA algorithm. The use of the P-RASO algorithm allows RAN to assign higher data rate and increases the level of user satisfaction with a lower request blocking rate. This is because the P-RASO algorithm focuses on the activities of user arrivals and departures, as well as the corresponding variation in the state of network resources. By dynamically implementing multiple policies and data rate adaptation schemes, service classes of the existing users can be appropriately adjusted to use the network resources to carry more traffic and improve user satisfaction in an effective way. By observing the simulation results of MUSE-VDA, severe fluctuations can be found in the values of data rate, objective function, and request blocking rate. These fluctuations are caused by lack of service adaptation scheme. Without a service adaptation scheme, the network resources cannot be adequately utilised to support the increased traffic load and improve the overall user satisfaction.

7.4 Concluding Remarks

In this chapter, several algorithms are proposed to facilitate an adaptive and efficient RAN selection. Section 7.1 presents the MIS algorithm and the MISSA algorithm, which constitute a simple and preliminary solution for a RAN selection algorithm. The service adaptation process performed in MISSA is inspirational and provides a direction for adjusting and optimising the utilisation of network resources. The RAO algorithm presented in section 7.2 implements an objective function to assess the perceived user satisfaction from each connecting candidate network, and selects the network which is able to provide the greatest user satisfaction. The RAO algorithm also improves the service adaptation process by applying the objective function to find out the best balance between the increased carried traffic and the user satisfaction. Section 7.3 presents an adaptive P-RASO algorithm which works with multiple policies. The P-RASO algorithm is able to cope with different scenarios, network conditions, and aims of operators by implementing different policies in a dynamic way. The P-RASO algorithm also enhances the service adaptation process to utilise the extra network resources for improving the user satisfaction in an effective and adequate way.

Chapter 8 Discussion and Conclusion

8.1 Discussion

This thesis presents a research that addresses the challenges in radio access network selection and optimisation of heterogeneous communication environments. The major contributions of this research include the network resource availability evaluation models, the architectural design of the decision-making system, and the adaptive selection and optimisation algorithms.

This thesis proposes network resource availability evaluation models for two radio access networks: UTRAN and IEEE802.11a/b based WLAN. The proposed evaluation models enable the implementation of call level simulations for the comparison and evaluation of the performance of the RAN selection and optimisation algorithms. Similar proposals were also presented in the literature. In [ZM06], besides the Multiservice Vertical Handoff Decision Algorithm, Zhu et al. presented the simulation settings where the available resources of a RAN were viewed as the bandwidth of the RAN. For example, assuming the bandwidth of a WLAN is B, after admitting a requesting user whose required data rate is d, the bandwidth and the remaining available resources in the WLAN are considered to be $B - d$. Such settings did not take into account the principle of the RAT. Although the simulations and the design of the algorithm were simplified by these assumptions, the effectiveness and the feasibility of the algorithm may be compromised. In contrast, the network resource availability evaluation models proposed in this thesis are based on the physical principles of the considered RANs. The UTRAN model studies the principle of resource utilisation in CDMA-based cellular networks and a new resource availability evaluation method is proposed. This evaluation method extends the work in [HT07] and implements a cell wrap-around model so as to avoid artefacts caused by boundary limitations. The WLAN model studies the principle of the CSMA/CA mechanism employed in IEEE802.11a/b standard and a novel resource availability evaluation method is proposed. The evaluation method defines a new parameter, the *expected number of contending packets,* and combines it with previous research on WLAN performance in order to measure the resource availability of a network for admitting or not real-time and non-real-time service requests. The proposed network resource availability evaluation models avoid employing assumptions which are too stringent abstractions of real network scenarios and provide a more realistic and efficient basis for the research of RAN selection and optimisation algorithms.

This thesis presents a new RAN selection and optimisation system which is compatible with the network architecture of future mobile communication systems. Different from other research, such as [ZM06] where the authors only presented an abstract and general architecture, the proposed RAN selection and optimisation system includes a comprehensive architectural design with extensive description of different functional entities, as well as the signalling and message exchange between them. The system consists of four types of functional entities: the Network Assisting Entity, the Terminal Assisting Entity, the Operation Entity, and the Database Entity. The Operation Entity and the Database Entity are placed in the IP backbone network and constitute the Cooperation Administrator. The Cooperation Administrator is responsible for the RAN selection and optimisation functionalities and it communicates with the Network Assisting Entity and the Terminal Assisting Entity for exchanging context information and decision commands. For reliability and scalability reasons, the proposed RAN selection and optimisation system may consist of more than one Cooperation Administrator and each Cooperation Administrator covers a certain area and cooperates with a limited number of RANs. Considering that the RAN selection and optimisation process can only be activated in five situations, including user service request initiation, remote connection initiation, handover, service/resource reconfiguration, and policy alternation, the proposed system also allows the legacy radio resource management entities in each wireless system to remain functioning. This measure can preserve the operator's previous investments and also it can help to reduce the unnecessary burden on the selection and optimisation system. Comparing with the related research, such as the Composite Radio Environment [DKKTOSPTVE-K03][DVE-KT04] where the decision-making entities were distributed in each RAN, the architecture of the proposed system presents a centralised approach for conducting RAN selection and optimisation without sacrificing the existing system specific intelligence. Such centralised approach provides advantages such as better understanding of the network and the user status, reduced signalling traffic load, efficient and effective operation, etc.

This thesis also proposes novel adaptive algorithms for RAN selection and optimisation. Four algorithms are proposed and evaluated: the Multiple Interface Selection algorithm, the Multiple Interface Selection and Service Adaptation algorithm, the Radio Access Optimisation algorithm, and the Policy-based Radio Access Selection and Optimisation algorithm. Although the Multiple Interface Selection algorithm and the Multiple Interface Selection and Service Adaptive algorithm are simple and preliminary, by observing the simulation results, they indicate that service adaptation is an encouraging direction for designing RAN selection algorithms and optimising the utilisation of network resources. Furthermore, if the main focus is only to serve the requesting user without considering the existing users and the utilisation of the network resources (similar to the conventional decision-making algorithms presented in sections 3.1, 3.2, and 3.3), the network selection decision might only be able to benefit a limited number of users, while the satisfaction levels of some users could be disturbed and overlooked. Also the network resources could be consumed in an ineffective way. This is because, the admission of a new user would inevitably affect the network resource status of the selected RAN and might impact the interests of the existing users.

The Radio Access Optimisation algorithm is based on an objective function to select a network and improve the service adaptation process. The proposed objective function measures the overall user satisfaction obtained from a candidate network and the Radio Access Optimisation algorithm selects the network which in overall can provide the greatest value for that objective function. When the network resources are insufficient to admit a new service request, the Radio Access Optimisation algorithm also carries out a data rate adaptation scheme which will attempt to decrease the data rate of selected users in order to obtain sufficient network resources. Such adaptation scheme aims to discover the best balance between the increased carried traffic and the user satisfaction and employs a heuristic method to simplify the complexity of the rate adaptation process. The simulation results demonstrate that the Radio Access Optimisation algorithm not only effectively utilises the network resources but also improves overall user satisfaction.

However, algorithms employing a single network selection and optimisation strategy or policy are not well fitted to cope with the ever-changing radio environment, traffic load, and network resource availability. A novel algorithm is required. The algorithm should be able to manage multiple policies and adaptively implement an appropriate policy according to the current scenario and network conditions. Therefore, this thesis proposes a Policy-based Radio Access Selection and Optimisation algorithm. The Policy-based Radio Access Selection and Optimisation algorithm implements multiple policies to deal with different scenarios, network conditions and aims of operators in an adaptive and dynamic way. Different policies have different considerations towards the definition of the gains obtained from a candidate network. When a policy is implemented, the corresponding gain is taken into account and the algorithm selects the network which can provide the greatest gain. In this thesis, the gains are represented by the values obtained from objective functions and the aim of each policy is to maximise the corresponding objective function. Furthermore, as the network resource availability changes, a data rate adaptation scheme may be executed by the active policy in order to optimise the network resource utilisation. Such adaptation scheme extents the Radio Access Optimisation algorithm to deal with the situation where extra resources become available. This thesis presents a scenario where the Policy-based Radio Access Selection and Optimisation algorithm has two different policies and the implementation of these two policies alternates when a predefined condition is triggered. The simulation results demonstrate that the proposed algorithm can effectively utilise the network resources and also provide high level of user satisfaction.

The RAN selection and optimisation system and algorithms proposed in this thesis present a centralised approach for conducting RAN selection and optimisation. However, this **does not** mean that the proposed system and algorithms can only be implemented in scenarios where services are coupled with a single network operator or where there is no competition between different network operators. In [OMM07], Ormond et al. foresee that the success of future wireless networks lies in offering a positive user experience, which results in a *Service-Oriented Heterogeneous Wireless Network Environment* (SOHWNE). In SOHWNE, service provision and delivery mechanism are no longer coupled together and service providers are placed into the wireless market as a third party. The service providers are responsible for producing novel services to satisfy the users' needs for communication, entertainment, and information. The decoupling of service provision and delivery mechanism also changes the current business model in the wireless market and one user no longer has to be bound with all network operators for ubiquitously obtaining the required service. Instead, different network operators compete to provide network access to the user, and the user can select the RAN which can best deliver the required service. The RAN selection and optimisation system proposed in this thesis also could be implemented in a SOHWNE. A network operator could implement the proposed system architecture to realise interworking and cooperation between its RANs. According to the requested service, the network operator's interest, and network conditions, the network operator could employ an appropriate policy in the selection and optimisation system in order to optimise the resource utilisation in its RANs and provide a list of candidate RANs (with supported QoS) to the requesting user to choose from. The proposed RAN selection and optimisation system not only considers the interests of the network operators but also helps to relax the complexity in network selection at the user side.

8.2 Conclusion

The overlapping of different wireless network technologies creates heterogeneous communication environments. Future mobile communication systems consider the technological and operational services of heterogeneous communication environments and enable a universal access via different wireless or even wired technologies. This foreseen universal access facilitates service convergence, joint resource management and adaptive quality of service but also brings in the challenges for RAN selection and optimisation.

The aim of this research was to address the above challenges and propose a novel context-aware RAN selection and optimisation system. This work investigated the system internetworking architectures supporting future mobile communication systems and conducted a feasibility study for performing context-aware operations. It also examined and summarised the algorithms employed in the selection process. Based on the above efforts, this thesis proposed an architectural and signalling design for the RAN selection and optimisation system. Although the decision-making in the proposed system architecture is centralised and network based, the distributed intelligence, such as the legacy radio resource management entities in each wireless system, still remains functioning. This work also proposed new network resource availability evaluation models for UTRAN and IEEE802.11a/b based WLAN. The proposed network resource availability evaluation models establish a more realistic and efficient basis for the RAN selection and optimisation research. Finally, this thesis proposed four different selection algorithms and evaluated their performances over user perceived data rates, objective function value, and blocking rates. The proposed Policy-based Radio Access Selection and Optimisation algorithm demonstrated its ability to change to different policies when different conditions are triggered. The simulation results show that, comparing with conventional utility based algorithms, this algorithm utilises the network resources more effectively and reduces the service blocking rate. This algorithm offers an adaptive and efficient solution for RAN selection and optimisation.

Overall, this thesis presents a systematic research for addressing the challenges in RAN selection and optimisation in heterogeneous communication environments. The proposed solutions are feasible and successful in achieving the aims of this research.

8.3 Future Work

In the thesis, the architectural and signalling design for the RAN selection and optimisation system was presented in chapter 4. Further simulations and test beds can be implemented to evaluate the design and optimise the performance of the proposed system architecture and the signalling scheme.

In this thesis, the network resource availability evaluation models considered two types of RANs: UTRAN and IEEE802.11a/b based WLAN. These two RANs constitute a basic heterogeneous communication environment. Nowadays, a variety of wireless broadband access technologies have been proposed and deployed to the communication industry, such as HSPA, WiMAX, IEEE802.11n based WLAN, etc. New resource availability evaluation models should be developed to simulate these networks and reflect a more diversified heterogeneous communication environment.

This thesis proposes RAN selection and optimisation based on heuristic algorithms. In future work, further study on the optimisation of resource utilisation could be carried out to propose a real-time, efficient, and effective algorithm. This thesis also uses an objective function to represent the user satisfaction level. Personalisation is the key to the success of future wireless networks. Therefore, a further study should be carried out on the representation of user's requirements and the application of user profile in RAN selection. In the simulations of chapter 7, it was assumed that all the users were covered by both RANs with acceptable radio channel conditions. Future work could consider more complex scenarios, such as user mobility and limited radio channel access, how these new complexities affect the selection and optimisation algorithms and the policies, and how the selection could be optimised to operate in this more realistic and complex environment.

Reference

[AIPN05] 3GPP TR 22.978 v7.1.0 (2005-06) Technical Report. “All-IP Network (AIPN) feasibility study (release 7)”

[B3G08] Beyond 3G – ultra-fast mobile radio networks of the future, http://cordis.europa.eu/ictresults/index.cfm/section/news/tpl/article/BrowsingType/Features/ID/89945

[BCB05] Bruno, R.; Conti, M.; Bregori, E.; “Modeling TCP Throughput over Wireless LANs”; Proc. 17th IMACS World Congress Scientific Computation, Applied Mathematics and Simulation, 2005

[BCB06] Bruno, R.; Conti, M.; Bregori, E.; “Performance Modelling and Measurements of TCP Transfer Throughput in 802.11-based WLAN”; Proceedings of the 9th ACM International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems, 2006, pages: 4 – 11

[BCB08] Bruno, R.; Conti, M.; Bregori, E.; “Throughput Analysis and Measurements in IEEE 802.11 WLANs with TCP and UDP Traffic Flows”; IEEE Transactions on Mobile Computing, vol. 7, no. 2, pages: 171 – 186, 2008

[BCB09] Bruno, R.; Conti, M.; Bregori, E.; “Average-Value Analysis of 802.11 WLANs with Persistent TCP Flows”; IEEE Communications Letters, vol. 13, issue 4, pages: 218 – 220, 2009

[BI03] Balasubramaniam, S.; Indulska, J.; “Handovers between heterogeneous networks in pervasive systems”; Communication Technology Proceedings, 2003. ICCT 2003. International Conference on Volume 2, 9-11 April 2003, Page(s): 1056 - 1059 vol.2

[Bia00] Bianchi, G.; “Performance Analysis of the IEEE 802.11 Distributed Coordination Function”; Selected Areas in

Communications. IEEE Journal on. Volume 18, Issue 3 (March 2000), 535-547

[BSMPR07] Bosch, P.; Samuel, L.; Mullender, S.; Polakos P.; and Rittenhouse, G.; “Flat Cellular (UMTS) Networks”; IEEE Wireless Communications & Networking Conference, 2007. WCNC 2007 (March 2007), 3861-3866

[CCPP07] The World Wide Web Consortium (W3C); CC/PP Information Page; http://www.w3c.org/Mobile/CCPP 2007

[CG00] Cali, F.; Conti, M.; Gregori, E.; “Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit”; Networking, IEEE/ACM Transactions on, Volume 8, Issue 6 (Dec. 2000), 785-799

[Che03] Chen, Y.; “Soft Handover Issues in Radio Resource Management for 3G WVDMA Networks”; Ph.D. Thesis, Queen Mary, University of London, 2003.

[CHH92] Shu-Jen Chen; Ching-Lai Hwang and Frank P. Hwang; “A New Approach to Fuzzy MADM Problems”, in: **Fuzzy Multiple Attribute Decision Making method and applications**, Springer-Verlag, 465-486 (1992).

[CK00] Chen, G.; Kotz, D.; “A survey of Context-Aware Mobile Computing Research”; Dratmouth Computer Science Technical Report TR2000-381, Department of Computer Science, Dartmouth College

[CLD07] Mainak Chatterjee, Haitao Lin, Sajal K. Das, “Rate Allocation and Admission Control for Differentiated Services in CDMA Data Networks”, IEEE Transactions on Moblie Computing, vol. 6, NO. 2, February 2007

[CLH04] Chen, W.; Liu, J.; Huang, H.; “An adaptive scheme for vertical handoff in wireless overlay networks”; Parallel and Distributed Systems, 2004. ICPADS 2004. Proceedings.

Tenth International Conference on 7-9 July 2004, Page(s): 541 – 548

[CRZO04] Carlton, A.; Rudolf, M.; Zuniga, J. and Olvera, U.; “MEDIA INDEPENDENT HANDOVER Functions and Service Specification”; http://www.ieee802.org/21/

[DA99] Dey, A.; Abowd, G.; “Towards a Better Understanding of context and context-awareness”; Technical Report GIT-GVU-99-22, Georgia Institute of Technology, College of Computing, June 1999

[Der03] R. Thomas Derryberry; 3GPP2 Technical Specification Group C. 3GPP2/TSG-C.R1002. 1xEV-DV Evaluation Methodology (V13) (2003)

[DGKLMSS01] Dinda, P.A.; Gross, T.; Karrer, R.; Lowekamp, B.; Miller, N.; Steenkiste, P.; Sutherland, D.; “The Architecture of the Remos System”; Proceeding of 10th IEEE International Symposium on High Performance Distributed Computing, 2001, Page(s): 252 – 265.

[Din06] Dinda, P.A.; “Design, Implementation, and Performance of an Extensible Toolkit for Resource Prediction in Distributed Systems”; IEEE Transactions on Parallel and Distributed Systems, Vol. 17, Issue 2, Feb. 2006, Page(s): 160 – 173.

[DKKTOSPTVE-K03] Demestichas, P.; Koutsouris, N.; Koundourakis, G.; Tsagkaris, K.; Oikonomou, A.; Stavroulaki, V.; Papadopoulou, L.; Theologou, M.E.; Vivier, G.; El-Khazen, K.; “Management of Networks and Services in a Composite Radio Context”; Wireless Communications, IEEE [see also IEEE Personal Communications] Volume 10, Issue 4, Aug. 2003, Page(s): 44 – 51

[DVE-KT04] Demestichas, P.; Vivier, G.; El-Khazen, K.; Theologou, M.; “Evolution in Wireless Systems Management Concept: From Composite Radio Environment to Reconfigurability”;

Communications Magazine, IEEE Volume: 42 Issue: 5, May 2004, Page(s): 90- 98

[E-KLGGB03] El-Khazen, K.; Lefevre, F.; Garrec, D.; Guiraudou, M.; Benali, O.; "Beyond 3G Demonstrator for Enhanced Service Signaling, Discovery and Management"; Computers and Communication, 2003. (ISCC 2003). Proceedings. Eighth IEEE International Symposium on 2003, Page(s): 1367 - 1372 vol.2

[ETSI98] ETSI TR 101 112 v3.2.0 (1998-04) Technical Report. Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of radio transmission technologies of UMTS (UMTS 30.03 version 3.2.0)

[ETSI01] ETSI Tech. rep. ETSI TR 101 957. Requirements and Architectures for Interworking between HIPERLAN/3 and 3rd Generation cellular Systems. 2001

[EUTRAN08] 3GPP TR 25.912 v8.0.0 (2008-12) Technical Report. "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Feasiblility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN) (Release 8)"

[GAN09] 3GPP TS 43.318 v8.4.0 (2009-02) Technical Specification. "3rd Generation Partnership Project; Technical Specification Group GSM/EDGE; Radio Access Network; Generic Access Network (GAN); Stage 2 (Release 8)"

[GAP-RSR08] Lorenza Giupponi; Ramon Agustí; Jordi Pérez-Romero; Oriol Sallent Roig; "A Novel Approach for Joint Radio Resource Management Based on Fuzzy Neural Methodology"; IEEE Transactions on Vehicular Technology, Vol. 57, Issue 3, 2008, Page(s): 1789 – 1805

[GLN09] Yanying Gu; Anthony Lo; Ignas Niemegeers; "A Survey of Indoor Positioning Systems for Wireless Personal

Networks"; IEE Communications Surveys and Tutorials. Vol. 11, No. 1, First Quarter 2009, Page(s): 13 – 32

[GY00] Garg, V. K.; Yu, O.T.W.; "Integrated QoS Support in 3G UMTS Networks"; Wireless Communications and Networking Conference, 2000. WCNC. 2000 IEEE Volume 3, 23-28 Sept. 2000 Page(s):1187 - 1192 vol.3

[HT07] Edited by Harri Holma and Antti Toskala; WCDMA for UMTS – HSPA Evolution and LTE; Wiley 2007

[IMMH03] Inoue, M.; Mahmud, K.; Murakami, H.; Hasegawa, M.; "MIRAI: A Solution to Seamless Access in Heterogeneous Wireless Network"; Communications, 2003. ICC '03. IEEE International Conference on Volume 2, 11-15 May 2003 Page(s): 1033 - 1037 vol.2

[IMMHM04] Inoue, M.; Mahmud, K.; Murakami, H.; Hasegawa, M.; Morikawa, H.; "Design and Implementation of Out-of-Band Signaling for Seamless Handover in Wireless Overlay Networks"; Communications, 2004 IEEE International Conference on Volume 7, 20-24 June 2004, Page(s): 3932 - 3936 Vol.7

[IMT-ADV08] International Telecommunication Union Radiocommunication Study Group Document IMT-ADV/1-E. "Background on IMT-Advanced"

[IRRH03] Indulska, J.; Robinson, R.; Rakotoirainy, A.; Henricksen, K.; "Experiences in Using CC/PP in Context-Aware Systems"; In 4th International Conference on Mobile Data Management, MDM'03, January 2003 LNCS (2515), Pages: 247-261, Melbourne

[Joh04] Alan B. Johnston; "SIP understanding the Session Initiation Protocol (second edition)"; Artech House, Inc. 2004

[JZCCL06] Huiling Jia; Zhaoyang Zhang; Peng Cheng; Hsiao-Hwa Chen; Shiju Li; "Study on Network Selection for Next-

Generation heterogeneous Wireless Networks”; IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, 2006, Page(s): 1 -5

[KAT07] Koundourakis, G.; Axiotis, D. and Theologou, M.; “Network-based Access Selection in Composite Radio Environment”; IEEE Wireless Communications & Networking Conference, 2007. WCNC 2007 (March 2007), 3877-3883

[Kin07] Kineto Wireless, Inc; "Universal Mobile Access: UMA Expands Beyond Dual-Mode Handset"; http://www.umatoday.com

[Kin08] Andrew King; “Website Optimization: Speed, Search Engine & Conversion Rate Secrets”; O’Reilly Media, Inc. 2008

[Li03] Li, W.; “A Service Oriented SIP Infrastructure for Adaptive and Context-Aware Wireless Services”; 2nd international conference on Mobile and ubiquitous multimedia Norkkoping, Sweden, 10/12 2003; http://www.ep.liu.se/ecp/011/014/ecp011014.pdf

[LJC08] Jong Min Lee; Seung Hyun Jeon; Jun Kyum Choi; “Adaptive Network Access Mechanism for Multi-interfaced Terminal over Heterogeneous Wireless Networks”; 14th Asia-Pacific Conference on Communications, 2008, Page(s): 1 - 4

[LL91] Chin-Teng Lin; C. S. George Lee; “Neural-Network-Based Fuzzy Logic Control and Decision System”; IEEE Transactions on Computers, Vol. 40, No. 12, 1991, Page(s): 1320 – 1336

[LL02] Latvakoski, E.J.; Laurila, P.J.; “Application based access system selection concept for all IP mobile terminals”; Global Telecommunications Conference, 2002. GLOBECOM '02.

IEEE Volume 3, 17-21 Nov. 2002 Page(s): 2984 - 2988 vol.3

[LMSGSS98] Lowekamp, B.; Miller, N.; Sutherland, D.; Gross, T.; Steenkiste, P.; Subhlok, J.; "A Resource Query Interface for Network-Aware Applications"; High Performance Distributed Computing, 1998. Proceedings. The Seventh International Symposium on 28-31 July 1998, Page(s): 189 – 196

[LTE08] 3GPP; UTRA-UTRAN Long Term Evolution (LTE) and 3GPP System Architecture Evolution (SAE); ftp://ftp.3gpp.org/Inbox/2008_web_files/LTA_Paper.pdf

[LWN05] Jaana Laiho; Achim Wacker; Tomas Novosad; "Radio Network Planning and Optimisation for UMTS (second edition)"; Wiley Blackwell 2005

[LWZ07] Litao Liang; Hui Wang; ping Zhang; "Net Utility-Based Network Selection Scheme in CDMA Cellular/WLAN Integrated Networks"; IEEE Wireless Communications & Networking Conference, 2007. WCNC 2007 (March 2007), Page(s): 3315-3319

[M.1645] RECOMMENDATION ITU-R M.1645. "Framework and overall objective of the future development of IMT-2000 and system beyond IMT-2000"

[MIH09] IEEE Standard for Local and metropolitan area networks- Part 21: Media Independent Handover, http://ieeexplore.ieee.org/servlet/opac?punumber=4769363

[MS00] Miller, N.; Steenkiste, P.; "Collecting Network Status Information for Network-Aware Applications"; INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE Volume 2, 26-30 March 2000, Page(s): 641 - 650 vol.2

[MWIMM02] Mahmud, K.; Wu, G.; Inoue, M.; Murakami, H.; Mizuno, M.; "Basic Access Network – the Signaling-only Network for Power-Efficient Mobile Multi-service Terminals in MIRAI Architecture"; Vehicular Technology Conference, 2002. Proceedings. VTC 2002-Fall. 2002 IEEE 56th Volume: 4, 2002, Page(s): 2385- 2389 vol.4

[MZ04] McNair, J.; Zhu, F.; "Vertical handoffs in fourth-generation multinetwork environments"; Wireless Communications, IEEE [see also IEEE Personal Communications] Volume: 11, Issue: 3, June 2004, Pages: 8 – 15

[Neg05] Michael Negnevitsky; Artificial Intelligence – A Guide to Intelligent Systems (second edition); Addison-Wesley 2005

[NGMN] Next Generation Mobile Networks, http://www.ngmn.org

[NGMN06] Board of NGMN Limited; "Next Generation Mobile Networks Beyond HSPA & EVDO A White Paper"; http://www.ngmn.org

[OMM07] Olga Ormond; Gabriel-Miro Muntean and John Murphy; "Dynamic Network Selection in Wireless LAN/MAN Heterogeneous Networks", in: **Mobile WiMAX: towards broadband wireless metropolitan area networks**, Auerbach Publications, 233-271 (2007).

[PNW03] Prehofer, C.; Nafisi, N.; Wei, Q.; "A framework for context-aware handover decisions"; Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on, Volume: 3, 7-10 Sept. 2003, Pages: 2794 - 2798 vol.3

[RGC01] Veselin Rakocevic, John Griffiths, Grahan Cope, "Performance Analysis of bandwidth Allocation Schemes in Multiservice IP Networks using Utility Functions", http://www.staff.city.ac.uk/~veselin/publications/Rakocevic_ITC01.pdf (2001)

[Saaty00] Saaty, TL.; Fundamentals of Decision Making and Priority Theory with the Analysis Hierarchy Process; RWS Publications 2000

[SAE08] 3GPP TR 23.882 v8.0.0 (2008-09) Technical Report. “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3GPP System Architecture Evolution: Report on Technical Options and Conclusions (Release 8)”

[SDA99] Salber, D; Dey, A.K. and Abowd, G.D.; “The Context Toolkit: Aiding the development of context-enabled applications”; in Proceedings of CHI'99, pp. 434-441. http://citeseer.ist.psu.edu/salber99context.html

[Ser08] 3GPP TS 122 105 v8.4.0 (2008-01) Technical Specification. “Universal Mobile Telecommunications System (UMTS); Services and service capabilities (3GPP TS 22.105 version 8.4.0 Release 8)”

[SFP02] Salkintzis, A. K.; Fors, C.; Pazhyannur, R., “WLAN-GPRS integration for Next-Generation Mobile Data Networks”, IEEE Wireless Communications, October 2002; Page(s): 112-124.

[SJ01] Henry Sinnreich and Alan B. Johnston; “Internet Communications Using SIP”; Wiley 2001

[SJ05] Qingyang Song; Abbas Jamalipour; “An Adaptive Quality-of-Service Network Selection Mechanism for Heterogeneous Mobile networks”; Wireless Communications and Mobile Computing, 2005, Vol. 5, Issue 6, Page(s): 607 – 708

[SM07] Milena Stoyanova; Petri Mähönen; “Algorithmic Approaches for Vertical Handoff in Heterogeneous Wireless Environment”; Proceeding of the 2007 IEEE Wireless Communications and Networking Conference, WCNC 2007, Page(s): 3780 – 3785

[SW02] Swany, M.; Wolski, R.; “Multivariate Resource Performance Forecasting in the Network Weather Service”; Proceeding of ACM/IEEE 2002 Conference on Supercomputing, 2002

[SZ08] Wei Shen; Qing-An Zeng; “Cost-Function-Based Network Selection Strategy in Integrated Wireless and Mobile Networks”; IEEE Transactions on Vehicular Technology, Vol. 57, Issue 6, Page(s): 3778 - 3788

[TMM-DK07] Elias Tragos; Albena Mihovska; Emilio Mino-Diaz; Pantelis Karamolegkos; “Access Selection and Mobility Management in a Beyond 3G RAN: The WINNER approach”; Proceedings of the 5th ACM International Workshop on Mobility Management and Wireless Access, 2007, Page(s): 168 – 175

[Tro06] G.D. Troxel et al.; “Adaptive Dynamic Radio Open-source Intelligent Team (ADROIT): Cognitively-controlled Collaboration among SDR Nodes”; Networking Technologies for Software Defined Radio Networks, 2006, SDR’06, 1st IEEE Workshop on, 8-17, 2006

[UMA] Universal Mobile Access, http://www.umatoday.com

[Vit98] Andrew J. Viterbi; “CDMA Principle of Spread Spectrum Communication”; Addison-Wesley 1998

[Wen03] Wenger, S.; “H.264/AVC Over IP”; IEEE Transactions on Circuits and Systems for Video Technology, Vol 13, No. 7, July 2003

[WINNER] Wireless World Initiative New Radio, http://www.ist-winner.org

[WINNER4.4] IST-2003-507581 WINNER D4.4 v 1.0. “Impact of Cooperation Schemes between RANs – Final Deliverable”, Wireless World Initiative New Radio, http://www.ist-winner.org

[WINNER6.13.14] IST-4-027756 WINNER II D6.13.14 version 1.1. "WINNER II System Concept Description", Wireless World Initiative New Radio, http://www.ist-winner.org

[Wis01] Wisely, D.R.; "SIP and Conversational Internet Applications"; BT Technology Journal; Publisher: Springer Science+Business Media B.V., Formerly Kluwer Academic Publishers B.V.; ISSN: 1358-3948 (Paper) 1573-1995 (Online); DOI: 10.1023/A:1011355410629; Issue: Volume 19, Number 2; Date: April 2001; Page: 107-118

[WLAN07] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. http://ieeexplore.ieee.org/. June 12 2007

[WLL05] Wang, W.; Liew, S. and Li, O.; "Solutions to Performance Problems in VoIP Over a 802.11 Wireless LAN"; Vehicular Technology, IEEE Transaction on. Vol 54, Issue 1, 366-384 (2005).

[WSG06] T. Weingart; D. C. Sicker and D. Grunwald; "A Method for Dynamic Configuration of a Cognitive Radio"; Networking Technologies for Software Defined Radio Networks, 2006, SDR'06, 1st IEEE Workshop on, 93-100, 2006

[WSP97] Wolski, R.; Spring N. and Peterson, C.; "Implementing a performance forecasting system for metacomputing: The network weather service"; Tech.Rep.TR-CS97-540, University of Sa Diego, May 1997

[Wu02] Gang Wu; Mizuno, M.; Havinga, P.J.M.; "MIRAI Architecture for Heterogeneous Network"; Communications Magazine, IEEE Volume 40, Issue 2, Feb. 2002, Page(s): 126 – 134

[WZ98] Weissman, J. and Zhao, X.; "Scheduling Parallel Applications in Distributed Networks"; Cluster Computing, vol.1, no.1, pp.95-108, May 1998

[XML04] The World Wide Web Consortium (W3C); Extensible Markup language 1.1; www.w3.org/TR/2004/REC-xml11-20040204/

[XSC01] M. Xiao; N.B. Shroff; E.K.P. Chong; "Utility-Based Power Control in Cellular Wireless Systems"; Proceeding of IEEE INFOCOM, vol. 1, page(s): 412 – 421, 2001

[Zha04] Zhang, W.; "Handover decision using fuzzy MADM in heterogeneous networks"; Wireless Communications and Networking Conference, 2004. WCNC. 2004 IEEE Volume 2, 21-25 March 2004, Page(s): 653 - 658 Vol.2

[ZM06] Fang Zhu and Janise McNair; "Multiservice Vertical Handoff Decision Algorithm"; EURASIP Journal on Wireless Communications and Networking, Vol. 2006, Page(s): 1 – 13

Appendix A — Author's Publications

Luo, W. and Bodanese, E. L.; "Intelligent Radio Access Network Selection System for Heterogeneous Wireless Network Environments"; Proceedings of the 2007 Networking and Electronic Commerce Research Conference (NAEC 2007), pp. 246-264, October 18-21, 2007

Luo, W. and Bodanese, E. L.; "Optimizing Access in an Integrated Wireless Network Environment", Proceedings of the 2008 (SPECTS 2008), pp. 151-158, June 16-18, 2008

Luo, W. and Bodanese, E. L.; "Radio Access Network Selection in a Heterogeneous Communication Environment", Proceedings of the 2009 Wireless Communications and Networking Conference (WCNC 2009), April 5-8, 2009

Luo, W. and Bodanese, E. L.; "Optimising Radio Access in a Heterogeneous Wireless Network Environment", Proceedings of the 2009 International Communications Conference (ICC 2009), June 14-18, 2009

Luo, W. and Bodanese, E. L.; "Adaptive and Efficient Radio Access Selection and Optimisation in a Heterogeneous Communication Environment", Proceedings of the 21st International Teletraffic Congress (ITC 21), September 15-17, 2009

Appendix B — Basic Operation of IEEE802.11 Distributed Coordination Function

The IEEE 802.11 standard supports two MAC schemes, Distributed Coordination Function (DCF) and Point Coordination Function (PCF). PCF is a centralized mechanism which uses a central coordinator. The central coordinator polls the wireless stations and provides a contention free (CF) access to the channel. However, many commercial products do not implement the PCF scheme. On the other hand, the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) based on DCF is widely used for supporting asynchronous data transfer on a best effort basis and provides fairness among the wireless stations.

The basic operation of DCF can be described as follows. Before transmitting the first packet, the wireless station will monitor the channel activity. If the channel remains idle for a DCF Interframe Space (DIFS) period, the station will transmit the packet. If the channel is sensed as busy either immediately or during the DIFS period, the wireless station will keep monitoring the channel until it is idle for a DIFS period. Then, in order to minimize the probability of collision with the packets transmitted by other wireless stations, a random backoff time is generated. In order to avoid channel capture, the station also will wait a random backoff time between two consecutive packet transmissions, even though the channel is sensed idle for a DIFS interval.

The backoff time is slotted. The size of a time slot (shown in Table B.1) depends on the physical layer and accounts for the propagation delay. At each packet transmission, the wireless station randomly selects the number of time slots in the range from 0 to Contention Window (*CW*) – 1. The value of *CW* depends on the number of unsuccessful packet transmissions. At the first transmission attempt, the value of *CW* is equal to $CW_{\min}$. $CW_{\min}$ is the minimum contention window and its value is shown in Table A.1. After each unsuccessful packet transmission, the value of *CW* will be doubled up to the maximum value $CW_{\max}$.

After sensing the channel as idle for a DIFS period, the wireless station begins to decrease the backoff time counter. Before the counter reaches zero, if the channel is sensed as busy, the wireless station will pause the counter, and reactivate the counting down only when the channel is sensed idle for a DIFS period. When the backoff time counter reaches zero, the wireless station transmits the packet.

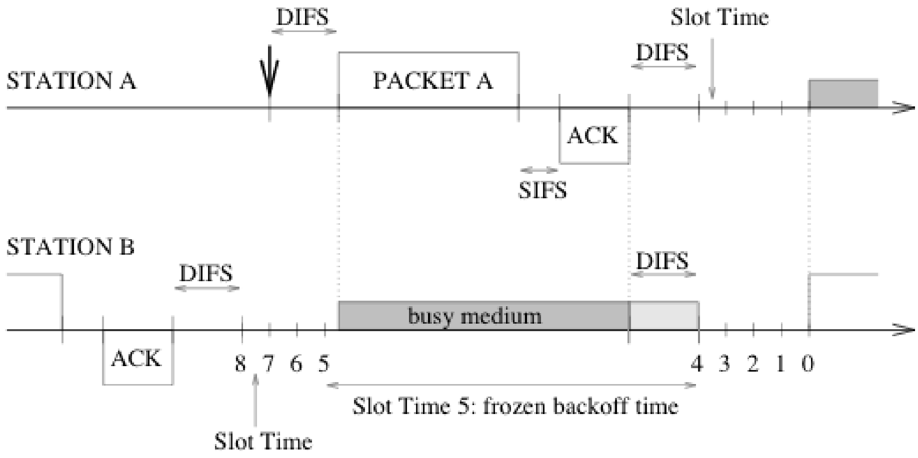


Figure B-1 Basic Operation of 802.11 DCF [Bia00]

Figure B-1 presents an example for the basic operation of 802.11 DCF. After transmitting a packet, station B waits for a DIFS and then, before transmitting the next packet, selects a backoff time comprised of eight time slots. Meanwhile, the first packet of station A arrives at the time indicated by the arrow shown in Figure A.1. Then, station A senses the channel as idle for a DIFS interval and transmits the packet. The packet transmission occurs in the middle of the time slot when the backoff value is five for station B. As a consequence, the backoff time counter is paused to the value of five and then reactivated again when the channel is sensed idle for a DIFS. Station B transmits the packet as the backoff time reaches zero.

In the wireless channel, the stations cannot detect a packet collision by hearing their own transmission. Therefore, after receiving a packet, the destination station waits for a Short Interframe Space (SIFS) period and then transmits an acknowledgement (ACK) back to the transmitting station to signal the successful packet reception. Because the SIFS period and the consequent propagation delay are shorter than the DIFS period, other wireless stations cannot sense the channel as idle for a DIFS period until the ACK is transmitted. If the transmitting station cannot receive the ACK within a specific timeout or it senses a different packet in the channel, it will consider the former packet transmission as unsuccessful and schedule a retransmission according to the backoff rules.

Parameter	Value
DIFS	50 μsec
SIFS	10 μsec
Slot Time	20 μsec
CW_{min}	32
CW_{max}	1023
Data Rate	1, 2, 5.5, 11 Mbps
Basic Rate	2 Mbps
PHY header*	192 μsec
MAC header	34 bytes
ACK*	248 μsec

- PHY header is transmitted at 1 Mbps, ACK comprises ACK frame + PHY header. The ACK frame is 14 bytes and is transmitted at basic rate, 2 Mbps, regardless of the data rate.

Table B.1 Parameter values of 802.11b DCF [WLL05]

For a multicast or broadcast packet, the transmitting station will not wait for the ACK because multicast destination stations do not transmit ACK. As a result, no retransmissions will be made for multicast and broadcast packets in IEEE 802.11 DCF. The transmitting station or AP will proceed to the next packet regardless of whether the previous packet has been successfully received. Typical values for the IEEE 802.11b DCF parameters are described in Table B.1.