Quality of Service (QoS) in 4G Wireless Networks

A thesis submitted in fulfilment of the requirements for

the degree of Doctor of Philosophy

ARN laboratory, iNext research centre

University of Technology Sydney

by

Fatima Furqan

Supervised by

Professor Doan B. Hoang

2015

DEDICATION

To my Husband, Furqan Naeem and Kids, Ayaan Furqan and Abdul Hadi Furqan

To my Parents, **Muhammad Yousaf Shah** (late) and Maimoona Yousaf

Thank you for your love and support

ACKNOWLEDGEMENTS

I am very thankful to Allah SWT who is the most merciful and beneficent. I sincerely express my deepest gratitude to my principal supervisor, Professor Doan B. Hoang, for his supervision and continuous encouragement throughout my whole PhD study. His guidance, wisdom, and enthusiasm have made me both more mature as a person and more confident to be a good researcher. He has been outstanding in providing insightful feedback and creating the perfect balance of my research engagement and my casual teaching work. From the beginning of determining the research direction to publishing fruitful research outcome, he always commits to foster my research skills. Without his guidance, I would still be in the marsh of research career. I feel so fortunate to have him as my supervisor for the past four and half years.

I thank Higher Education Commission (HEC) of Pakistan for offering me the Human Resource Development Scholarship (HRD). I thank the University of Technology Sydney for offering me an IRS scholarship. I also thank the iNext research centre for providing valuable resources, including funding for attending conferences. Special thanks go to the School of Computing and Communication at the University of Technology Sydney for offering me the tutorship that has significantly increased my teaching experience in academia. I thank Dr. Ian Collings, Deputy Chief CSIRO, to co-supervise me. I am very thankful to him for his valuable feedback into the thesis and also offering the top up scholarship. Thanks to these funding and support, this enabled me to concentrate on my research work without the burden of living. I also thank Dr. Priyadarsi Nanda for his support in getting the license of LTE module of OPNET.

My special thanks to my department in charge in Fatima Jinnah Women University, Nadeem Fakhar. My thanks also go to the staff members and research students in the ARN lab for their help, suggestions, friendship and encouragement: special thanks to Eryani Tjondrowalujo, Najmeh Kamyabpour, Lingfeng Chen, Noor Faizah Ahmad, and Dang Thanh Dat. My special thanks to my parents-in-law, Khawaja Muhammad Naeem and Saeeda Naeem. I also thank to my siblings and friends, Sehr Saood, Naveed Ahmed, Asma Naveed, Usman Naeem, Noman Naeem, Sehrish Noman, Ali Shah, Fatima Ali, Ali

Chaudhary, Fiza Ali, Ahsan Naeem, Maryam Naeem, Tasneem Memon and Muneera Bano Sahibzada, for their support.

Furthermore, I thank my parents for their upbringing and encouragement to succeed in my study. Then, I express my gratitude and appreciation to my husband for his love and support. His selfless sacrifices and commitment to the family, made it possible for me to finish my PhD studies. Finally, I would like to thank my family, my parents and parents-in-law for their encouragement, and thank all the people who helped me and contributed to this study.

CERTIFICATE OF ORIGINAL AUTHORSHIP

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

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

Signature of Student: Production Note: Signature removed prior to publication.

Date: 22nd May 2015

THE AUTHOR'S PUBLICATIONS

International Conference Publications and Proceedings

FURQAN, F. & HOANG, D. B. Analysis of Parameters Contributing Performance and Coverage of Mobile WiMAX with Mix Traffic. 12th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), 20-22 Oct 2011. 313-318.

FURQAN, F. & HOANG, D. B. WFICC: A new mechanism for provision of QoS and Congestion Control in WiMAX. IEEE Consumer Communications and Networking Conference (CCNC), 11-14 Jan 2013. 552-558.

FURQAN., F. & HOANG., D. B. Wireless Fair Intelligent Congestion Control -- A QoS Performance Evaluation. *In:* 13th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), 14-16 Dec, 2012. 3-9.

FURQAN, F. & HOANG, D. B. Wireless Fair Intelligent Admission Control -- WFIAC. IEEE 27th International Conference on Advanced Information Networking and Applications (AINA), 25-28 March 2013. 1001-1008.

FURQAN, F. & HOANG, D. B. 2014. LTE_FICC: A New Mechanism for Provision of QoS and Congestion Control in LTE/LTE-Advanced Networks. *Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer.

FURQAN, F., HOANG, D. B. & COLLINGS, I. B. "LTE-Advanced fair intelligent admission control LTE-FIAC", IEEE 15th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2014. IEEE, 1-4.

FURQAN, F., HOANG, D. B. & COLLINGS. B. I. "Effects of Quality of Service Schemes on the capacity and dimensioning of LTE Networks", 33rd IEEE International Performance Computing and Communications Conference (IPCCC), 5-7 December 2014.

International Journals

FURQAN, F. & HOANG, D. B. "Wireless Fair Intelligent Congestion Control — A QoS Performance Evaluation, Journal of Interconnection Networks, vol. 14, p. 1360001, 2013.

FURQAN, F., HOANG, D. B. & COLLINGS. B. I. “LTE/LTE-Advanced Fair Intelligent Admission Control – LTE-FIAC”, Journal of IEEE Computer Networks (submitted, under review)

ABSTRACT

Quality of Service (QoS) of 4th Generation Broadband Wireless Access (BWA) networks is directly affected by two factors: congestion in the network caused by changes in population density and application demand distribution; and varied attributes of network traffic such as minimum rate and delay requirements.

The current 4G BWA specifications define QoS parameters for each type of traffic, but do not provide QoS mechanisms including Radio Admission Control (RAC), scheduler and congestion prevention mechanism to ensure the QoS to existing and new connections within the network. Significant amount of research is dedicated to provide QoS and control congestion using RAC and scheduler. Current QoS mechanisms are inadequate to deal with network congestions and provide fairness among the traffic flows.

In this thesis, we have proposed a QoS framework and control algorithms for 4G BWA networks, Mobile WiMAX and Long Term Evolution (LTE). The framework includes a new load control mechanism, the Fair Intelligent Congestion Control (4G-FICC) and an intelligent admission control, the Fair Intelligent Admission Control (4G-FIAC), based on the QoS architecture of 4G BWA networks.

4G-FICC avoids and controls congestion at the base station of WiMAX and LTE networks, respectively. It avoids congestion through traffic balancing, while handles congestion when unavoidable, allocates resources fairly and minimizes resource underutilization. It estimates fair share of bandwidth for each type of service based on its current resource utilization, QoS constraints and load at the network. It ensures that the traffic is scheduled in a way that fairness is guaranteed among the traffic flows, without violating the QoS requirements of connections.

We have identified critical parameters of 4G-FICC and discuss the impact of various settings of these parameters on the network performance. Detailed and comprehensive simulations are performed in ns-2 and OPNET. The results show that 4G-FICC is always active in the network, whether the network is overloaded or underutilized. It performs extremely well in allocating resources fairly among different type of services, yet preserving

their QoS requirements in terms of throughput, delay and jitter. Furthermore, 4G-FICC is simple to implement, robust and relatively insensitive to parameter settings**.**

To ensure end-to-end delay and QoS, we propose a predictive RAC, the Fair Intelligent Admission Control for 4G networks (4G-FIAC). It admits or rejects an incoming connection based on the resource availability and the current load in the network. The key idea is to utilise feedback from the load control module to determine load in the network. The proposed RAC is based on the bandwidth borrowing and degradation of over provisioned connections in order to minimise blocking probability and maximise resource utilisation in the network.

Therefore, 4G-FIAC along with 4G-FICC avoids congestion in the network to guarantee QoS to end-users. Detailed and comprehensive simulations are performed in ns-2 and OPNET to show the efficiency of the proposed RAC scheme. Extensive simulations demonstrate that 4G-FIAC outperforms existing schemes in terms of blocking probability of different service classes and fair resource allocation.

In this thesis, we have performed a comprehensive study of parameters that affect both the capacity and coverage of 4G networks. It serves as a basis for designing effective QoS schemes for dynamic and mixed distribution of services. With thorough investigation of the impact of QoS schemes on the capacity and dimensioning of 4G networks, we have presented a general and efficient approach for the network operators to determine the extent to which current network configurations can effectively manage the dynamic variations in the access and core side of the network.

Different scenarios are presented in the thesis to evaluate the effects of QoS schemes on the capacity of the network. The results are valuable in assisting the network operators to determine the optimum point for re-dimensioning the network to minimise cost and ensure the QoS of connections in terms of throughput and delay.

The research results are not limited to 4G networks in particular, but can be applied to other next generation wireless technologies, to ensure QoS to users in the covered area.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AC	Admission Control
ACR	Allowed Class Rate
ARP	Allocation and Retention Priority
ASN-GW	Access Service Network-Gateway
BE	Best Effort
BER	Bit Error Rate
BP	Blocking Probability
BS	Base Station
BUR	Buffer Utilization Ratio
BW	Bandwidth
BWA	Broadband Wireless Access
CAP	Connection Arrival Procedure
CBR	Constant bit Rate
CC	Congestion Control
CCCH	Common Control Channel
CCM	Congestion Control Module
CDP	Connection Departure Procedure
CoB	Class of Bearers
CoS	Class of Service
CP	Cyclic Prefix
CPS	Common Part Sublayer
CS	Complete Sharing
CSN	Connectivity Service Network
DL	Downlink
DSL	Digital Subscriber Line
eNodeB	Enhanced NodeB
EPC	Evolved Packet Core

EPS	Evolved Packet System
ER	Expected Rate
ERRM	Extra Resource Reservation Module
ertPS	Extended Real time Polling Service
E-UTRAN	Evolved -Universal Terrestrial Radio Access Network
f(Q)	Queue control Function
FDD	Frequency Division Duplex
FIAC	Fair Intelligent Admission Control
FICC	Fair Intelligent Congestion Control
FTP	File Transfer Protocol
FTTH	Fiber To The Home
GBR	Guaranteed Bit Rate
GPC	Grant Per Connection
GPRS	General Packet Radio Service
GPSS	Grant per Subscriber Station
GSM	Global System for Mobile
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSUPA	High Speed Uplink Packet Access
HTTP	Hypertext Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IMT-Advanced	Internal Mobile Telecommunication- Advanced
IP	Internet Protocol
ITU	International Telecommunication Union
LE	Load Estimation
LTE	Long Term Evolution
MAC	Medium Access Control
MACR	Mean Allowed Class Rate
MBR	Maximum Bit Rate

MCS	Modulation and Coding Scheme
MME	Mobility Management Entity
MRTR	Minimum Reserved Traffic Rate
MSTR	Maximum Sustained Traffic Rate
NBN	National Broadband Network
NIST	National Institute of Standards and Technology
Non- GBR	Non Guaranteed Bit Rate
nrtPS	Non Real Time Polling Services
ns-2	Network Simulator-2
OFDMA	Orthogonal Frequency Division Multiple Access
OH	Overheads
PAPR	Peak-to-Average Power Ratio
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PER	Packet Error Rate
PF	Proportional Fair
P-GW	PDN GW
PRACH	Physical Random Access Channel
PRB	Physical Resource Block
PUCCH	Physical Uplink Control Channel
Q_0	Target Operating Point
QCI	QoS Class Indicator
Qlen	Queue Length
QoC	QoS Class
QoS	Quality of Service
RAC	Radio Admission Control
RB	Resource Block
RE	Resource Element

RLC	Radio Link Control
ROHC	Robust Header Compression
RR	Round Robin
RRM	Radio Resource Management
rtPS	Real Time Polling Services
SAE	System Architecture Evolution
SC-FDMA	Single Carrier- Frequency Division Multiple Access
SDF	Service Data Flow
S-GW	Serving Gateway
SINR	Signal-to-Interference-to- Nose Ratio
SLA	Service Level Agreement
SNR	Signal-to-Noise Ratio
SS	Subscriber Station
TB	Transport Block
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TFT	Traffic Flow Template
ToS	Type of Service
TTI	Transmission Time Interval
UE	User Equipment
UGS	Unsolicited Grant Services
UL	Uplink
UMTS	Universal Mobile Telecommunication System
VBR	Variable bit Rate
VNI	Visual Networking Index
VoIP	Voice over IP
VP	Virtual Partitioning
WCDMA	Wideband Code Division Multiple Access
WFIAC	WiMAX Fair Intelligent Admission Control

WFICC	WiMAX Fair Intelligent Congestion Control
WiMAX	World Wide Interoperability for Microwave Access
WRR	Weighted Round Robin
3GPP	3rd Generation Partnership Project

Chapter 1 Introduction

Cisco's latest Visual Networking Index (VNI) reports that Internet traffic will increase to four times its current size by 2016, hitting an impressive figure of 1.1 zettabytes (10^{21}) of data per year (Cisco, 2014). The proliferation of tablets, mobile phones, and other smart devices are driving up the demand for connectivity. The report forecasts that by 2016 there will be nearly 18.9 billion network connections—almost 2.5 connections for each person on earth.

Keeping in view the future demand of the Internet, National Broadband Network (NBN) Australia is providing high-speed Fiber To The Home (FTTH) to 93% of premises. Due to the cost and physical limitations, NBN utilizes fixed wireless and satellite to cover 4% and 3% of premises, respectively.

Mobile WiMAX and LTE technologies have been considered for wireless broadband access as they evolve towards the 4G networks that can deliver up to 100Mbps and beyond. Mobility support and provision of high data rates make 4G networks the choice of the future broadband Internet access. In rural areas where only the deployment of wireless is feasible and cost effective the 4G networks are inevitable. The wireless broadband portion of the NBN covers many rural communities of different shapes and sizes from the small areas with high population density to vast geographical areas that are sparsely populated. Clearly, appropriate dimensioning of the whole access network is crucial to ensure the overall cost is minimized under a constraint that the specified data rates are guaranteed to all the users regardless of their geographical locations.

Once the network has been dimensioned, its characteristics may change. For instance, the load at the core network, which connects a base stations to the Internet increases or the traffic demand of a cell increases beyond the forecast due to changes in population density or application distributions. In these situations, a network without appropriate QoS schemes may not be able to manage these dynamic changes and as a result QoS of connections degrades. Consequently, it imposes a requirement on the network operators to re-dimension the network. Whereas, with the employment of appropriate QoS schemes the network operators are able to manage increase in

the data demand and the load at the core network to a certain extent, without the need to re-dimension.

The 3GPP and IEEE 802.16-2005 standards define specific QoS parameters for each type of service but do not provide QoS schemes including RAC and scheduler and left them as an open research area. The standard also does not specify any mechanism for congestion prevention in the network.

The aim of this dissertation is to contribute to the development of a QoS framework for 4G networks that effectively manage the dynamic variations in the demography and traffic demands of the network. Following factors motivates the proposed QoS architecture.

Firstly, when performing the network dimensioning, the bandwidth requirements of network connections are planned based on the condition that the network traffic is pre-determined with limited room for variations. Whereas, some services such as file transfer and email, which employs Transmission Control Protocol (TCP) at the transport layer, are elastic. They adjust their source rates according to the available bandwidth in the network. In the current 4G standards bandwidth allocation is specified by a range, which is stated by the two parameters, minimum and maximum rate of a connection. The proposed QoS schemes utilize the ability to adapt the rate allocated to the existing connections to manage the increase in load and demand in the network.

Secondly, user's applications differ in their QoS constraints. Some applications such as video streaming, online gaming and VoIP require a minimum bandwidth and delay guarantee. Whereas, services such as buffered streaming, FTP and Web do not have stringent delay requirements. Our proposed QoS schemes differentiate among the services to ensure QoS and fair resource allocations among the connections.

Finally, the network service providers are facing increase in the demand of more efficient resource management and load control. The service providers can improve their network service quality by employing the proposed QoS schemes, which provide service differentiation to ensure delay and bandwidth requirements of connections. It enables the service providers to provide

QoS to existing and new connections in a cost effective manner even when the load at the access and core side of the network increases.

This chapter outlines key issues of this research and presents the aims and objectives of this thesis. Furthermore, it summarizes main contributions of this research and discusses research model and methodology. Finally, it provides an overview of the remainder of this thesis. Before presenting the research aims, the chapter provides a brief introduction to the BWA networks and the QoS.

1.1 Defining Broadband Wireless Access Networks (BWA) and QoS

BWA networks provide high-speed Internet access to the users in a wide area with wireless technology. Internal Mobile Telecommunication - Advanced (IMT-Advanced) has given the requirements for 4G BWA networks, which are listed below (ITU-R, 2008).

- High Spectral Efficiency to offer peak data rates of 100 Mbps for high mobility access, and 1 Gbps for low mobility access.
- Packet switched optimized
- High level of mobility and security
- Optimized terminal power efficiency
- Supports scalable bandwidths

LTE and IEEE 802.16 WiMAX (Worldwide Interoperability for Microwave Access), both are considered as candidate technologies for the 4th Generation (4G) of mobile networks. Both of the technologies offer

- Fully integrated IP solutions.
- “Anytime” “Anywhere” access.
- Spectrally efficient system.
- High QoS to various applications

Long Term Evolution (LTE) is the latest standard of the 3rd Generation Partnership Project. LTE radio access, Evolve UMTS Terrestrial Radio Access Network (E–UTRAN), supports high system capacity, low latency and user’s mobility. LTE deploys two separate access techniques

for uplink and downlink transmissions. It employs Orthogonal Frequency Division Multiple Access (OFDMA) on its downlink. It uses Single Carrier- Frequency Division Multiple Access (SC-FDMA) on its uplink, with a view to reduce Peak-to-Average Power Ratio (PAPR) to save the battery power at a User Equipment (UE). LTE physical supports transmission in both Time Division Duplex (TDD) and Frequency Division Duplex (FDD) modes. LTE physical supports scalable bandwidths from 1.25 to 20 MHz. LTE supports peak data rates of 150 Mbps and 50 Mbps in Downlink (DL) and Uplink (UL), respectively.

IEEE 802.16 series of standards aim to provide BWA over a long distance. Mobile WiMAX deploys OFDMA access technique for both uplink and downlink transmissions. WiMAX physical supports transmission in both TDD and FDD modes. Mobile WiMAX physical supports scalable channel bandwidths from 1.25 to 20 MHz. It supports peak data rates of 128 Mbps and 56 Mbps in DL and UL, respectively.

Table 1.1. Comparison between Mobile WiMAX and LTE

	Mobile- WiMAX	LTE
Multiplexing	OFDMA	OFDMA, SC-FDMA
Duplexing Mode	FDD, TDD	FDD, TDD
Modulation	QPSK, 16-QAM, 64-QAM	QPSK, 16-QAM, 64-QAM
Channel BW (MHz)	1.25 - 20 MHz	1.25 - 20 MHz
Frequency Bands	2.3-2.4 GHz 2.5-2.7 GHz 3.3-3.4 GHz 3.4-3.6 GHz	2 GHz 2.6 GHz 700MHz 900 MHz
Downlink Peak Data Rate	128 Mbps 1Gbps (Rel. 2)	100 Mbps 1 Gbps (LTE-ADV)
Uplink Rate	56 Mbps	50 Mbps
Coverage	5-10 km, 50 km	5km, 5-30 km, 100 km

WiMAX is the candidate 4G technology, evident from the recent incident of Holborn fire on 1st April, 2015. The fire resulted in major disruption to broadband services in the affected area. The businesses in the affected area turned to WiMAX to obtain the broadband services (Scroxton, 2015).

WiMAX continues to grow across Africa and Middle East regions. Furthermore, Afghanistan ministry of communication and information technology offered WiMAX licenses to operators in 2012 and expected it will grow to 80% of Afghanistan population in only two years (Cyprien, 2012). Pakistan has the highest WiMAX penetration globally. According to the WiMAX Forum president Declan Byrne " WiMAX technology has achieved a penetration rate of 50% of all broadband connections in urban centers in Pakistan" (Rehman, 2012).

WiMAX operators with TDD spectrum in Japan, Korea, Malaysia and USA will continue to grow. WiMAX technology offers an optimal solution for private networks serving dedicated segments such as energy utilities, smart grid industrial applications of telemetry, measurements and managements of critical systems that require real time and high security such as aviation, transportation and oil gas industries. WiMAX is expected to harmonize and integrate with LTE such that WiMAX networks serve the specialized segments while LTE provides the public networking (Aldmour, 2013).

QoS refers to the overall quality of the applications experienced by the network users. QoS of connections in the network can be measured in terms of several parameters. In this thesis, throughput, queuing delay, packet loss and fair resource allocation are considered.

a) **Throughput:** It is the data rate (bits per second) of the successfully received traffic on the network.

b) **Queuing Delay:** It indicates delay of packets at the queue of a base station's buffer. In situations, when the packets departure rate from the buffer is less than the arrival rate to the buffer, the queuing delay increases. Queuing delay has a significant impact on the performance of real time applications, such as voice, live streaming and online gaming.

c) **Packet Loss:** In this thesis it refers to the packet loss at the output buffer due to an overflow. It happens only when the rate at which packets arrive in the buffer is more than the rate at which they leave the buffer. Similar to queuing delay, packet loss has a significant impact on the QoS of real time applications.

d) **Fair Resource Allocation:** This attribute indicates the ability of the QoS schemes to differentiate among different type of services and allocate resources to meet the QoS requirements of each service type. It also enables the network to allocate the same amount of resources to the connections with the same QoS requirements and hence ensure fairness in the network.

1.2 Motivation and Research Issues

The NBN is providing the first largest Australian broadband network, and there is not any significant research to show the experience on how wireless can cover the whole rural area of Australia. We want to investigate the dimensioning and QoS aspects of BWA part of the NBN. The aim is to perform network dimensioning in terms of coverage and capacity and to offer cost effective and efficient dimensioning solutions. As network service providers have to ensure provision of guaranteed service level under the dynamic changes in population density and traffic distribution. So we aim to propose QoS mechanisms that ensure efficient network resource management and load control with varying population and application distributions.

Efforts have been dedicated to provide QoS and to control congestion. However, current researches do not offer a sufficient and comprehensive solution for the QoS in BWA networks, due to the following reasons.

- There is no efficient congestion relief scheme. Current approaches to perform load balancing are either based on thresholds or are applicable to a specific protocol, such as TCP only (Qiu. et al., 2011). QoS provisioning schemes, which are based on thresholds, cannot ensure QoS to connections. It is because they are active only when the network is already heavily loaded. They do not operate when the network is approaching to congestion. This mode of operation results in inefficient utilization of the network resources.

- The parameters that had been defined and employed in the existing schemes are mostly set manually (Kwan. et al., 2010, E. O. Lucena et al., 2010, Emanuel B Rodrigues and Francisco Rodgrigo P Cavalcanti, 2008). These schemes lack the rationale behind the

selection of thresholds and the impact of variations in parameter settings on the network performance. Hence, these schemes are not directly applicable to the networks, which are changing dynamically.

- Fairness among the service flows of different service types is rarely considered. Furthermore, current approaches do not consider fair resource allocation among the flows of the same type of service.

- Current approaches of RAC consider only the network resources and do not consider the load at the core network. Most of these schemes (Anas et al., 2008) are not able to differentiate among the connections of the different service types when resources are limited in the network. Several schemes provide service differentiation (Delgado and Jaumard, 2010, Bae. et al., 2009, Kwan, 2010, Lei. et al., 2008, Tung et al., 2008) , but they employ thresholds and resource reservation and often lead to inefficient utilization of network resources.

  The existing schemes based on pre-emption or degradation rarely ensure fair bandwidth allocation among the service flows at the same as well as different priority levels (Kwan. et al., 2010, Priya and Franklin, 2012, Qian et al., 2009, Khabazian et al., 2012, Khabazian et al., 2013). Some of the RAC schemes employs fixed size degradation step and do not offer rationale behind the selection of the value (Wang et al., 2005, Suresh. et al., 2008). Additionally, these RAC schemes in the admission process do not consider the current congestion state at the core network.

- RAC schemes, which reserve additional resources with an incoming connection to offset changes in the user demand due to variations in channel conditions, are also proposed in the literature (Mehdi. et al., 2012). None of these schemes can be directly applied to 4G BWA networks, which already allow the existing connections to take additional resources above their minimum requirements to gain their maximum rate.

Therefore, existing approaches are not sufficient to ensure QoS to connections and to cope with dynamic changes in the 4G BWA networks.

1.3 Research Aims and Objectives

The primary focus of this research is to ensure QoS to the users of 4G BWA networks. The aim is to develop a QoS framework that enables the network operators to effectively manage an increase in traffic demand and load at the core network, and provide the requested QoS to the end users. The existing QoS schemes have fixed value parameters, therefore, these schemes are unable to cope with dynamic variations in demand and ensure fair resource allocation in the network. We propose innovative and intelligent QoS schemes that are always active in the network. The schemes are scalable and robust in changing network scenarios, as the parameters involved are functions of the current network usage and the current network state.

Furthermore, the research identifies the factors that affect both the capacity and coverage of the network such as bandwidth, frequency, signal-to-noise-ratio, system and protocol overhead, and Modulation and Coding Scheme. These factors are detailed in 0.

The objectives of this thesis are identified to address the above research issues, and are described as follows:

- Focus on provisioning of QoS in IEEE 802.16 based Mobile WiMAX and 3GPP LTE, 4G wireless networks.
- Develop a comprehensive framework to efficiently manage the dynamic changes in the network.
- Propose and implement QoS mechanisms that ensure provision of QoS in terms of throughput, delay, jitter and particularly fairness among the connections.
- Achieve network optimization in terms of efficient resource utilization.
- Propose schemes that are scalable, reliable, robust and insensitive to minor mistuning of the parameters.
- Develop the QoS mechanisms that are based on QoS constraints, current resource usage of each type of the service and the current network state.
- Perform network dimensioning to determine coverage and capacity of the system for a mix traffic distribution. Determine the effective parameters that specifically affect both coverage and capacity of the network.

- Apply the proposed QoS schemes to BWA scenarios to determine the effective capacity of the network.

Proposed QoS Framework

Figure 1.1 depicts our proposed QoS framework for mobile WiMAX and LTE networks. In the proposed framework the QoS schemes including congestion control, admission control and scheduler interact and cooperate to keep the network in a stable state.

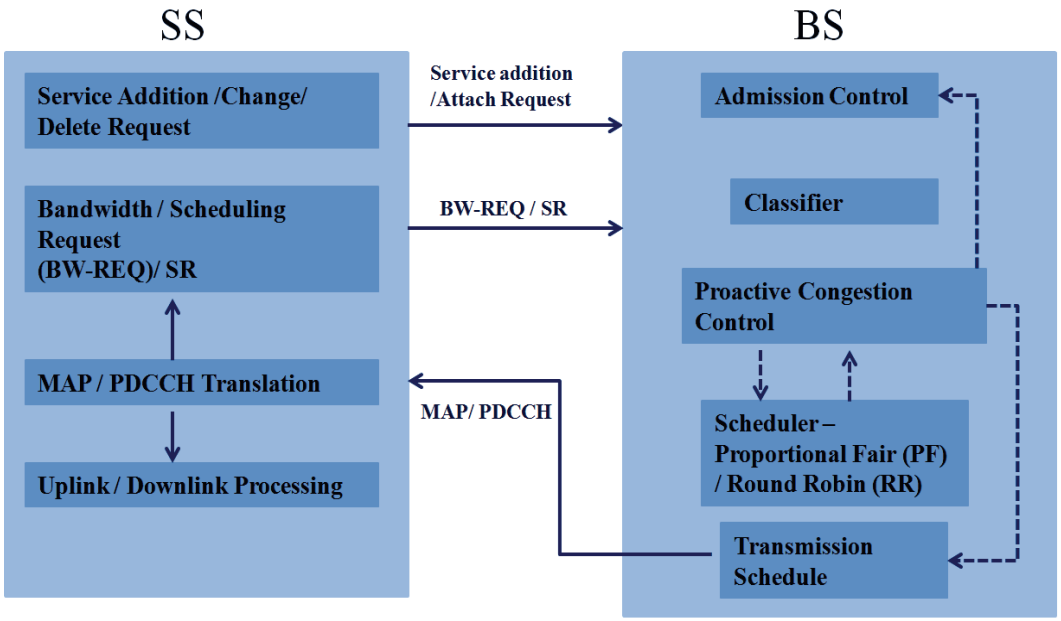


Figure 1.1 Proposed QoS Framework

In relation to the research objectives described above, major outcomes of this thesis are as follows.

a) A solution to estimate the capacity of the network with a specific network setup, application distribution and population density.

b) A Congestion Control (CC) mechanism that controls the load and ensures the QoS of all service types. It coordinates with the scheduler to ensure the traffic is scheduled in such a way that load can be managed efficiently with fair resource allocation. The CC is always active whether the network is overloaded or underutilized and ensures the network resources are optimally utilized.

c) A RAC scheme that work in conjunction with the CC scheme to determine the admission of an incoming connection. The proposed RAC employs an intelligent degradation procedure. The proposed degradation procedure ensures two vital requirements, service differentiation and fair resource allocation among the connections at the same as well as different priority levels.

1.4 Research Contribution

4G networks like IEEE 802.16 and LTE are competitive alternative to wire line BWA in provisioning QoS to different traffic types. According to Cisco VNI report, the 4G networks traffic is expected to grow to 51 % of total mobile traffic by 2018. Both IEEE 802.16 and 3GPP leave the resource management to be vendor specific.

Provisioning guaranteed service level is more challenging when the network changes in terms of population density and traffic distribution. Despite the significant research available on the QoS in computer networks, the QoS schemes for the 4G BWA networks are still progressing. Our novel QoS framework brings in new ways of provisioning QoS in 4G BWA networks. The major contributions of this study are listed below:

- The research provides a new way of QoS provisioning in 4G networks, where the QoS mechanisms namely RAC, load control and scheduler work together to provide QoS in terms of throughput, fair resource allocation, delay and jitter.

- The proposed QoS schemes conform to the original QoS architecture of the candidate 4G technologies. They do not necessitate any changes in the existing system architecture. The schemes involve information sharing only among the modules at a base station. Consequently, the schemes do not require sending feedback to the user equipment and hence avoid control information overheads.

- The proposed QoS framework avoids and effectively controls the congestion at the core side of the BWA networks. It accommodates a new mechanism, 4G-FICC, to estimate the fair share of bandwidth for each type of service based on its current average usage of

the resources and the current congestion state. To utilize network resources effectively, it oversells the bandwidth to connections when the network resources are underutilized.

- The RAC component of the proposed framework, 4G-FIAC, guarantees QoS to existing and new connections in the network. It intelligently predicts the available resources using the degradation scheme. It determines the admission using the feedback from proposed congestion control module, 4G-FICC, to avoid congestion and to guarantee QoS to applications.

- The propose schemes, 4G-FICC and 4G-FIAC, use simple algorithms involving few parameters. This study thoroughly investigates the effect of variation in parameter settings on the performance of the schemes. This analysis of parameters makes the schemes realizable and scalable in the 4G QoS architecture.

- The study discusses ways to effectively reduce system and protocol overhead. It provides a detailed study of the parameters, which affect the capacity and coverage of the networks. It discusses the effect of the proposed QoS schemes on the offered capacity of the network. The results of this research are significant for the network designers to effectively determine the network capacity and to update different aspects of the network to provide the sufficient QoS for a specific traffic mix and population distribution.

1.5 Research Model and Methodology

The research strategy adopted in this thesis comprises of three main stages – problem definition, developing new approaches and evaluation.

To address each significant and key problem, we developed following research questions for this thesis.

Characterization:

- Are the existing QoS mechanisms offering the comprehensive QoS solution that is applicable to the different network scenarios?

Based on our research, no overall QoS provisioning approach for the 4G networks exists currently. The current schemes do not efficiently utilize the specifications of the standards to control the load and the increased demand. It is expected, the current approaches do not enable the network operators to effectively deal with the dynamic changes in the network.

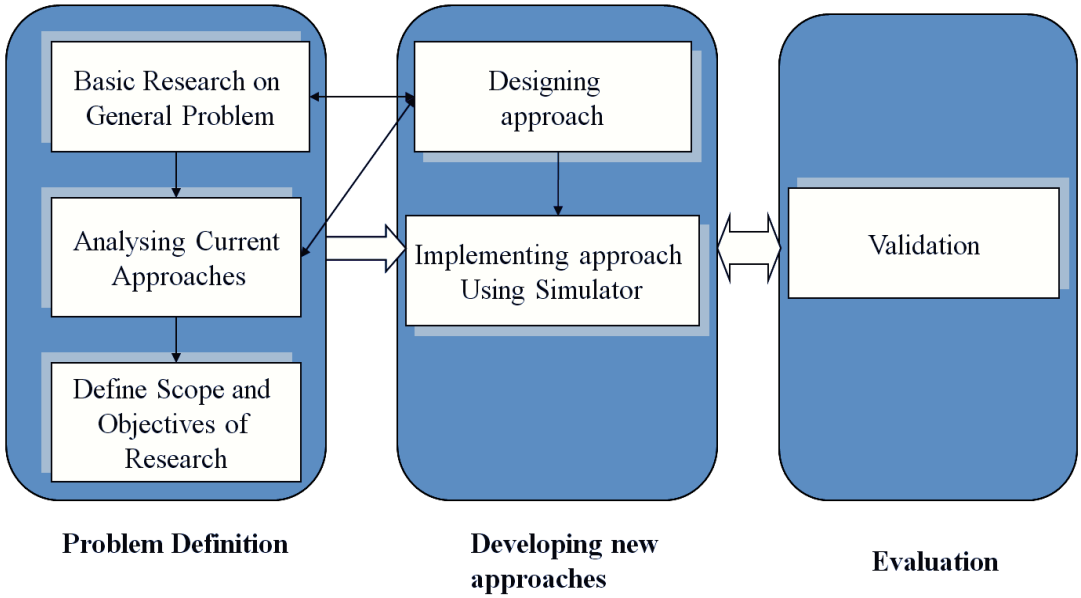


Figure 1.2 Research Methodology

Feasibility:

- How the proposed mechanism is used in integration with other 4G networks?

  The proposed control mechanisms for the provision of QoS are employed as separate modules at a base station. They are applicable to any last mile point to multipoint (PMP) architecture of 4G networks.

Selection:

- What are the most feasible methods for the validation of proposed QoS framework?

  Analyze the effectiveness of the proposed scheme through the simulations. Determine the optimum values of the parameters involved in the proposed mechanisms through the simulator to achieve the desired level of QoS. The proposed mechanisms and the

optimum values should be applied on different network scenarios and configurations to evaluate the effectiveness and robustness of schemes.

Generalization:

- Can the proposed mechanisms work in conjunction with existing mechanisms?
- Can they operate without changing underlying schemes on a base station?

The proposed mechanisms are determining the appropriate values of different important parameters in the network (such as data rate) and pass them to the existing mechanism (scheduler, admission control), which makes it suitable and flexible to work with any existing schemes for the efficient and fair resource allocation.

Risks:

- Is the problem of provision of the QoS in 4G networks effective especially when the network changes dynamically in terms of population, density and application distribution?

The proposed schemes use parameters that are function of the current network usage of each type of service and the current network state, which makes the scheme scalable and robust in changing network scenarios.

The designing and evaluation stages can be further classified in the following steps.

Step 1

- Extracting the parameters that effect both capacity and coverage of the network.
- Using simulation, perform network dimensioning and determine capacity and coverage of the system for different parameters settings.

Step 2

- It comprises of three sub phases, which are related to the design and development of the proposed algorithms.

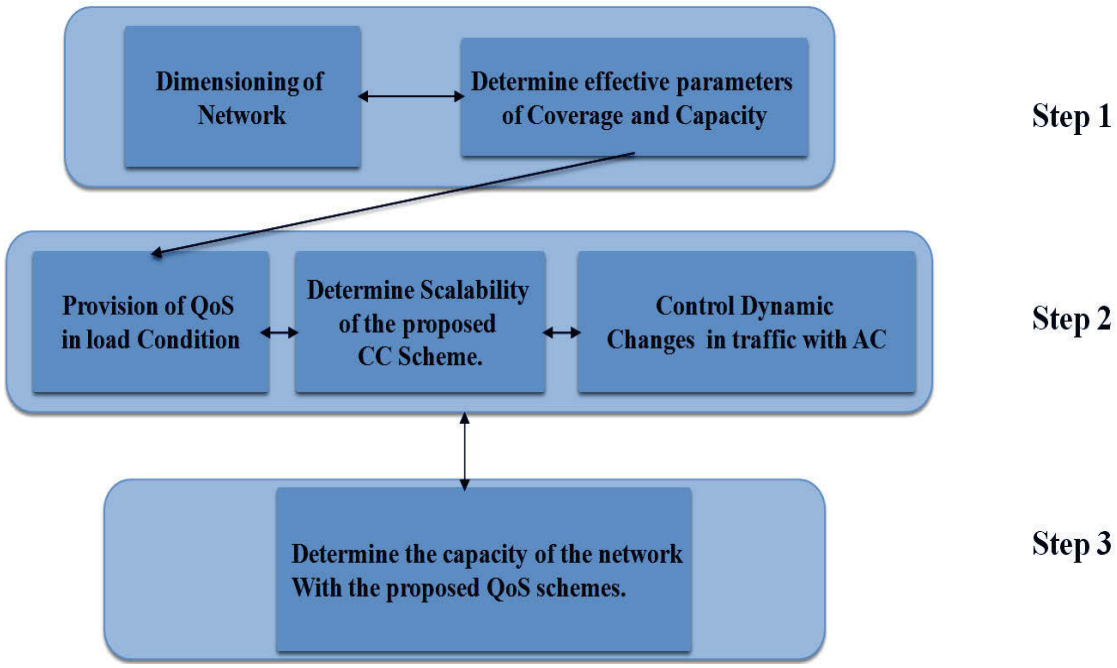


Figure 1.3 Steps of Designing and Implementation Phase

Step 2a

- Design and implement the CC mechanism, 4G-FICC, to ensure the network controls the load and also maintains the QoS of connections in the network.

Step 2b

- Analyze the performance of the network using the proposed CC scheme in terms of throughput, delay, jitter and fairness under various parameter settings using the simulation.

Step 2c

- Design and implement a RAC module, 4G-FIAC, that coordinates with the CC module, 4G-FICC, to admit an incoming connection.
- Evaluate the performance of the proposed RAC in terms of blocking probability of new connections, throughput, delay, jitter, fairness among the existing connections, and the load in the network.

- Evaluate the capability of the proposed RAC scheme to deal with the changes in resource demand of connections due to channel fluctuations.

Step 3

- Applying dimensioning and the proposed QoS mechanisms to provide an efficient and cost effective solution to the network operators to deal with the dynamic fluctuations in access and core side of the network.

Validation:

The proposed research idea is justified based on the current approaches. Figure 1.4 provides the QoS framework evaluation strategy.

The following steps are considered in the validation phase:

1. The idea is justified with the way current schemes provide QoS in the network and how the parameters are set for these schemes.
2. The proposed schemes, 4G-FICC and 4G-FIAC, are evaluated on the simulator.
3. The robustness and sensitivity of the parameters of the schemes are evaluated to verify the schemes are effective under the varying network conditions.

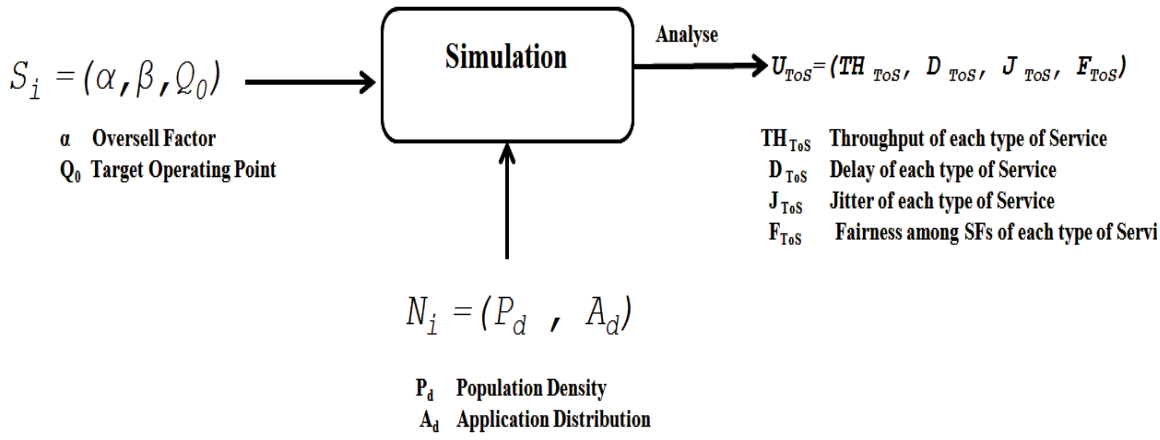


Figure 1.4 QoS Framework Evaluation Strategy

1.6 Structure of the Thesis

Chapter 1 is an introductory chapter and provides an overview of the complete research. It encompasses the, introduction to the BWA networks and the QoS, research questions formulated by the motivation provided by the issues of QoS provisioning in the 4G networks, research aims and objectives, research contributions, and research model and methodology of this thesis.

Chapter 2 presents a review of the literature of the existing QoS schemes for wireless networks. First we introduce system architecture of the candidate 4G technologies, WiMAX and LTE. Furthermore, we discuss the QoS architecture of both the technologies. Subsequently, we identify and explore existing solutions of QoS provisioning in the wireless networks.

Chapter 3 introduces the proposed QoS framework and control algorithms for the 4G networks. First we discuss basics of an intelligent CC algorithm for the 4G networks. We then present an overview of the proposed intelligent admission control algorithm for the 4G networks.

Chapter 4 investigates the CC algorithm specifically designed for the WiMAX QoS architecture. The chapter provides comprehensive evaluation of the proposed scheme. This chapter discusses the analysis of parameters involved in the proposed scheme. Afterwards, it thoroughly investigates the impact of parameters variation on the network performance.

Chapter 5 focuses on the CC algorithm specifically designed for the LTE QoS architecture. It discusses the modifications required in the scheduler to provide fair resource allocation in the network. It introduces the simulation model and discusses the changes required in the current simulator to implement the proposed scheme. The chapter provides comprehensive evaluation of the proposed scheme.

Chapter 6 discusses and presents an intelligent admission control for the WiMAX networks. The chapter addresses two important issues: the blocking probability and the load management in the core network. The chapter presents the information sharing between the load control and the admission control module to decide an admission in the network. Afterwards, the chapter provides a detailed evaluation of the proposed scheme and compares results with a basic admission control.

Chapter 7 discusses and presents an intelligent admission control designed for LTE networks. The chapter addresses three important issues: the blocking probability, the load management in the core network, and the extra resource reservation to handle the channel fluctuations. To address these issues the chapter presents load control, admission control and extra resource reservation modules. It discusses the communication between these modules to decide an admission of a connection in the network. Afterwards, the chapter compares the proposed scheme with the work of other researchers discussed in chapter 2 from various aspects and features. Towards the end, it discusses the effect of load estimation and extra resource reservation on the network performance.

Chapter 8 discusses the dimensioning of the 4G networks, WiMAX and LTE. It provides coverage and capacity analysis of these networks. It demonstrates the impact of various factors on the coverage and the capacity of the network. The chapter thoroughly discusses the impact of our proposed QoS schemes on the capacity of the network.

Chapter 9 summarizes the research work presented in this thesis, outlines the major contributions of the research, and maps direction for the future research plans and work.

Chapter 2 Literature Review and Related Work

This chapter presents the background of the 4G technologies, LTE and WiMAX. It provides the concepts necessary for understanding the proposed QoS framework and control algorithms. The background covers the system level architecture, protocols and QoS architecture of the technologies. The chapter systematically reviews the literature available for QoS mechanisms. Researchers have proposed various mechanisms to control load and provide QoS in 4G networks, including load balancing schemes; admission control scheme; scheduling approaches; combined Radio Admission Control (RAC) and scheduler schemes; and also combined load control, RAC and scheduler approaches. It facilitates to better understand and appreciate the distinctive features of the proposed control algorithms in the thesis.

The structure of this chapter is as follows: Sections 2.1 to 2.5 covers the background and introduces system architecture of the candidate 4G technologies, Mobile WiMAX and LTE, and their QoS architectures in detail. Section 2.6 focuses on the existing solutions for the QoS provisioning and points out the gaps that this thesis aims at filling. Section 2.7 concludes the chapter.

2.1 System Level Architecture of LTE and WiMAX

Long Term Evolution (LTE) is specified by the 3rd Generation Partnership Project (3GPP), which also provides GSM, GPRS, WCDMA and HSPA. The GSM release 98 was completed at the end of year 1998 and was followed by WCDMA release 99 at the end of year 1999. WCDMA specified the first Universal Mobile Telecommunication Services (UMTS) and offered data rate of 2 Mbps. The UMTS was upgraded to HSDPA and HSUPA in March 2002 and December 2004, respectively. The evolution continued to HSPA+ version 7 in 2007. The version 8 of LTE was approved at the end of year 2007 and its backward compatibility started in 2009. The 3GPP introduces the new architecture recognized as Evolved Packet System (EPS) as part of the two work items, namely System Architecture Evolution (SAE) and Long Term Evolution (LTE). The SAE provides specifications of all IP based Core Network (CN) called Evolved Packet Core (EPC). The LTE gives specifications of the radio access network termed as Evolved

Universal Terrestrial Radio Access Network (E-UTRAN) (Christopher Cox, 2012). The specifications of LTE-Advanced were approved in 2010 (Harri Holma and Antti Toskala, 2009).

IEEE 802.16 series of standards aim to provide BWA over a long distance. IEEE 802.16-2004 focused on fixed applications. The WiMAX forum announced its first product based on IEEE 802.16-2004 in early 2006 (Jeffery G. Andrews et al., 2007). IEEE 802.16-2005 provides specifications for mobile data networks. Mobile WiMAX release 1.0 system profile is defined for TDD mode of operation. The air interface profile specifications release 1.5 enables the mobile WiMAX also in FDD mode. The release 2.0 of mobile WiMAX system profile focused on the major enhancements in the spectrum efficiency, latency and scalability of the technology to wider bandwidths (Kamran Etemad, 2008).

Figure 2.1 shows the system architecture of WiMAX and LTE networks.

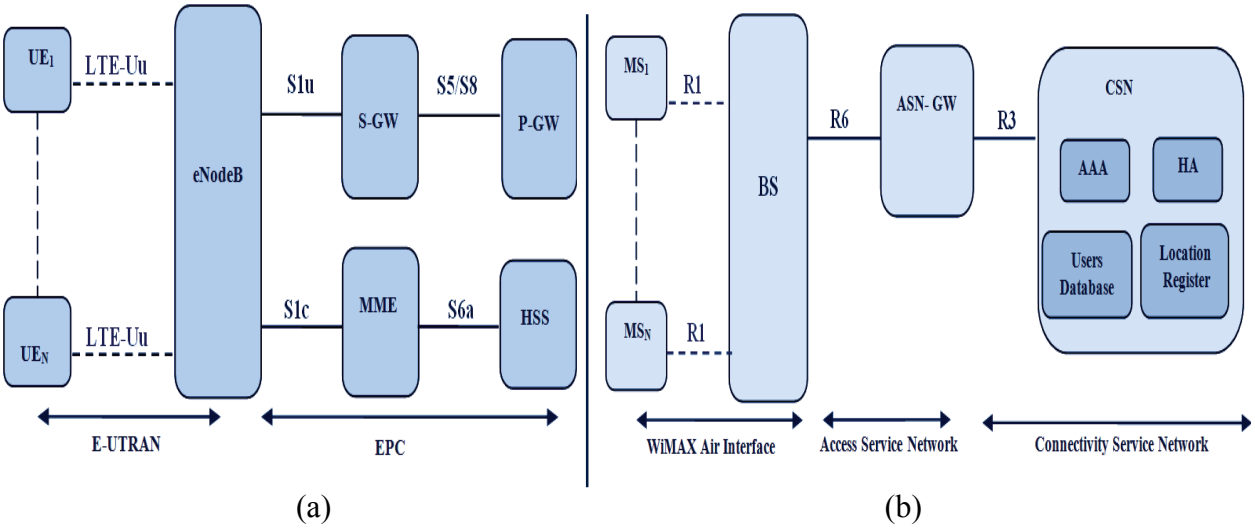


Figure 2.1 System architecture of (a) LTE Networks (b) WiMAX Networks

LTE handles communication between a subscriber device called User Equipment (UE) and an Evolved Node B (eNodeB). In E-UTRAN, an eNodeB is connected to UEs through LTE-Uu interface. In WiMAX air interface, a Base Station (BS) connects to Subscriber stations (SSs) through R1 interface. Both, LTE eNodeB and WiMAX BS, transmit traffic to the network users on Downlink (DL) and receives user's traffic on Uplink (UL). They perform functions of user's authentication, admission control, Hybrid Automatic Repeat Request (HARQ), DL/UL

scheduling, QoS policy enforcement, traffic classification and link adaptation on user's plane based on user's channel conditions (Bhandare, 2008).

In LTE system architecture, S1 interface connects LTE E-UTRAN to the core network. The packet core of LTE networks includes Home Subscriber Server (HSS), Mobility Management Entity (MME), Serving gateway (S-GW) and Packet Data Network (PDN) gateway (P-GW). In an EPC, the MME is a signaling entity that manages control plane traffic. The S-GW manages user plane traffic. The MME provides functions of authentication, security, mobility management, bearer management and idle state management. It also performs the functions of UE paging, UE location update and the selection of S-GW and P-GW (Bhandare, 2008). The MME is connected to HSS via S6a interface. The HSS is a database that stores user's subscription record.

In WiMAX system architecture, R1 interface connects the air interface to an Access Service Network Gateway (ASN-GW). The ASN-GW provides functions of both the S-GW and the MME entities of LTE networks. Both, LTE eNodeB and WiMAX BS, provide interface into all IP network and uses IP tunnels to transmit user's traffic to the S-GW (Christopher Cox, 2012, Pazhyannur, 2010) and the ASN-GW, respectively. The ASN-GW is connected to a Connectivity Service Network (CSN) through R3 interface. The CSN mainly includes Home Agent (HA), AAA server and Location Register. The AAA server provides functions of authentication, authorization and accounting. The location register stores information about user's current location and login. The P-GW of LTE and the HA of WiMAX, support mobility among the access gateways.

LTE employs GPRS Tunneling Protocol (GTP) for S1u interface and S1 Application Protocol (S1-AP)/Stream Control Transmission Protocol (SCTP) for S1c interface. It uses GTP based protocol or Proxy Mobile IP Version6 (PMIPV6) for interface between access gateways, S-GWs and P-GWs. Mobile WiMAX uses Generic Routing Encapsulation/User Datagram Protocol (GRE/UDP) as a tunneling protocol and UDP as control plane transport on its R6 interface. It employs (MIPV4) on the interface between access gateway and HA (Bhandare, 2008).

2.2 LTE Protocol Architecture

The 3GPP specification separates User Plane (UP) and Control Plane (CP) in the core network. The MME is the only control plane entity in an EPC. It controls user plane tunnel establishment and also establishes UP bearers between an eNodeB and the S-GW. The user plane bypasses the MME directly to a S-GW. The control plane carries control-signaling messages to support the user plane functions.

2.2.1 User Plane Protocol layered Architecture

The user plane carries user's applications traffic only. A UE has all layers including application, transport, IP and access stratum. An Access Stratum (AS) is a layer between a UE and the radio access network such as E-UTRAN. It manages the radio resources and is responsible to control the transportation of data over the radio interface LTE-U_u. It is also responsible for the transportation of Non Access Stratum (NAS) protocols over the radio interface. Figure 2.2 shows the user plane protocol stack for LTE network architecture (3GPP 23.401).

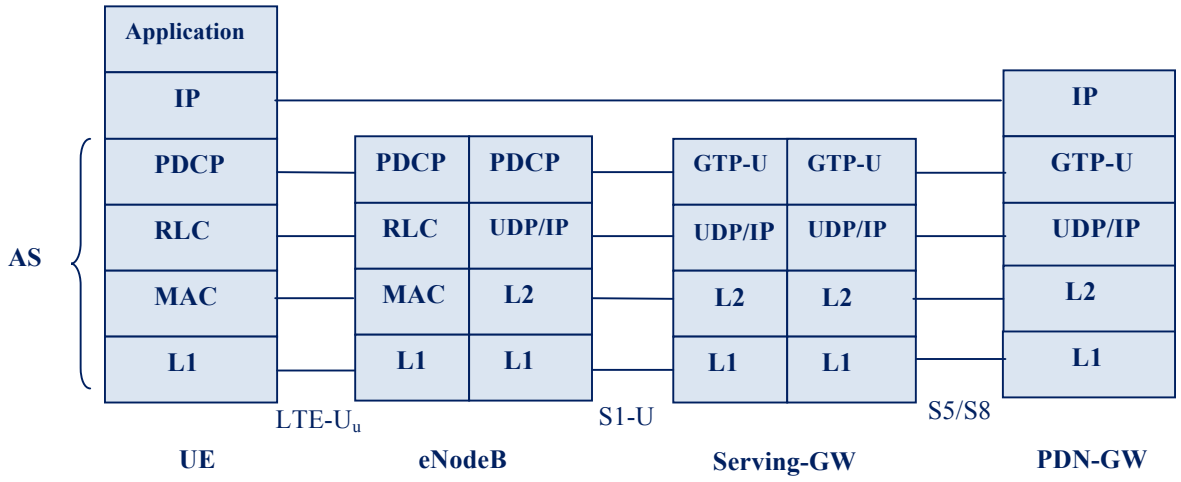


Figure 2.2 User Plane Protocol Stack (UE - P-GW) of LTE Networks

2.2.1.1 Access stratum (AS) Architecture

The UP Access Stratum (AS) protocol stack includes:

- Packet Data Convergence Protocol (PDCP).
- Radio Link Control (RLC)
- Medium Access Control (MAC)
- Physical layer / L1

Packet Data Convergence Protocol (PDCP): The PDCP layer provides the services of the transfer of user plane and control plane data, header compression and decompression of IP data packets using Robust Header Compression (ROHC) protocol, maintenance of PDCP sequence numbers, in-sequence delivery of upper layers PDUs at the re-establishment of lower layers and duplicate elimination of lower layer SDUs (3GPP TS 36.322).

The PDCP performs ciphering and deciphering of user plane and control plane data. It also performs integrity protection and integrity verification of control plane data. A fixed size PDCP header of 16 bits is added to all SDUs. All packets entering LTE go through PDCP encapsulation.

Radio Link Control (RLC): The RLC layer provides services of transfer of PDUs, RLC re-establishment, segmentation and concatenation of RLCs SDUs to form PDU of the size determined by a scheduler. The RLC operates in any of the three modes, Transparent Mode (TM), Unacknowledged Mode (UM) and Acknowledged Mode (AM). The TM-RLC does not support segmentation and concatenation so RLC header is not included in this mode. The UM-RLC and the AM-RLC support in-sequence delivery of SDUs to higher layers, duplicate detection and RLC SDU discard. The AM-RLC additionally provides protocol error detection and correction through Automatic Repeat Request (ARQ) (3GPP TS 36.322).

Medium Access Control (MAC): 3GPP MAC at both, a UE and an eNodeB, performs mapping between logical channels and transport channels. It supports error correction through Hybrid ARQ (HARQ). It also performs multiplexing and demultiplexing of SDUs from one or different logical channels onto the Transport Blocks (TBs) to be delievered to the physical layer. At an eNodeB, it additionally provides the services of transport format selection, prioirty handling between UEs and also among the logical channels of one UE. The MAC layer at a UE additionally performs logical channel prioritisation and scheduling information reporting for the UL (3GPP TS 36.321).

The MAC layer at an eNodeB uses a scheduler to generate the MAC Protocol Data Units (MPDUs) for the downlink subframe and to create the uplink grants for an uplink subframe. For FDD physical profile, the scheduler executes in every subframe and creates uplink grants for the $n+4^{th}$ subframe. For TDD physical profile, the scheduler executes only when the current subframe is of downlink.

2.2.1.2 Physical Layer:

The physical layer of 3GPP LTE employs OFDMA in DL and Single Carrier SC-FDMA in UL.

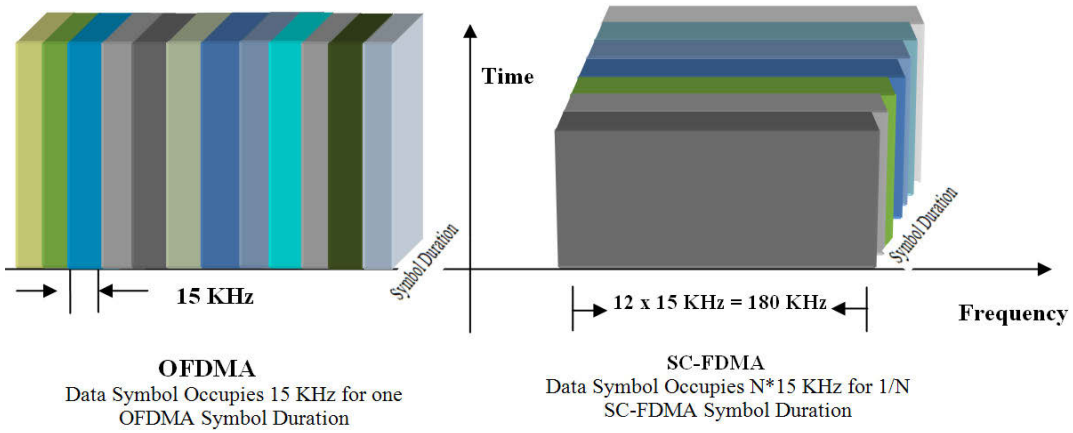


Figure 2.3 OFDMA and SC-FDMA

Both OFDMA and SC-FDMA uses a number of narrowband subcarriers. A major difference between them is how they use data symbols between the subcarriers. An OFDMA data symbol occupies narrowband subcarrier (15 KHz) extending for the entire duration of symbol (66.7 µs). A SC-FDMA data symbol has a shorter symbol time, but it occupies a wider bandwidth (N X 15 KHz) and visually appears as single carrier as indicated in Figure 2.3. The SC-FDMA employs single carrier modulation and as a result its Peak to Average Power (PAPR) is lower compared to OFDMA transmission.

LTE utilizes both TDD and FDD physical profiles. Most of the current implementations of LTE are in FDD mode, so in this thesis we confine our discussion to FDD mode. In LTE, frame structure type 1 is applicable to both full duplex and half duplex FDD.

Frame structure type 1

The frame structure type 1 is described in both time and frequency domain. In Time domain, each radio frame is 10 ms long. It consists of 20 slots of length 0.5 ms each, numbered from 0 to 19 as shown in Figure 2.4. A subframe is defined as two consecutive slots. In time domain, each slot consists of either 6 or 7 OFDM symbols for long or short cyclic prefix (CP), respectively (3GPP 36.211). In FDD, uplink and downlink transmissions are separated in the frequency domain. So for FDD mode, 10 subframes are available for both DL and UL transmissions in each 10 ms interval.

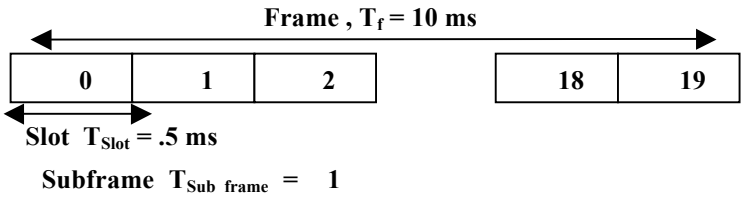


Figure 2.4 Frame Structure Type 1, Reproduced from (3GPP 36.211)

In frequency domain, each slot is described by a Resource Block (RB) and consists of 12 subcarriers each over 6 or 7 OFDM symbols. A pair of two RBs is termed as Transport Block (TB). A TB is the minimum unit used by a scheduler for the resource allocation. A single subcarrier for one symbol duration [0.5 ms (slot length) / 6 or 7= 66.7 μs] is represented as a Resource Element (RE). Therefore a RB has 84 or 72 REs depending on the configured cyclic prefix.

The number of available RBs depends on the bandwidth configuration in a cell. Table 2.1 provides the number of Physical Resource Blocks (PRBs) per subframe for different bandwidths.

Table 2.1. Available Resource Blocks per Subframe for Different channel Bandwidths

Channel Bandwidth (MHz)	**1.4**	**3**	**5**	**10**	**15**	**20**
Used Subcarriers	72	180	300	600	900	1200
N_{RB}	6	15	25	50	75	100

2.2.2 Control Plane Protocol layered Architecture

The control plane consists of protocols for control traffic and also to support user plane functions (3GPP 23.401). It manages the network access functions such as attaching to and detaching from an E-UTRAN. It controls the attributes of an established network access connection such as activation of an IP address and the establishment of routing path to support user mobility. It also manages the assignment of the network resources to meet changing user demands. Figure 2.5 provides control plane protocol stack for LTE networks (3GPP 23.401).

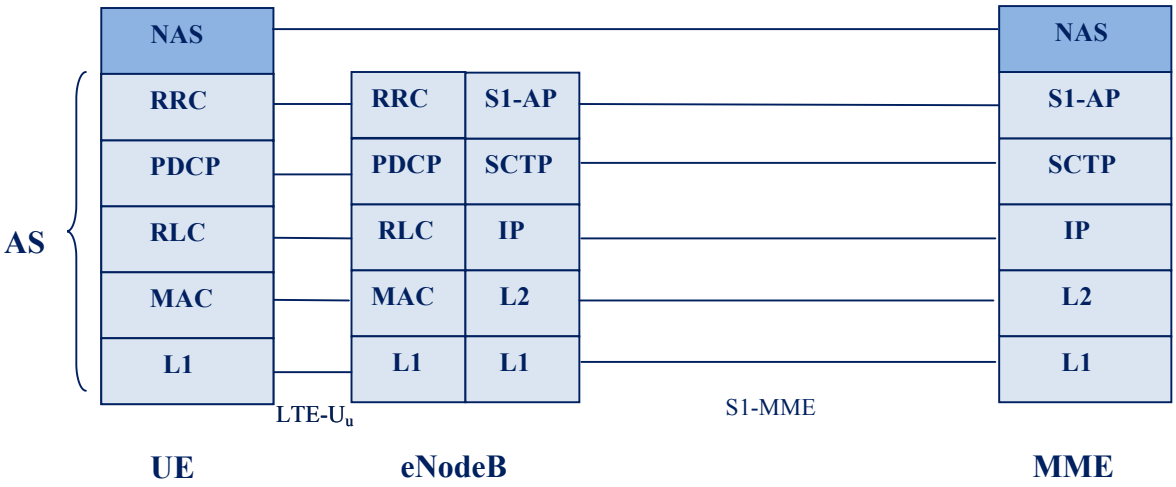


Figure 2.5 Control plane protocol Stack (UE – MME) of LTE Networks

The CP protocol stack at an eNodeB includes:

NAS: Non Access Stratum (NAS) is the top most layer in the CP. It is a functional layer directly between a UE and the MME in the core network for an E-UTRAN access. An eNodeB is only involve in the transportation of signaling messages (3GPP 23.401). The NAS protocols include EPS Mobility Management (EMM) and EPS Session Management (ESM). The EMM supports mobility management within the system. It is also responsible of ciphering and integrity protection of NAS signaling. The ESM controls the user plane bearer activation, modification and deactivation (3GPP 23.401).

Radio Resource Control (RRC): Similar to the access stratum protocols of the user plane, control plane's access stratum includes PDCP, RLC, MAC and L1 layers. The CP AS

additionally includes Radio Resource Control (RRC) layer. The RRC layer controls UE signaling and data connections. It also manages the functions for handovers.

2.2.3 Control Channels Overhead

The control channels carry control information in the cell. They have a direct impact on the resources available to the radio admission control to allocate to users. Table 2.2 shows the number of TBs per second (TBPS) and REs per TB for FDD physical profile after considering the control channels overhead with 5, 10, 15 and 20 MHz bandwidths, respectively. It clearly indicates that the number of TBPS increases as the bandwidth increases.

Table 2.2. Transport Blocks (TBs) per second and Resource Elements (REs) per TB

	Bandwidth							
BW	**5**		**10**		**15**		**20**	
	DL	**UL**	**DL**	**UL**	**DL**	**UL**	**DL**	**UL**
TBPS	25000	22400	50000	47400	75000	72400	100000	97400
Avg. RE per TB	123	139	124	139	125	139	125	139

In Table 2.2, the number of TBPS for UL is estimated taking into account the TBs required by Physical Uplink Control Channel (PUCCH) and Physical Random Access Channel (PRACH). The PUCCH carries control information such as user's Scheduling Request (SR), Hybrid ARQ (HARQ) ACK/NACK and Channel Quality Indication (CQI). To match the requirements and in accordance to the simulator employed (OPNET, 2012), in this thesis, 20 TBs per frame are assumed for PUCCH. The PRACH carries user's Random Access (RA) preamble and offers an interface between unsynchronized UEs and the uplink radio access (Telesystem Innovations Inc, 2010). In this thesis, we applied RA preamble format 0 and the number of RA resources per frame is set to 1. As a result, in the current analysis, RA channel occupies 6 TBs per frame.

Table 2.2 also shows the average number of REs per TB. On average there are 168 (12 X 7 X 2) and 144 (12 X 6 X 2) REs per TB for normal and extended CP, respectively. The estimation of REs takes into account the overhead of DL and UL reference symbols (pilot symbols). The

reference symbols are inserted in a subframe for channel estimation and signal demodulation at the receiver (Telesystem Innovations Inc, 2010). The reference symbols occupy 12 and 4 REs in each RB for UL and DL, respectively.

For DL, the estimation of REs per TB further takes into account Physical Downlink Control Channel (PDCCH) symbols per subframe. The PDCCH mainly carries DL assignments and UL grants, PRACH responses and UL power control commands. According to 3GPP specifications it can take 1, 2 or 3 symbols per subframe. For the estimation of REs per TB, the PDCCH symbols per subframe are set at 3 (21% OH). According to LTE specifications, the TB to bits mapping table is constructed by considering 120 reference REs in a TB, so an average number of REs per TB is estimated to match the specifications (OPNET, 2012).

The actual number of TBPS available in the admission procedure also depends on UL and DL loading factor. The loading factor indicates a fraction of total resources available to the RAC. In this work, UL and DL loading factors are set at 1. It indicates the maximum possible resources are available to the admission control algorithm for both UL and DL.

2.2.4 Protocol Overhead

A significant portion of the total resources available to the RAC is used to accommodate the protocol OH. The total amount of protocol OH is based on the scheduler in use (So-In. et al., 2010).

The Real-Time Transport Protocol/User Datagram Protocol/ IP (RTP/UDP/IP) and TCP/IP overheads (OH) per packet in an uncompressed mode are 40 and 60 bytes each for IPv4 and IPv6, respectively. The PDCP layer at the top of the radio interface protocol stack compresses higher layers overhead. It uses ROHC protocol, which is specified by Internet Engineering Task Force (IETF). In this work, in accordance to the simulator used for LTE (OPNET, 2012), the maximum compressed header sizes per packet are considered to be 4, 2 and 4 bytes for RTP/UDP/IP, UDP/IP and TCP/IP, respectively. As discussed earlier, the PDCP layer has a fixed overhead of 16 bits. The RLC layer has an overhead of 2 bytes. With segmentation or concatenation the RLC overhead is assumed to increase to 4 bytes. The MAC layer overhead is assigned a typical value

of 6 bytes. Also, the physical layer adds 3 bytes Cyclic Redundancy Check (CRC) checksum to data packet (Petter Edstrom, 2007).

2.3 QoS in LTE Networks

To guarantee service differentiation and provision of Quality of Service (QoS), LTE/ LTE-Advanced employ concept of bearer. A bearer identifies packet flows from one network element to another. The most important bearer is an Evolved Packet system (EPS) bearer, which carries data between UEs and a P-GW with predefined QoS characteristics. Each bearer is identified with an EPS Bearer Identity (EBI) also known as Radio Bearer ID (RB-ID). Figure 2.6 illustrates EPS bearer architecture.

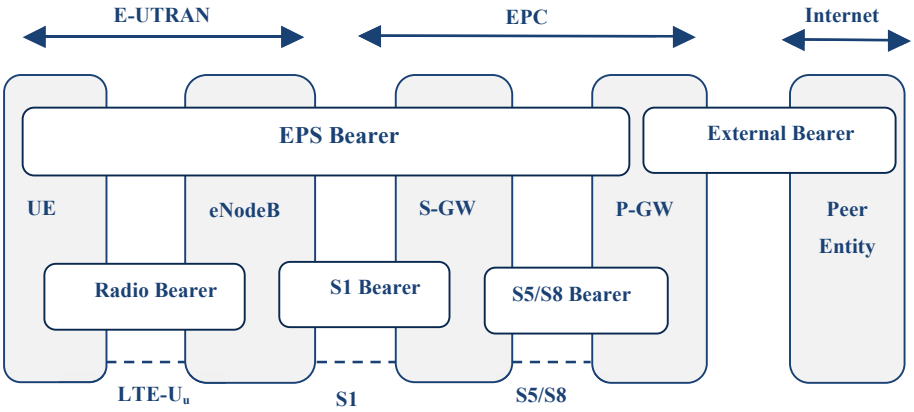


Figure 2.6 EPS Bearer Service Architecture (3GPP TS 36.300)

2.3.1 QoS Parameters of EPS bearers

The QoS profile of an EPS bearer mainly includes QoS Class Identifier (QCI) and Allocation and Retention Priority (ARP).

2.3.1.1 QoS Class Identifier (QCI)

The EPS QoS concept is class based, so each bearer is associated with only one QCI. A QCI is a scalar value that classifies a QoS class to which a bearer belongs. Each QCI is a reference to a set of QoS attributes including resource type (GBR or non-GBR), priority, packet delay budget and an acceptable packet error loss rate. It determines packet forwarding treatments such as

scheduling weights, admission thresholds, queue management thresholds, and link layer protocol configurations between a UE and a P-GW (3GPP TS 36.300, 3GPP 23.401).

Standardized QCIs specified in 3GPP specifications (3GPP 23.203) are shown in Table 2.3.

Table 2.3. Characteristics of LTE Standardized QCIs

QCI	Resource Type	Priority	Delay	PELR	Examples
1	GBR	2	100 ms	10^{-2}	Conversational Voice
2		4	150 ms	10^{-3}	Conversational Video (live streaming)
3		3	50 ms	10^{-3}	Real-time games
4		5	300 ms	10^{-6}	Non Conversational Video (buffered Streaming)
5	Non-GBR	1	100 ms	10^{-6}	IMS Signaling
6		6	300 ms	10^{-6}	Video (buffered streaming), Web, Email, ftp
7		7	100 ms	10^{-3}	Voice, Video (live streaming), interactive games
8		8	300 ms	10^{-6}	Video (buffered streaming), Web, Email, ftp
9		9			

With respect to resource type, LTE/LTE-Advanced supports two types of bearers, *Guaranteed Bit Rate* (GBR) and *Non Guaranteed Bit Rate* (non-GBR).

Guaranteed Bit Rate (GBR) bearers

A GBR bearer is established or modified only on demand and reserves network resources corresponding the GBR value associated with it.

Each GBR bearer is associated with the following bearer level QoS parameters.

Guaranteed Bit Rate (GBR): It is the minimum guaranteed bit rate to be provided by a GBR bearer to a UE. It is specified separately for uplink and downlink (3GPP 23.401).

Maximum Bit Rate (MBR): It indicates the maximum bit rate to be provided by a GBR bearer. It limits the bit rate that can be provided by a GBR bearer to a UE (3GPP 23.401). It is also specified separately for uplink and downlink.

Non Guaranteed Bit Rate (non-GBR) bearers

A non-GBR bearer does not reserve network resources and can experience congestion related packet losses. Each non-GBR bearer is associated with the following bearer level QoS parameters.

APN-AMBR: Each Access Point Name (APN) is associated with Aggregate Maximum Bit Rate (APN-AMBR). It limits the aggregate bit rate that can be provided to all Non-GBR bearers of the same APN.

UE-AMBR: Each UE is associated with Aggregate Maximum Bit Rate (UE-AMBR). It is the maximum bit rate that is assigned to all non-GBR bearers of the same UE.

Every QCI is associated with a priority level. The priority levels differentiate among the flows of the same UE and also among the flows from different UEs (3GPP 23.203). In this thesis, QCIs are grouped into Class of Bearers (CoBs) according to the resource type, GBR and non-GBR. The GBR CoB groups the priority levels from 2 to 5. The non-GBR CoB groups the priority levels from 6 to 9.

2.3.1.2 Allocation and Retention Priority (ARP)

An ARP parameter is an integer in the range 1-16. It is used for the prioritization of EPS bearers during the admission procedure. The ARP provides information about a bearer's priority level, pre-emption capability and pre-emption vulnerability. The admission control uses ARP to determine whether a bearer establishment or modification request can be granted or needs to be rejected due to the resource limitations.

The admission control uses the priority level information of the ARP to ensure that request of higher priority bearer is preferred. During the exceptional resource limitations, an EPS bearer's ARP pre-emption capability information indicates whether to drop bearers with lower priorities to obtain its required resources. An EPS bearer's ARP pre-emption vulnerability information determines whether a pre-emption capable bearer, with a higher priority ARP, can drop it. The ARP only impact bearer's admission decision, it does not effect on bearer level packet forwarding treatments (3GPP 23.401).

An EPS bearer is established when a UE connects to a PDN. It is referred to as default bearer. The default bearer is always allocated an IP address from the PDN. The default bearer remains established throughout the lifetime of the PDN connection and provides a UE with always-on IP connectivity to that PDN. It is always a non-GBR (best effect service) bearer. Any additional bearer of a UE that is established with the same PDN is referred to as a dedicated bearer that can be GBR or non-GBR. The dedicated bearers are established with different QoS parameters to provide QoS to different applications such as voice, video, games, email and ftp.

An EPS bearer carries traffic flow aggregate(s) consisting of one or more *Service Data Flows* (SDFs), each of which transports packets for a specific service such as video application. Each SDF is mapped to a bearer based on the traffic flow template. An EPS bearer's traffic flow template (TFT) is set of all packet filters associated with that EPS bearer. A packet filter is based on packet header characteristics such as source IP address, destination IP address, source port, destination port, source port range, destination port range, Type of Service (ToS), and protocol etc. Every dedicated EPS bearer is associated with a TFT. A default bearer may or may not have a TFT. It typically uses a "match all filter". Any SDF that does not match to any existing dedicated bearer's TFT is normally matched to the default bearer.

All SDFs mapped to the same EPS bearer receive the same packet forwarding treatment. So, separate bearers are required to provide different packet forwarding treatments. One bearer exists per combination of a QoS class and an IP address of a UE.

2.4 Layered Protocol Architecture in WiMAX Networks

WiMAX 802.16 standard defines only two lowest layers of the OSI model, Physical layer and MAC layer, which is main part of data link layer.

2.4.1 MAC Layer

MAC layer consists of three sublayers, Convergence Sublayer (CS), Common Part Sublayer (CPS) and Security Sublayer.

The **Convergence sublayer (CS)** is just above the MAC CPS and uses services provided by it. It accepts higher-layer packets also known as Service Data Units (SDUs). The standard defines CS specification for two types of higher layers, Asynchronous Transfer Mode (ATM) and the packet. The CS classifies and maps SDUs into appropriate Connection Identifier (CIDs), which is the basic function of the QoS of 802.16. It may also optionally perform the function of Payload Header Suppression (PHS). While performing PHS it suppresses the repetitive parts of the payload headers at a sender and restores them at a receiver.

The MAC CPS resides in the middle of MAC layer and performs the core MAC functions of system access, bandwidth allocation, connection establishment and maintenance. The standard provides management and transfer messages for communication between SSs and a BS. The management messages are exchanged before and during the establishment of connections. The transfer messages are used for the transfer of user's data, once the connection has been established.

The CPS receives data from various CSs, through the MAC Service Access Points (SAP), classified to a particular MAC connection. It applies appropriate fragmentation and concatenation over SDUs to form MAC Protocol Data Units (PDUs). Depending on the size of the payload, multiple SDUs can be carried on a single MAC PDU, or a single SDU is fragmented to carry over multiple MAC PDUs. The QoS is taken into account at this layer for the transmission and scheduling of data over the physical layer.

The MAC Security Sublayer handles security issues by providing Extensible Authentication Protocol (EAP) based authentication, secure key exchange, and Advance Encryption Standard – Counter with CBC-MAC (AES-CCM) - based encryption.

2.4.2 Physical Layer

Mobile WiMAX uses OFDMA, which is a multiuser version of OFDM and allows sharing the bandwidth among multiple users by doing multiplexing additionally in frequency domain. An OFDMA symbol structure mainly consists of 'data subcarriers' that carry user data; 'pilot subcarriers' that are used for synchronization and various estimation purposes like signal strength

estimation; and ‘null subcarriers’, which do not carry any transmission at all such as DC subcarriers together with guard subcarriers (used for guard bands).

In OFDMA, SSs are allocated resources in terms of “slots” or “subchannels”, as discussed in LTE physical layer. The composition of a subchannel that is the number and distribution of the subcarriers in a subchannel depends on the permutation mode. The permutation modes can be categorized into two major types:

1. Contiguous permutation: It takes subcarriers that are adjacent to each other and allows system to exploit multiuser diversity making it suitable for fixed environments.

2. Diversity permutation: It takes subcarriers pseudo-randomly to form a subchannel. It provides better frequency diversity, which helps in inter cell interference averaging making it suitable for mobile environments. It further includes two types of permutations, FUSC (Fully Used Sub-Carrier) and PUSC (Partially Used Sub-Carrier).

The exact definition of slots structure depends on the transmission direction and the permutation in use. In this thesis, our analysis for WiMAX employs PUSC permutation mode. In PUSC permutation mode on the DL, active subcarriers are grouped into clusters. Each cluster consists of 14 subcarriers (12 data and 2 pilot) over two symbols time. A slot consists of two clusters. In PUSC permutation mode on the UL, active subcarriers are grouped into tiles. Each tile consists of 4 subcarriers over 3 symbols time. Out of 12 subcarrier-symbol combinations, 4 are used for pilot and 8 are used for data. A group of 6 tiles form a slot in the uplink (So-In. et al., 2010).

Most of the WiMAX deployments are in TDD mode, so in this thesis we confine our discussion to TDD mode. In TDD mode, a frame comprises of DL and UL subframes separated by Transmit to Receive Gap (TTG) and Receive to Transmit Gap (RTG) as shown in Figure 2.7.

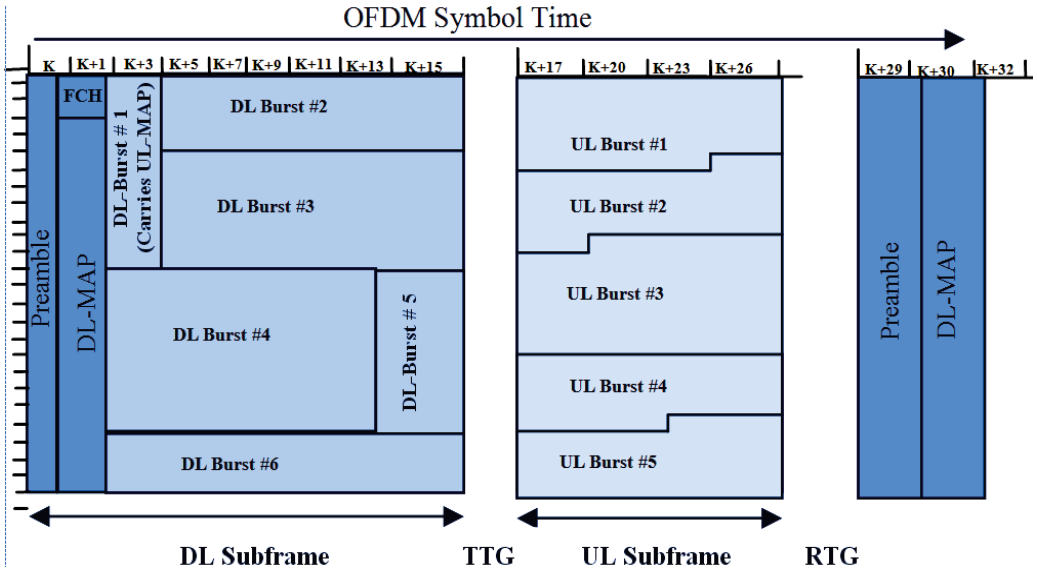


Figure 2.7 Frame Structure of Time Division Duplex (TDD) in WiMAX Networks

2.4.3 MAC and Physical layers Overhead

The overall overhead of a DL frame can be represented in the following equation (Yuehong et al., 2008).

$$OH_{DL} = P + FCH + MPDU(OH) + MAP(OH) + ACK_{DL} + PC_{DL} \quad 2.1$$

In Eq. 2.1, preamble (P) is the first OFDM symbol column in a TDD frame. It provides a mechanism for channel estimation and synchronization at the start of a TDMA frame and symbol time. The Frame Control Header (FCH) is immediately after the preamble to provide information to decode subsequent DL-MAP message. A 24-bit FCH is sent at the most robust Modulation and Coding Scheme (MCS), QPSK 1/2, so that all subscribers in a cell can receive it. The FCH carries information about sub channelization, coding and the size of DL-MAP. Before being mapped to FCH, 24 bit-DL prefix is duplicated to form a 48-bit block, which is the minimal Forward Error Correction (FEC) block size (802.16-2005, 2004).

$$FCH(OH) = \left\lceil \frac{FCH\ in\ FEC\ Block\ Size}{s_{QPSK1/2}} \right\rceil * r\ (in\ slots) \quad 2.2$$

In Eq. 2.2, ‘r’ is the repetition factor.

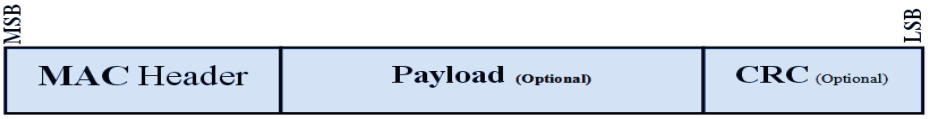


Figure 2.8 MAC PDU Format

Figure 2.8 shows the format of a MAC PDU (802.16-2005, 2004). A fixed length MAC header termed as Generic MAC header (GMH) of 6 bytes is added at the start of each MAC PDU and is optionally followed by the payload and a 4 byte CRC. The payload, if present consists of zero or more subheaders (fragmentation or packaging) each of 2 bytes, and zero or more MAC SDUs (802.16-2005, 2004). Hence, MPDU OH in Eq. 2.1 is estimated using Eq. 2.3.

$$\text{MPDU(OH)} = \sum_{i=1}^{N_{DL}} \left\lceil \frac{\text{GMH} + \text{CRC32} + \text{Frgamentation or Packaging header}}{s_i} \right\rceil * \text{r (in slots)} \qquad 2.3$$

In Eq. 2.3, N_{DL} is the number of users scheduled for downlink and s_i is the slot size with the MCS used for the i^{th} user.

In Eq. 2.1, ACK_{DL} is used to send ACK/NACK for uplink transmissions and on average uses 5 bytes (Yuehong et al., 2008). The PC_{DL} is used for fast power control and its size is set to 3 bytes per frame considering only open loop power control (Yuehong et al., 2008). In this thesis, overhead for ACK and PC is not considered. In Eq. 2.1, MAP OH indicates overhead of DL-MAP and UL-MAP messages. The MAP messages carry resource allocation information on a per connection basis. It can use either standard or compress overhead format. The two formats are discussed as follows.

Standard MAP format

Each standard DL-MAP and UL-MAP is of 104 bits and 64 bits long, respectively. It consists of information elements (IEs) of 60 bits and 32 bits for DL and UL, respectively. The MAP entries such as FCH are repeated 4 times and use only QPSK-1/2 MCS, causing a significant amount of overhead.

$$DL\_MAP(OH) = \left\lceil \frac{DL_{MAP} + \sum_{i=1}^{N_{DL}} DL\_MAP\_IE}{S_{QPSK1/2}} \right\rceil * r \text{ (in slots)} \quad 2.4$$

In Eq. 2.4, $S_{QPSK1/2}$ is the slot size for QPSK-1/2. The values of slots sizes for different MCS are given in (So-In. et al., 2010).

$$UL_{MAP(OH)} = \left\lceil \frac{UL_{MAP} + \sum_{i=1}^{N_{UL}} UL\_MAP\_IE}{S_{QPSK1/2}} \right\rceil * r \text{ (in slots)} \quad 2.5$$

In Eq. 2.5, 'N_{UL}' is the number of users scheduled for uplink.

Compressed MAPS with SUB-DL-UL-MAP

To reduce the overhead involved in MAP message, a base station in IEEE 802.16e can use compressed DL-MAP and UL-MAP. The compressed DL-MAP does not use GMH, so the amount of overhead reduces. The size of compressed DL-MAP and UL-MAP is reduced to 88 bits and 48 bits, respectively. A 32 bits CRC is added to the compressed map(s) that is computed across all bytes of the compressed map(s).

In IEEE 802.16e, a BS can further reduce overhead using Sub-MAP messages within the MAP message (Sanker et al, Oct. 15 2009). The SUB-DL-UL-MAP messages are used only with the compressed DL and appended UL MAP structure (802.16-2005, 2004). The basic idea is that while a BS broadcasts the main MAP message with the most robust MCS and high repetition factor, it multicasts Sub-MAPS using different MCS and repetition factors to different groups of users. In mobile WiMAX, a SS periodically communicates its channel conditions to a BS using Channel quality Indicator Channel (CQICH) in uplink. Based on the channel conditions of a SS, the BS determines its MCS. The BS groups SSs according to their MCS for the transmission of SUB-MAPS. There are some limitations on the use of SUB-MAPs like maximum number of SUB-DL-UL-MAP messages per 5ms frame is three. In addition to that SUB-MAP messages also creates fix overhead of 40 bits and is referred to as $SUB\text{-}MAP_{Fixed}$ in this thesis. Still in most of the scenarios it helps to reduce the overhead involved in the MAP message. A CRC-16 value is appended to the end of SUB-DL-UL-MAP message.

The compressed overhead (OH) is divided into two parts, fixed and variable. The main MAP part includes compressed DL-MAP and UL-MAP each followed by a 32 bits CRC. It is broadcasted using QPSK-1/2 with a repetition of 4.

Main MAP using Compressed OH

$$\text{Main MAP} = DL\_MAP(OH) + UL\_MAP(OH) \tag{2.6}$$

$$DL\_MAP(OH) = \left\lceil \frac{DL_{MAP}+CRC32}{s_{QPSK1/2}} \right\rceil * r \ \ (in\ slots) \tag{2.7}$$

$$UL\_MAP(OH) = \left\lceil \frac{UL\_MAP+CRC32}{s_{QPSK1/2}} \right\rceil \ * r\ (in\ slots) \tag{2.8}$$

The variable part of compressed OH includes SUB-DL-UL-MAP for each group modulated using QPSK-1/2 and with a repetition that depends on its group. The SUB-DL-UL-MAP comprises of fixed overhead of 40 bits, followed by a 16-bit CRC for each group as in Eq. 2.9.

Sub-MAPS using SUB-DL-UL-MAP

$$Sub\_MAP_{OH} = \left\lceil \frac{Sub\_MAP_{Fixed} + CRC16 + \sum_{i=1}^{N_{G\#}} DL\_MAP\_IE + \sum_{i=1}^{N_{G\#}} UL\_MAP\_IE}{s_{QPSK1/2}} \right\rceil * r \ \ (in\ slots) \tag{2.9}$$

In Eq. 2.9, 'r' is the repetition factor and $N_{G\#}$ is number of users in each group.

2.5 QoS in WiMAX Networks

The QoS support in WiMAX is provisioned at MAC layer of a BS and a SS. In IEEE 802.16, MAC uses connection-oriented approach and all data communication is in the context of unidirectional connection that is established between peer MAC entities at a BS and a SS. Each unidirectional connection is identified by a 16 bit identifier, CID. The MAC convergence sublayer associates packets traversing MAC interface with Service Flows (SFs) to be delivered over the connection. A service flow is the central element of WiMAX QoS framework. It identifies a unidirectional flow of packets between a SS and an ASN-GW with a particular set of QoS parameters and is identified by 32 bits Service Flow ID (SFID) (802.16-2005, 2004).

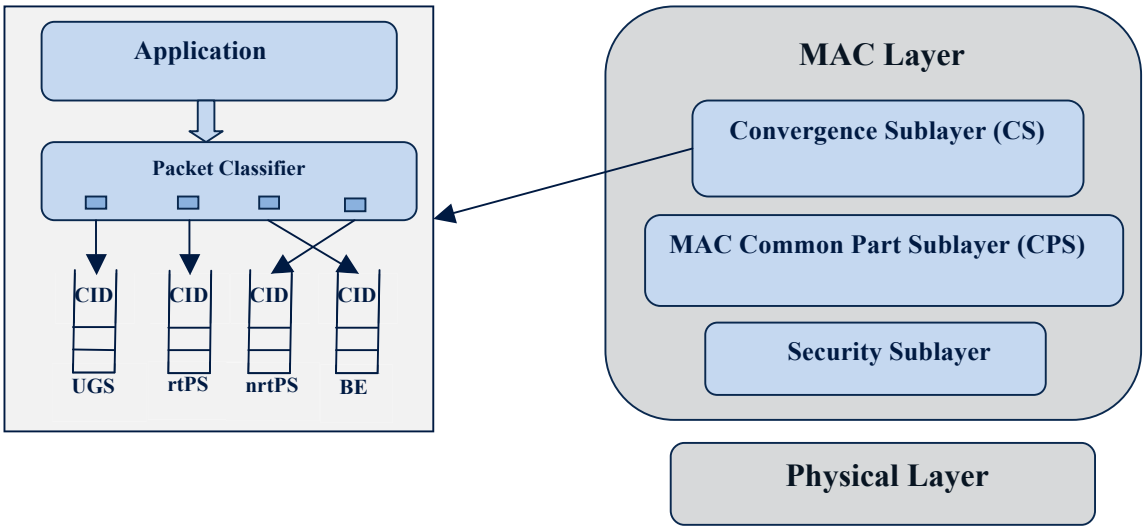


Figure 2.9 WiMAX QoS Architecture

Figure 2.9 shows that the packet classifier is use at the CS to map the packets to corresponding SFs based on some criteria such as five-tuples (source address, destination address, source port, destination port and protocol) and Differentiated Service Code Points (DSCP). For DL, packet classifier is employed at an ASN-GW and for UL it is employed at a MS. When a frame does not match to any criteria, the WiMAX system discards it.

In WiMAX networks, either the network that is a BS, or a SS initiates a service flow using MAC management messages. The MAC management messages include Dynamic Service Addition (DSA), Dynamic Service Change (DSC) and Dynamic Service Deletion (DSD) for SF addition, modification and deletion, respectively.

Similar to LTE default bearer, the WiMAX standard defines an Initial Service Flow (ISF), to establish IP connectivity during network entry before any packet transmission and reception starts.

To ensure the QoS, WiMAX standard has defined following 5 different Class of Services (CoSs).

Unsolicited Grant Service (UGS): The UGS Class of Service is designed to support applications that generate fixed size packets at constant bit rate (CBR). Example includes T1/E1 and VoIP. The QoS parameters of UGS CoS include Maximum Sustained Rate (MSTR), Maximum Latency tolerance and Jitter tolerance.

Extended Real Time Polling Service (ertPS): This CoS supports VoIP with activity detection. For ertPS connections, a base station reserves bandwidth during call setup. The QoS parameters of ertPS CoS include Maximum Sustained Rate (MSTR), Minimum Reserved Traffic Rate (MRTR), Maximum Latency tolerance and Jitter tolerance.

Real time polling service (rtPS): The rtPS CoS supports applications that generate variable size real time data on periodic basis. Example of such applications includes streaming audio/video. The QoS parameters of rtPS CoS are Maximum Sustained Rate (MSTR), Minimum Reserved Traffic Rate (MRTR), Maximum Latency tolerance and Jitter tolerance.

Non Real time polling service (nrtPS): The nrtPS CoS supports applications with variable size real time data on non-periodic basis with no delay guarantee. File download/upload is an example of such applications. The QoS Parameters of nrtPS CoS are Maximum Sustained Rate (MSTR), Minimum Reserved Traffic Rate (MRTR) and traffic priority

Best Effort (BE): This QoS class supports services such as web that neither requires a minimum data rate nor any delay guarantee.

2.6 Current Approaches for QoS Provisioning

Researchers have proposed various mechanisms to provide the QoS in 4G networks, which can be categorized as load balancing schemes; stringent RAC schemes; RAC schemes with preemption or degradation; combined RAC and scheduling approaches; and combined load control, RAC and scheduling approaches.

2.6.1 Load Balancing Schemes

To provide a satisfactory QoS, (Tung et al., 2008) proposed a dynamic call admission control and a QoS aware bandwidth allocation algorithm. The dynamic call admission control is based on

class of services and uses thresholds. The bandwidth allocation algorithm is triggered when blocking and dropping probabilities increase beyond the defined thresholds. It applies only to rtPS, nrtPS and BE CoSs. To apply these schemes, system needs to effectively define thresholds. Also, they do not consider the delay requirements.

The researchers (Li et al., 2011) proposed a congestion control for networks with delay sensitive and insensitive traffics. They suggested a new utility function, which considers both the rate and delay constraints of the traffic. The authors proposed a distributed algorithm, which allocates resources in a manner that the utility of voice and data traffic is equally optimized. In a network with the priority queuing, the algorithm can result in higher R-factor of voice traffic and higher throughput of data traffic at the cost of high packet delay of data traffic.

The authors (Chen and Khorasani, 2011) proposed a robust decentralized congestion control strategy for large scale fixed topology networks with differentiated services (Diff-Serv) traffic. The researchers (Chen and Khorasani, 2014) proposed decentralized and distributed Markovian jump - guaranteed cost congestion control (MJ-GCC) strategies for mobile Diff-Serv large-scale networks. They modeled the changes in the network topology caused by nodes' mobility using a Markovian jump process. By applying the GCC principle, the congestion controller for a node includes information from only the nearest neighboring nodes. The proposed distributed congestion control approach demonstrates an improvement over the performance of the decentralized control approach regarding the QoS performance and control efforts characteristics.

To ensure a congestion free system the authors (Al-Manthari et al., 2011) proposed an admission-level bandwidth management scheme. The scheme comprises of three components, a monitoring component, a CAC component and a dynamic pricing component. The monitoring component measures the available bandwidth in the network at the end of each time window. When the available bandwidth is different from the previously measured value, it invokes the CAC component. The CAC component estimates the optimum number of connections for each service class considering the maximum resource utilization and fairness among different service types. The dynamic pricing component estimates the prices of bandwidth units for each service type depending on the user demand so that the actual connections admitted in the new time

window are in the range of optimal values.

The decision as when to carry ASN-GW relocation is determined by (Liu. and Chen., 2012), which is based on traditional admission control and Wiener Process-based prediction algorithm. They proposed a Gateway Relocation AC (GRAC), which combines ASN-GW relocation and AC algorithms to maximize system capacity. Whenever, there are no resources for a newly arrived MS, the proposed GRAC requests an anchored MS to perform ASN-GW relocation provided there are anchored MSs in the system. The handover MSs are sensitive to call dropping and handover latency so they proposed a prediction algorithm based on wiener process to request anchored MS to perform the ASN-GW relocation early. Thus handover MS are not dropped when the system is overloaded. In addition, the handover MSs do not need to wait for the completion of ASN-GW relocation so handover latency reduces.

A base station load balancing scheme is proposed by (Casey et al., 2008) for mobile WiMAX. The proposed scheme states that when a base station is overloaded, it starts handover of overlapping terminals to the base station with light load until the resource utilization reaches an average value or the overlapping MSs have exhausted or a load balancing cycle ends. The scheme states that a base station is in overloaded state if its resource utilization exceeds a threshold, which is the sum of average resource utilization and a hysteresis margin. The results of the scheme show that setting a hysteresis at a very high value causes the system to remain in the load state for a very long time. Furthermore, if an upper limit is defined to initiate a load-balancing scheme, it works only for static environments.

The researchers (Vulkan. and Heder., 2011) presented a congestion control scheme to provide fairness in heterogeneous radio access networks comprising of legacy and flat system architecture with shared transport infrastructure. The proposed congestion control scheme alleviates congestion by selective drops at Radio Network Controller (RNC) in legacy system, and at an eNodeB in flat system architecture. The proposed scheme is applicable to provide fairness between legacy and flat system in DL and is limited to TCP based applications only.

A congestion control (CC) mechanism is proposed by researchers (Kwan. et al., 2010) that can be triggered due to congestion or by an AC pre-emption. The proposed load control procedure mitigate load in the network by removing low priority bearers until the system load reaches a

predefined target value. The priority of a bearer is determined based on its ARP value and load contribution. The paper does not discuss clearly how the target load can be defined for the network. The proposed scheme does not ensure fairness among bearers with the same ARP and the same load contribution.

The researchers (Qiu. et al., 2011) introduced a congestion control mechanism for LTE networks to protect an eNodeB output buffer overflow by controlling TCP advertisement window. The RLC layer at an eNodeB monitors buffer utilization and sets congestion flag once it reaches the threshold. On detecting congestion, the proposed CC looks in TCP header of UL ACK and reduces the value of advertisement window. The proposed scheme violates the protocol layer design principles. The proposed approach has effect for TCP based applications only.

To ensure that RBs allocated to users are fully utilized, authors (Zolfaghari. and Taheri., 2012) proposed queue aware scheduling technique in which a scheduler is also aware of packet availability in a queue during scheduling process. They presented performance of various queue aware scheduling schemes with an end-to-end congestion control scheme that controls the rate of elastic traffic and consequently affects the buffer status. The authors also proposed individual flow-based congestion measure for the wireless part of LTE networks to compel an end-to-end congestion control to follow the rates estimated by the scheduler. The scheme is proposed for DL only. The end-to-end congestion scheme used in their work involves the overhead of congestion field that is added in every packet. As in wireless network the resources are already limited so it will affect the overall performance of the network.

The above mentioned schemes to perform load balancing are either based on thresholds, or are applicable to a specific protocol, such as TCP only. The existing load balancing schemes merely discuss fairness among the flows of the same QoS class. This thesis proposes a congestion control scheme based on the queue length at an output buffer of a base station. It uses a queue control function instead of fixed thresholds to achieve an acceptable buffer delay. The proposed scheme neither involves changing the basic scheduling algorithm nor requires any additional field in the packets header to control congestion in the network. Rather, it adds a new module at a base station that estimates a rate for each type of service based on its QoS requirements and current output queue status to provide fair bandwidth allocation.

2.6.2 RAC Approaches

RAC approaches in the literature can be classified in different categories based on their admission strategies.

2.6.2.1 Stringent RAC Schemes without preemption/degradation strategy

The researchers (Anas et al., 2008) proposed a single cell Admission control (AC) algorithm. An incoming connection is admitted only if total number of Physical Resource Blocks (PRBs) per Transmission Time Interval (TTI) requested by an incoming and the existing connections is less than the total PRBs in the system bandwidth. The algorithm uses fractional power control formula agreed in 3GPP to calculate the number of PRBs required by new and active connections.

With the aim to recognize and priorities voice traffic in the packet switched networks, an AC with a voice classifier based on Deep Packet Inspection (DPI) is proposed by (Nageshar and van Olst, 2011). The AC reserves bandwidth for non-voice services to avoid their QoS degradation. It does not discuss the classification of other types of traffics.

The approach proposed by researchers (Antonopoulos and Verikoukis, 2010) is based on bandwidth reservation concept and is executed under busy hour condition. In the proposed approach, the bandwidth for UGS traffic is reserved only under busy hours condition, when the arrival rate of UGS connections exceed a specified threshold. The remaining bandwidth is provided to rtPS and nrtPS traffic. The BE connections are always accepted as they don't have any QoS requirements.

To enlarge the service area and to improve the wireless transmission quality, LTE- Advanced and IEEE 802.16j specifications approved the mobile multi-hop relaying (MMR). The researchers (Chang et al., 2013) proposed a Dynamic Cost-Reward-based (DCR) admission control for the MMR networks. It includes two main mechanisms: an exponential cost function based on Markov Decision Process (MDP) and a dynamic reward function, which cooperates with MDP cost function. The MDP cost function formulates the carrying cost of an incoming connection. The dynamic reward function utilizes different reward functions for different type of nodes (mobile station / relay station) and types of connections (real time or non real time). The

DCR admits the high priority connections that are with high reward.

The authors (Kaur and Selvamuthu, 2014) proposed Joint CAC for LTE-UMTS networks. The JCAC is proposed to balance the load between two interfaces, LTE and UMTS. When a request arrives it first determine its type of service. When the request is of voice, it tries to admit the call to UMTS provided its load is within the threshold range. In case the load of UMTS is higher than the threshold range, it tries to admit voice call to LTE system. Whereas, when the request is of data traffic, it tries to admit the call to LTE interface provided its load is within the threshold range. Otherwise, it tries to admit voice call to UMTS system.

The researchers (Chen, 2013) proposed an end-to-end measurements based admission control for VoIP networks. They considered two types of measurements methods, blocking percentage calculation (BPC) and blocking percentage calculation with instant feedback (BPC-IF). To avoid the drawbacks of probe-based approaches, the Endpoint admission control (EAC)-like scheme is utilized. It uses the dynamic threshold value, which is adjusted based on the passive measurements of the blocking rate received from each node. When a call request is admitted, ingress router selects the optimal path by calculating blocking rates across all source–destination paths. The work demonstrated that BPC is simple and scalable since the core nodes do not maintain the per-flow states. Whereas, BPC-IF has the scalability problem as the core must send up-to-date data to all ingresses.

The researchers (Bae et al., 2009) proposed a resource-estimated CAC algorithm to meet the QoS requirements for packet delay in LTE system. The CAC estimates the required number of PRBs for the requesting call based on the type of service and the current MCS of UE. At the time of connection request, the CAC estimates the available resources based on the PRBs usage of on-going calls. The CAC results in better performance in terms of packet delay but the average data rate and PRB utilization of the scheme is lower than the non-CAC.

The authors (Ovengalt et al., 2014) proposed a model based on a Type-1 Fuzzy Logic Controllers (FLCs). It is designed to enable the CAC to deal with uncertainty caused by the parameters such as the time-varying nature of wireless links, which affect the QoS of the real-time traffic. The inputs to the FLCs include latency and packet loss. Using these inputs the model

defines the attainable throughput of the wireless channel. The output of FLCs is the throughput / spectral efficiency of the channel, which is converted to estimate the exact amount of bandwidth required by each type of service. To guarantee the required throughput of a service, it performs channel aggregation where one or combination of channels is assigned to the UE.

The researchers (Ramkumar et al., 2012) proposed a crossed layer AC, which manages the dependence of user's QoS on the channel characteristics and queue characteristics. The proposed AC operates in two phases. In the first phase, it estimates the minimum required resources for a new user considering the effects of Adaptive Modulation and Coding (AMC) in the physical layer and queuing in the MAC layer. In the second phase the AC estimates the mean number of resources occupied by the existing users considering their buffer conditions.

To handle handoff connections, an adaptive RAC is proposed by (Chaudhry and Guha, 2007). It prioritizes handoff connections by reserving an adaptive temporal channel bandwidth based on the most recent requests. When the network is moderately loaded, the proposed dynamic guard channel scheme performed better than fixed guard channel scheme in terms of new connection blocking, handoff dropping probability and resource utilization.

These RAC schemes admit connections only when enough resources are available in the network. Consequently, in times of resource scarcity, the Blocking Probability (BP) for connections is very high with these admission schemes.

2.6.2.2 Threshold-based RAC Schemes

Researchers (Kwan, 2010) proposed a predictive AC scheme. It admits an incoming connection when load of all active connections and incoming connection is less than a predefined threshold. It uses different thresholds based on QCIs and request type. When the scheme is used with channel quality indication, overall cell throughput increases with high blocking rate at the cell edge and leads to unfairness in the network.

Admission control proposed by (Delgado and Jaumard, 2010) grants a new request only if the minimum throughput and the maximum delay requirements of already admitted bearers and incoming bearer can be guaranteed. It maintains a target blocking or dropping probability and a target system load. When the system is in load, it admits bearers based on their priorities. It confirms gain for the most sensitive traffic but does not show improvement in the total system throughput.

The researchers (Bae. et al., 2009) proposed a delay aware call admission control to provide service differentiation and to prevent congestion in the network. It maintains two thresholds for PRB utilization. It admits an incoming connection based on the values of thresholds. It updates the values of thresholds depending on the measured packet delay for each type of service and current PRB utilization at an eNodeB. It prioritizes handoff calls over new calls when the network encounters an increase in packet delay. The algorithm guarantees delay for each service type but with lower utilization of the resources.

The authors (Lei. et al., 2008) proposed a Call Admission Control (CAC) based on the RB usage, as it's easy to implement and can adaptively adjust thresholds for different users and traffics based on the network condition.

The authors (Lim et al., 2013) proposed a joint uplink/downlink connection admission control scheme for WLAN/cellular integrated systems. The scheme adaptively determines the admission control thresholds based on the uplink and downlink bandwidth characteristics of cellular networks and WLAN, and the bandwidth requirements of services. To evaluate the performance, they developed a multi-dimensional Markov chain and derived the blocking probability and the total blocking cost.

The RAC schemes, which utilize thresholds to differentiate between services often lead to inefficient utilization of the network resources.

2.6.2.3 Reservation-based RAC to handle channel fluctuations

Due to user mobility or channel quality variations, during the connection holding time, a bearer can demand extra resources beyond the amount allocated by the RAC procedure. To avoid

QoS degradation of ongoing sessions, several researches suggested admission control schemes that limit the admission rate by reserving additional resources.

The researchers (Lakkakorpi and Sayenko, 2009) proposed two admission schemes to control the number of real time connections in the network. In the first scheme, it utilizes the averaged number of free slots as an input to decide the admission of real time connections. To admit connections arriving in batches, the second method is employed, which periodically adjusts the admission thresholds according to the current load of the network. The authors (Ukil and Sen, 2010) introduced a proactive resource reservation scheme, which accepts or rejects an incoming connection based on the congestion notification sent by the traffic prediction module. The traffic prediction module estimates future traffic based on the current QoS states and the QoS demand of an incoming connection.

The researchers (Mehdi. et al., 2012) proposed an admission scheme that reserves additional resources at the time of connection admission to offset the changes in radio resource demand. To determine the amount of extra resources needed, it uses an average of extra resources utilized at the end of each mobility period.

In the 4G networks when the resources are available, an incoming connection is also granted resources to meet its maximum bit rate. The network is required to guarantee only the minimum rate of a connection, so the resources allocated above the minimum rate can be effectively used to handle the channel fluctuations. The schemes proposed by (Mehdi. et al., 2012, Lakkakorpi and Sayenko, 2009, Ukil and Sen, 2010) while reserving extra resources for the connections do not take into account these additional resources allocated to the existing connections above their minimum rate requirements. So, these schemes are not directly applicable to 4G networks where they can over reserve the resources to deal with the channel fluctuations and results in an increase in connection's BP.

2.6.2.4 RAC Schemes with preemption or degradation strategy

In times when the network resources are limited, to admit connections of high priority an admission control can degrade the rate allocated to the existing connections. Additionally, it can

pre-empt the existing connections to attain the resources for new connections of high priority services.

The researchers (Wang et al., 2005) proposed a dynamic CAC scheme for fixed WiMAX networks. It is based on reservation and degradation. It reserves resources for high priority UGS connections. To reduce blocking probability of rtPS and nrtPS connections, it reduces the bandwidth of only over provisioned nrtPS connections. The step size of degradation is assigned a fixed value of 32 kbps.

To provide service differentiation and to maximize the system revenue, the authors (Hou et al., 2006) introduced a RAC, which exclusively reserves bandwidth for connections of high priority CoSs such as UGS and rtPS. On the arrival of UGS, rtPS and nrtPS connections, if resources are not sufficient, degradation is applied to connections of nrtPS and BE CoSs only. They do not discuss the value of the step size of degradation.

The authors (Wang et al., 2007) proposed a RAC scheme that admits new connections only if the system has enough resources. With handover connections, if the resources are not sufficient, it applies degradation on existing connections of lower and same priority CoSs. The degradation process reduces the bandwidth of a connection to its minimum reserve rate. The degradation is not equally applied on all connections of a CoS that leads to unfairness in network.

For handover connections, the authors (Ge and Kuo, 2006) proposed a guard channel scheme. When the resources are not enough, it applies a bandwidth borrowing on connections of lower and same priority CoS to minimize blocking and dropping probabilities. It does not apply a step-wise degradation but it reduces the bandwidth of a connection directly to its minimum reserve rate.

The researchers (Murawwat et al., 2009) proposed a CAC based on degradation. It degrades the bandwidth of all nrtPS connections to their minimum reserve rate only when a request of high priority CoSs such as UGS and rtPS arrives and resources are not adequate in network. The scheme also proposed to upgrade nrtPS connections when the network is lightly loaded. The proposed degradation does not provide fairness among the connections at the same priority level.

The authors (Suresh. et al., 2008) proposed a RAC that satisfies both bandwidth and delay requirements to admitted connections. For new and handoff UGS and rtPS connections, if enough bandwidth is not available in the network, a bandwidth degradation process is applied to nrtPS connections. If the requested bandwidth is not gained, the degradation is applied on the admitted rtPS connections only for handoff UGS and rtPS connections. The step size of degradation is assigned a fixed value of 256 and is same for each CoS. After allocating the bandwidth, the delay requirements of rtPS connections is investigated. A new connection is rejected if it effects the delay requirements of existing rtPS connections.

To avoid resource starvation, authors (Jiang and Tsai, 2006) introduced a RAC that uses a combination of thresholds and degradation schemes. The proposed mechanism defines threshold for each CoS. To minimize blocking probability, if the resources are not sufficient in the system, degradation is applied to connections of lower and same priority and even high priority CoSs only if connections of the respective CoSs are using bandwidth above their respective thresholds. The proposed degradation process leads to unfairness among the connections of a CoS as to steal the resources from a CoS above its threshold, degradation is applied only to some connections of the respective CoS. Furthermore, they do not clearly define the degradation step.

The authors (Luo et al., 2009) proposed a CAC scheme that defines a priority table and degradation groups based on the priority and precedence values in the CID field. The scheme defines threshold for each CoS as in (Jiang and Tsai, 2006). When a CoS is above its threshold, any incoming connection of that respective CoS is marked as downgraded flow that can be preempted when the network is congested.

The researchers (Khabazian et al., 2012) presented a pre-emption based admission scheme with two phases. In the first phase, it applies a partial pre-emption to obtain the extra-required resources. For partial pre-emption it defines a contribution metric for each bearer, which takes the bearer's priority and the amount of over-provisioned resources as inputs. The partial pre-emption degrades the bandwidth allocated to bearers until their minimum guaranteed bit rate. The second phase of full pre-emption is applied only when enough resources cannot be obtained after the first phase completes. In the second phase, resources allocated to bearers are fully pre-empted one-by-one until the required resources are obtained. The partial pre-emption scheme

followed by the full pre-emption scheme improves fairness among the connections at the same priority level compared to the pre-emption scheme in (Kwan. et al., 2010), but still it results in unfairness to the connections with low priority.

The researchers (Khabazian et al., 2013) proposed an improvement in the scheme proposed by (Khabazian et al., 2012). In the proposed bandwidth adaptation scheme the contribution metric includes three parameters namely bearer's priority, QoS over-provisioning and the bearer's channel quality. The proposed full pre-emption schemes in (Kwan. et al., 2010, Khabazian et al., 2012, Khabazian et al., 2013) are unfair to bearers of low priority as they are altogether removed from the network. Also, starting from the lowest priority, it removes bearers one-by-one until the target resources are realized. So, it is possible that at the end of pre-emption, at a certain priority level some connections are removed while other connections with the same priority are still in the network. Hence, the proposed full pre-emption is unfair to connections at the same priority level.

The authors (Priya and Franklin, 2012) proposed A Dynamic Bandwidth Allocation (DBA) based predictive RAC. When the resources are limited, it uses DBA to degrade the connections of Non Real Time (NRT) services to their minimum to accommodate an incoming service. It also employs a DBA departure algorithm that allocates the bandwidth, which becomes available after the departure of a connection, to the existing connections to the increase system utilization.

The researchers (Qian et al., 2009) proposed a RAC that uses service degradation scheme to prioritize bearers with high priority. The degradation schemes by (Priya and Franklin, 2012, Qian et al., 2009), starting with the lowest priority degrades bearers one-by-one to their minimum rate until the required resources are obtained. It is possible that at the end of degradation, some connections are degraded to their minimum rate while other connections at the same priority level are still over-provisioned. Hence, these proposed degradation schemes are unfair to the connections at the same priority level. Also, they use the same degradation level and degrade connections at any priority level instantly to their minimum rate. Consequently, during the degradation, they do not provide differentiation among the connections at different priority levels. Also, the DBA departure algorithm proposed by (Priya and Franklin, 2012) does not discuss the fairness among the connections at the same priority level.

The authors (Borodakiy et al., 2014) proposed a pre-emption based AC for two guaranteed bit rate services, video conferencing (VC) and video on demand (VoD). The scheme defines a higher priority level for VC compared to VoD. In case of resource limitations, the AC allows an incoming high priority VC connection to pre-empt the low priority VoD connections, which are selected randomly. The scheme in times of limited resources also allows the VC quality degradation from a high to standard definition to allow the admission of low priority VoD connections. The authors employed a recursive algorithm to calculate the system probability distribution to analyze the model performance measures.

The authors (Tsiropoulos et al., 2011) proposed a probabilistic framework for CAC for the Next Generation Networks (NGN). The CAC admits the low priority service classes (SCs) with a variable imposed probability. The probabilistic framework is considered under a bandwidth-centric approach named probabilistic bandwidth reservation scheme (PBRS), which is based on the total number of bandwidth units (BUs) occupied in a cell. By utilizing markovian chains the multiple SCs, which corresponds to different call specifications, are supported and analytical expressions for the call blocking probabilities (CBPs) are derived. The PBRS adjusts the admission rate of the low priority SC calls according to traffic variations. When the network traffic increases, the admission probability of low priority SC calls is reduced. Thus, as the available network resources are reduced, the number of low priority calls is gradually reduced. Whereas, calls of high priority SCs are always admitted unless all BUs are occupied.

The authors (Carvalho et al., 2013) proposed a Joint Call Admission Control (JCAC) for the next generation wireless networks (NGWNs). The JCAC decides admission of a service request based on the admission constraint. It also determines the radio access technology (RAT) to which the new request will be connected. They proposed a cost function, which makes use of a blocking cost function and an alternative acceptance cost function for the optimal decision. The JCAC supports both real time and non real time services. It selects the biggest RAT for real-time service class and the smallest one for non-real-time service class. The Bandwidth Adaptation (BA) mechanism can increase the system capacity, by reducing the blocking probabilities. Consequently, the authors for the future work considered developing an integrated JCAC and the Bandwidth Adaptation (BA) for NGWNs.

The authors (Ivesic et al., 2014) proposed a resource management mechanism to improve session establishment success and the network resource management with acceptable end-user quality of experience level. They proposed two approaches, an admission control and resource allocation. The proposed admission control utilizes Media Degradation Path (MDP) to perform the admission decision. In situations when available resources are not sufficient for the optimal configuration of the new multimedia session, the proposed admission control selects an alternative configuration from the MDP. Consequently, it increases the admission probability and also provides user satisfaction at an acceptable level. The proposed MDP-based resource allocation procedure deals with the problem of resource allocation in times of reduced resource availability in the network. It degrades sessions to less resource demanding configurations.

2.6.3 Scheduling Approaches

The authors (El-Shinnawy et al., 2010) proposed scheduling algorithm that switches between different scheduling criteria's so that a WiMAX scheduler can consider multiple QoS aspects to ensure the QoS in an overloaded system. It utilizes an algorithm, which employs multi-queuing system. So, defining the purpose of each algorithm and condition to switch between them is very crucial. It defines two layers of priority. In the first layer, priority is given to any flow that needs resources to meet its minimum requirements. In the second layer, the algorithm allocates the resources to any flow that needs to maximize its throughput.

The authors (Fang-Chang et al., 2012) proposed a resources allocation scheme for the Uplink of LTE networks. The scheme with the aim to improve the user's satisfaction allocates RBs to the UEs according to the data rate granted by the CAC. The proposed resource allocation scheme prioritizes and selects the users for resource allocation based on the shortage ratio of the average data rate. To avoid wastage of the RBs, the scheme considers the queue length of UE reported in BSR.

2.6.4 Combined Load Control, RAC and scheduling approaches

The researchers (Rodrigues and Cavalcanti, 2008) introduced a QoS-driven adaptive framework that combines functionalities of load control, admission control and scheduling for

High Speed Packet Access (HSDPA) system. The CC framework considers VoIP frame erasures (FER) as the measure of QoS, and does not directly consider the resources available in the network. The load control function depending on whether the VoIP FER is less than or greater than the target value, adjusts the parameters of RAC and packet scheduler to define the priority level between VoIP and other low priority services. The proposed scheme manually selects the values for the step sizes of degradation or upgradation; and the parameters of RAC and scheduler.

The authors (E. O. Lucena et al., 2010) proposed the generalization of CC framework proposed by (Rodrigues and Cavalcanti, 2008) to work with OFDMA. They mainly proposed a delay based prediction mechanism to avoid high peaks of FER. They proposed a mechanism for early detection of overload situation based on the packet delay of RT flows in a framework called Delay-Based Prediction. The scheme updates parameters of RAC and packet scheduler based on the increasing delay information from the delay based prediction framework. The values of thresholds and other parameters such as the step sizes of degradation/ upgradation are defined in the manual way same as in (Rodrigues and Cavalcanti, 2008). These schemes do not consider the rationale behind the selection of parameter values and the effect of parameter settings on the system performance.

Our proposed RAC scheme does not define thresholds for different QoS classes. To differentiate among different QoS classes in times of resource scarcity, it applies bandwidth-borrowing scheme. The degradation step size is defined based on the current network resource utilization and the resources requested by an incoming connection. The degradation is applied to obtain only the requested resources. To deal with channel fluctuations, it reserves additional resources with an incoming connection. When performing extra resource reservation, it considers the resources, which are already allocated to the existing connections above their GBR requirements. Consequently, it avoids over reservation of the resources.

In this thesis, we proposed a novel QoS framework for the 4G networks. The framework includes a new load control mechanism and an intelligent admission control. Compared with existing QoS schemes, our proposed algorithms do not rely on thresholds to perform reactive operations such as dropping existing connections, restricting admission of certain QoS classes or adapting the parameters of admission control and scheduler etc. Our proposed control algorithms

are proactive and allow the network to employ the same control policies in all network situations. They enable the network operators to ensure stability in the network. This thesis presents a comprehensive framework, which requires the congestion control, admission control and scheduler to coordinate to effectively perform fair resource allocation in the network, as introduced in 0.

2.7 Summary

In this chapter, from the system architecture, layers profile, and the QoS architecture, we then moved to the characteristics of QoS. Finally, recent studies in the literature on the QoS in 4G networks or related scenarios are reviewed. We introduced some related work and pointed out the differences and features of our framework. By investigating the QoS issues in 4G context and current approaches, we analyzed the causes of these issues and illustrated desirable properties and objectives for our proposed QoS framework for the 4G networks.

Chapter 3 Proposed QoS framework and Control Algorithms for 4th Generation Networks

Existing QoS schemes are far from optimal in several key aspects; most of the schemes utilize thresholds to detect congestion, employ different control schemes in states of congestion and non-congestion and do not provide fair resource allocation in the network. In this chapter, we provide an overview of our proposed QoS framework and control algorithms for the 4G networks. The novelty of this framework compared to other mechanisms and technologies is that our framework allows the network operators to exercise controls and to ensure the QoS of the admitted traffic flows when the traffic demands at the access side and the load at the core side of the network vary dynamically. It efficiently prevents congestion and delivers QoS to users in terms of fair resource allocation, throughput and delay. The main components of the QoS framework include: the Congestion Control (CC), the Radio Admission Control (RAC) and the scheduler.

In accordance with the key issues discussed in Section 1.2, and the desirable properties and the objectives of the QoS framework described in Section 1.3, we propose a comprehensive QoS framework adapted to the 4G networks. In this chapter, Section 3.1 provides an overview of the proposed QoS framework and control algorithms. Section 3.2 introduces an intelligent CC algorithm for the 4G networks. It deals with both the unfair bandwidth allocation and the congestion issues encountered in the 4G networks. Section 3.3 describes an intelligent admission control algorithm for the 4G networks. The RAC intelligently admits or rejects a connection to manage load in the network. Section 3.4 briefly discusses the required enhanced features of a scheduler. These schemes form the basis of the control algorithms designed to address the specific QoS architecture of each 4G LTE and Mobile WiMAX technologies.

3.1 Proposed QoS Framework

In this thesis, we propose a QoS framework and load control algorithms for the 4G networks, mobile WiMAX and LTE. The framework includes a new load control mechanism, the Fair

Intelligent CC based on the QoS architecture of the 4G networks (4G-FICC). It avoids and controls congestion at the base stations of WiMAX and LTE networks, respectively. 4G-FICC is always active in the network, whether the network is overloaded or underutilised. It avoids congestion through the traffic balancing and minimises the resource underutilisation. It defines a target operating point and maintains the network traffic around this point. It estimates the fair share of bandwidth for each Type of Service (ToS) based on its current usage of resources, QoS constraints and load at the network. It ensures fairness is guaranteed among the traffic flows, without violating the QoS requirements of the connections. The scheduler allocates resources based on the feedback from the load control module, 4G-FICC. The aim is to maintain the network traffic load around the target operating point to minimise delay and stabilise throughput.

Moreover, to ensure end-to-end delay and the QoS, we propose a predictive RAC, the Fair Intelligent Admission Control for the 4G networks (4G-FIAC). It admits or rejects an incoming connection based on the resource availability and the current load in the network. The key idea is to utilise feedback from the load control module to determine the load in the network. 4G-FIAC is based on the bandwidth borrowing and applies a step-wise degradation on the over provisioned connections in order to minimise blocking probability and to maximise resource utilisation in the network. So, 4G-FIAC along with 4G-FICC avoids congestion in the network to guarantee QoS to end users.

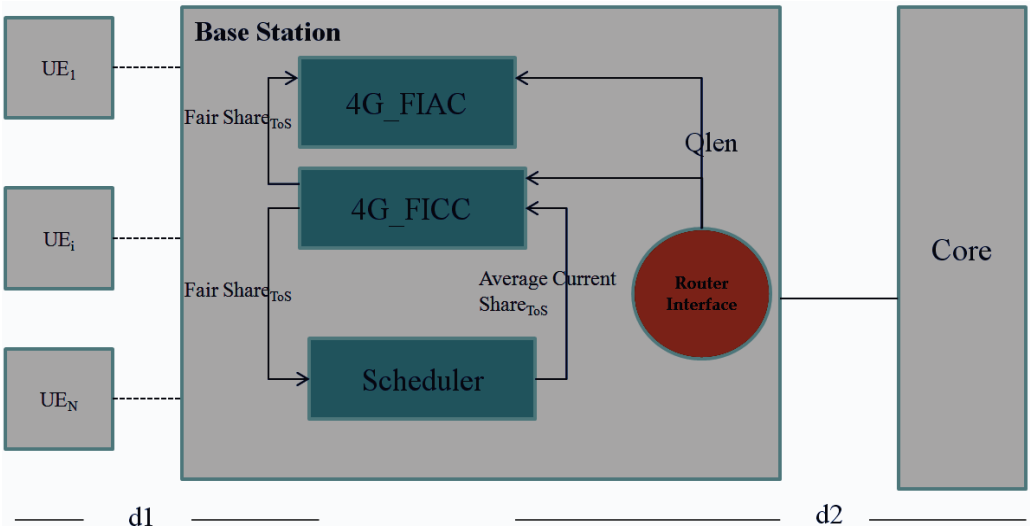


Figure 3.1 Proposed QoS Framework

3.2 Fair Intelligent CC for 4G Networks (4G-FICC)

In this section, we introduce an intelligent CC scheme for 4G networks, mobile WiMAX and LTE. It is based on the basic principles of the explicit feedback CC, the FICC introduced by (Phan. and Hoang., 2005) for wired networks. FICC main idea is as follows.

3.2.1 Description of FICC

Normally, CC schemes employ thresholds to define congestion and non-congestion states. Importantly, these schemes employ different rate allocation schemes in these two network states. In FICC, instead of the fixed thresholds, a target operating point is employed. When the network operates below the target point, it is regarded as in non-congested state. Otherwise, it is in congested state. FICC employs similar and consistent rate allocation policies for both, the congested and the non-congested states.

In FICC, egress router uses the Resource Discovery (RD) protocol to probe the available network resources from the routers inside the Diff-Serve domain. The main aim of FICC is to calculate an Expected Rate (ER), which reflects how much traffic could be handled by a transit router within the network. The calculated ER is sent as a feedback to the source nodes. The sources set their rates to the ER as indicated in the feedback. The ER is a function of an optimal class rate and the current queue utilization.

Main features of FICC are as follows. Firstly, it tries to maintain the queue length (Qlen) at a router close to the target operating point (Q_0), so neither the queue reaches the buffer's maximum capacity, nor it becomes empty. Consequently, the link is never idle unnecessarily. Secondly, FICC adjusts an Allowed Class Rate (ACR) to reduce variations in the buffer queue length and in packet delays. FICC employs an efficient function of queue size called queue control function (f(Q)) given in Eq. 3.1. The queue control function encourages traffic sources when the queue length is less than the target operating point and discourages sources if the switch operates beyond the target operating point. Thirdly, it solves unfairness problem, as it shares bandwidth equally among the connections of the same class and distributes the left over bandwidth fairly among the connections that can use the additional share. To achieve this objective, FICC

oversells the bandwidth to connections in need when the network operates below the target operating point. Fourthly, FICC maintains per output queue information to make it more scalable. It is kept simple to reduce its implementation complexity.

FICC: At each router interface

```
IF (receive RD (CCR,ER,DIR) == forward)
   IF (Qlen > Q0)
       IF (ACR < MACR)
               MACR=MACR + β * (ACR-MACR)
       End IF
   Else
       MACR=MACR + β * (ACR-MACR)
   End IF
Else IF (receive RD (CCR,ER,DIR)== backward)
   IF (Qlen > Q0)
       f(Q) = (Buffer_Size−Qlen) / (Buffer_Size−Q0)
   Else
       f(Q) = (α−1)∗(Q0− Qlen) / Q0 + 1
   End IF
End IF
ER = Max (MinER, min(ER, f(Q) * MACR))
```

3.1

3.2.2 Design Goals of FICC for 4G Networks (4G-FICC)

We aim to develop a CC scheme for the 4G networks. It must be designed to match the QoS structure of the 4G networks, including mobile WiMAX and LTE/LTE-Advanced. The resources are limited in the 4G networks, so it must prevent overhead involved in the resource discovery feedback in the basic FICC. Additionally, it must provide service differentiation and fair resource allocation to connections with different service types. The main ideas behind the proposed CC scheme are 1) maintains information per QoS class and avoids per connection accounting to keep

the algorithm simple 2) must be scalable to effectively manage the dynamic changes in load of the network and 3) avoids resource underutilization.

In this section the development of the new CC scheme for the 4G networks is discussed with respect to the above discussed design goals.

Its key features are summarized in this section.

1. In 4G networks, the output queue status of a base station depends on the output link capacity as well as the load at the core network. So, a large queue length at the base station's buffer serves as an indication that the core network is congested. Instead of using fixed thresholds to detect and to control congestion at the core network, FICC for the 4G networks defines a target operating point at an output buffer of a base station. It employs an efficient function of queue size called queue control function (f(Q)) same as in Eq. 3.1. Its aim is to maintain the network traffic around the target point to achieve the maximum resource utilization and reduced delay.

2. 4G networks are constraint by limited spectrum availability, so opposed to the basic FICC scheme, the proposed FICC scheme for 4G architecture does not employ the resource discovery feedback control mechanism. 4G-FICC is employed as a separate module in a base station to help manage the scarce and valuable bandwidth resources. It sends the feedback to the scheduler on a base station, which grants resources to connections. In this way, it avoids overhead involved in the feedback mechanism.

3. In basic FICC for wired networks, there are no minimum or maximum rate constraints. Therefore, it can adjust the ER without any limit to manage the load and to keep the network traffic around the target operating point. The 4G standards, IEEE 802.16 and 3GPP, define different QoS classes (QoCs) to ensure QoS for each service type (sections 2.3 and 2.5). FICC for the 4G networks takes into consideration different QoS parameters of each QoC defined by the network operators such as the maximum rate, the minimum rate, delay and Packet Error Rate (PER).

3.2.3 Estimation of Expected Rate of Each QoS Class (ER_{QoC}):

In order to keep the queue length around the target operating point, 4G-FICC based on the average bandwidth allocated to the connections of each QoS class estimates the expected rate for each QoS class (ER_{QoC}) using the queue control function. The ER_{QoC} indicates the fair share of bandwidth of each QoS class. It reflects how much traffic a BS can handle due to the load at the core network and is defined as follow.

$$ER_{QoC}(t) = MACR_{QoC}(t) * f(Q) \quad 3.2$$

In Eq. 3.2, $MACR_{QoC}$ is the Mean Allowed Class Rate of each QoS class. It indicates the average resource allocation to the connections of each QoS class. 4G-FICC maintains information per QoS class. Consequently, it utilizes one MACR for each QoS class ($MACR_{QoC}$). To obtain accurate MACR per QoC, the scheduler at the BS updates the $MACR_{QoC}(t)$ as follows.

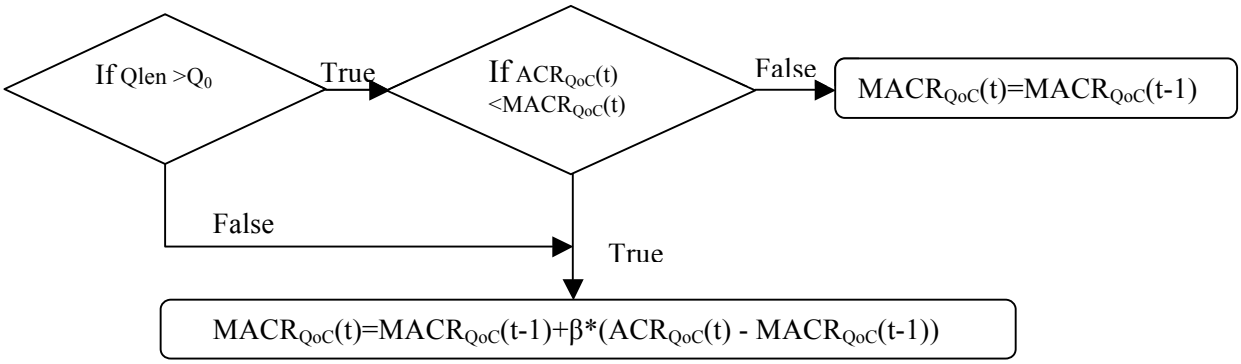


Figure 3.2 Calculation of Mean Allowed Class Rate per QoS Class ($MACR_{QoC}$)

In Figure 3.2, 'β' is an exponential average factor. It determines how fast the $MACR_{QoC}(t-1)$ converges to the $ACR_{QoC}(t)$. The $ACR_{QoC}(t)$ is an Allowed Class Rate, which represents the actual rate allocated to an active connection of a specific QoC in time (t). Thus, the $MACR_{QoC}(t)$ reflects the mean of the allowed class rate of all active connections per QoS class.

3.2.4 Restriction on Expected Rate (ER) of each QoS Class

In the basic FICC scheme, there is no restriction on reducing or increasing the ER to enable the network to maintain the queue length around the target operating point. Whereas, in 4G networks, the CC ensures that the ER of a QoS class must not be lower than its minimum rate to

preserve its minimum throughput and delay requirements. Additionally, 4G-FICC does not allow increasing the ER of a QoS class above its maximum rate.

3.2.5 Queue Control Function (f(Q))

4G-FICC utilizes the same function of queue size as employed by the basic FICC given in Eq. 3.1. The queue control function expresses the degree of network congestion that a base station can tolerate. For simplicity of implementation this function uses a linear function to maintain the queue operation around the target operating point (Q_0). The linear function as a candidature for 4G-FICC scheme is defined in Figure 3.3.

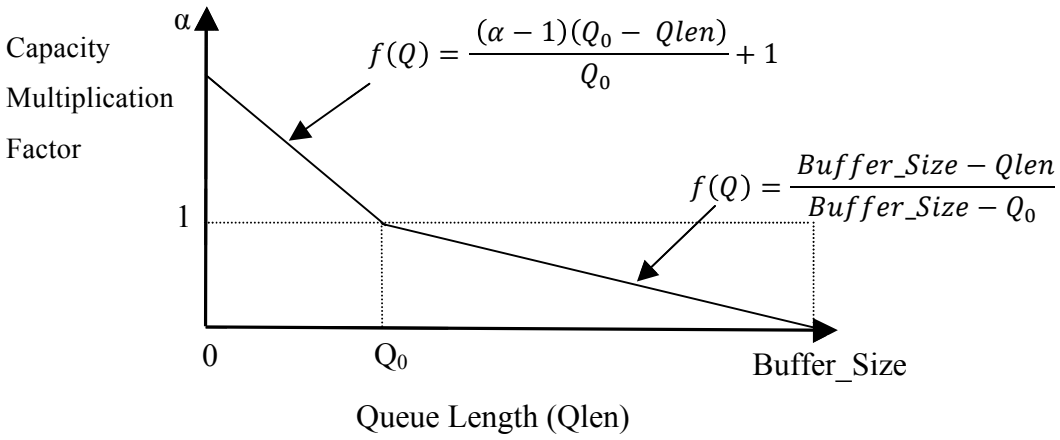


Figure 3.3 Queue Control Function (f(Q))

In Figure 3.3, 'Buffer_Size' is the total size of the buffer at the base station, 'Qlen' is the current queue length at the BS output buffer and Q_0 is the target operating point. The parameter 'α' can be considered as an oversell factor when the network is underutilized. The value of oversell factor (α) determines how much traffic is encouraged to put inside the network when it is underutilized. It must has a value higher than 1 to allow an increase in the ER.

Function of queue is a linear function with values between α to 1 for queue lengths in range of [0, Q_0], and between 1 to 0 for queue lengths in range of [Q_0, Buffer_Size]. The two lines intersect at Q_0, where the value of f(Q) is 1. Consequently, when the queue length is less than the target operating point (Qlen < Q_0), the ER of each service type increases from $MACR_{QoC}$ up to either the value of α*$MACR_{QoC}$, or the maximum class rate, whichever is the minimum. When

the queue length is greater than the target operating point (Qlen > Q_0), the expected rate of each service type reduces from $MACR_{QoC}$ to the maximum of either '0' or the minimum class rate (i-e max (0, minimum class rate)).

3.2.6 Step-Wise Degradation and Upgradation

In times of congestion in the core network, instead of pre-empting ongoing sessions or degrading their rate directly to the minimum rate, 4G-FICC applies a step-wise degradation to keep the queue length close to the target point. In situations when it estimates that the network is underutilized, it stepwise upgrades the rate of each QoS class with a view to ensure stability in the network in terms of maximum throughput and minimum delay. The step size of degradation and upgradation is determined by the function of queue. By applying step-wise degradation or upgradation, it provides fair resource allocation among the connections at the same as well as different priority levels.

Following subsections discuss the way 4G-FICC operates to provide fair resource allocation in the network.

3.2.7 Fair Resource Allocation among Flows of Different QoS Classes

To ensure service differentiation among connections of different QoS classes, it considers the priority of connections while estimating the expected rate of each QoS class.

In times when there is congestion, the queue control function returns a value less than 1 indicating the network is operating beyond the target point. In this situation, 4G-FICC starts control from the lowest priority QoS class and step-wise reduces its ER. When the ER of the lowest priority QoS class reaches its minimum, it then moves to the next lower priority QoS class. It keeps on moving to the next lower priority QoS classes and reduces their expected rate until either the value of function of queue becomes greater than or equal to 1 or the ER of all QoS classes reaches their respective minimum rates. In this way, it ensures that at the end of the degradation procedure more degradation is applied on the ER of lower priority and over provisioned QoS classes. As a result, it guarantees fair resource allocation among the connections of different priority classes.

Additionally, in times when the network is underutilized, f(Q) returns a value greater than 1 indicating the network is operating below the target operating point. In this situation, 4G-FICC starts operation from the highest priority QoS class and step-wise upgrades its ER. Once the ER reaches the class maximum rate, it moves to the next higher priority QoS class. It keeps on moving to the next higher priority QoS classes and increases their ER until either f(Q) becomes less than or equal to 1 or the expected rate of all QoS classes reaches their respective maximum rates. In this way, it ensures that at the end of the upgradation procedure more upgradation is applied on the ER of higher priority and less provisioned QoS classes. Consequently, it guarantees fair resource allocation among the connections of different priority classes.

To estimate the new expected rate 4G-FICC does not consider the channel conditions of individual users. It is because the scheme is kept simple and estimates the expected rate per QoS class rather than per connection.

3.2.8 Fair Resource Allocation among Flows of the Same QoS Class

It is necessary that all users of the same QoS having the same QoS requirements be treated fairly. Fairness determines that users are getting the fair share of resources and hence the same throughput in the network. The 4G-FICC scheme reduces or increases the ER of each QoS class depending on the congestion state in the network. It passes the estimated expected rate of each QoS class to a scheduler making it flexible and easy to operate with any underlying basic scheduler at a base station.

The scheduler allocates resources to flows of the same QoS class corresponding the expected rate of the respective QoS class. To attain the same expected rate, the scheduler assigns higher resources to connections having bad channel conditions compared to users with good channel conditions. Consequently, 4G-FICC manages load and provides the fair share of bandwidth to flows of the same QoS class.

The proposed 4G-FICC takes into account the current usage of the network resources, the load at the core network and the QoS attributes of different QoS classes, in determining the fair share of the bandwidth of each service type. As a result, it ensures efficient buffer occupancy in the network as each service in the network occupies the fair share of the buffer.

4G-FICC Degradation Procedure

```
IF f(Q) < 1 // indicate congestion in the network
   i= index of the minimum priority QoS class
   While f(Q) < 1
         IF (ER_i > Minimum_rate_i)
               Degrade Expected rate of 'i' QoS Class
         Else IF (i+1 < MAX QoS Classes)
               Move to the next lower priority QoS class 'i+1'
         Else
               Break
         End IF
   End While
```

4G-FICC Upgradation Procedure

```
Else IF f(Q) > 1 // indicate network can send at higher rate
   i= index of the maximum priority QoS class
   While f(Q) > 1
         IF (ER_i < Maximum_rate_i)
               Upgrade Expected rate of 'i' QoS Class
         Else IF (i-1 < MAX Classes)
               Move to the next higher priority QoS class 'i-1'
         Else
               Break
         End IF
   End While
End IF
```

Figure 3.4 Algorithms of Degradation and Upgradation Procedures of 4G-FICC

3.2.9 Parameters of 4G-FICC

4G-FICC employs only a small number of parameters to keep it simple and scalable. Following parameters are utilized in the scheme.

`

Buffer Utilization Ratio (BUR)

The parameter BUR defines a target operating point (Q_0) as follows.

$$Q_0 = Buffer\_Size * BUR \quad 3.3$$

The decision as whether to restrict or encourage the ER of each QoS class is determined by BUR (buffer in use/buffer capacity), not by a fixed threshold regardless of the actual buffer capacity (Hoang and Wang, December 1999). When the queue length operates beyond the target operating point, 4G-FICC restricts the rate of connections. Otherwise, it encourages the sources to send more traffic.

BUR also determines the slope of the queue control function that is the rate at which f(Q) decreases or increases the ER of each QoS class. When the target point is set at a higher level, remaining buffer (Buffer Size – Q_0) is not large enough to drain the traffic effectively, so a tight queue control function is required to drain the queue. The resulting f(Q) with the large target point satisfies this requirement by increasing the slope of the function and hence restricts more aggressively the ER of each QoS class (Hoang and Wang, December 1999). The larger values of the target point avoid link underutilization by allowing more traffic in the network, but increase queuing delay.

When the target point is set at a lower level, the queue control function operates smoothly and the network is expected to have a short queuing delay (Hoang and Wang, December 1999). At a very low level of the target point, the remaining buffer (Buffer Size * (1-BUR)) absorbs large amount of traffic when the network is heavily loaded. However, the buffer may be empty when the traffic load is light, and results in underutilization of resources and low throughput.

Thus, BUR should be defined in a way that the resultant target operating point maximizes resource utilization and throughput in the network.

Overselling Factor (α)

The parameter 'α' is used in the queue control function (refer to Figure 3.3) when the network operates below the target level. It indicates how much extra capacity can be allocated to existing

connections, which can take the extra capacity. A high value of 'α' increases the over allocation rapidly and results in long queues and high delays in the network. Therefore, small values of 'α' are preferable by the queue control function. However, it must be little higher than 1 to encourage small allocation above the current allowed rate, and to avoid resource underutilization in the network (Hoang and Wang, December 1999).

Averaging Factor (β)

The parameter 'β' is an exponential average factor. It determines how fast the MACR converges to the new allowed class rate. When the value of 'β' is set to 1, MACR becomes equal to ACR. Whereas, when its value is set at 0, MACR is never updated. So, to allow MACR to gradually converge to ACR, the value of 'β' must be set in the range of 0 to 1, exclusive.

3.3 Fair Intelligent Admission Control for 4G Networks (4G-FIAC)

FIAC for the 4G networks (4G-FIAC) operates and implements active functions to react to the load at the core network and the limited resource availability in the access part of the LTE networks. 4G-FIAC combines the idea of Complete Sharing (CS) and Virtual Portioning (VP) resource allocation schemes. To guarantee maximum resource utilization, it avoids resource reservation for any specific type of service. Consequently, when resources are available in the network, it admits connection of any QoS class. Hence, 4G-FIAC uses a variation of VP that does not reserve even nominal bandwidth for any traffic group.

4G-FIAC decides an admission of a connection using the feedback from a congestion control module. It ensures the QoS of existing connections in terms of bandwidth and delay guarantee by rejecting incoming connections until the network is congested. Thus, it trade-offs between the incoming connections demand and the QoS of existing connections. As soon as the load reduces, it starts admitting connections to maximize the network resource utilization.

4G-FIAC differentiates among connections of different priority services by employing the bandwidth borrowing scheme. It also ensures fairness among connections at the same priority level. In this section, we detail 4G-FIAC, as well as its interaction with 4G-FICC.

The key contributions are as follows:

3.3.1 Load Estimation

4G-FIAC coordinates with a CC module, 4G-FICC, with a view to maintain the queue length at an output buffer fluctuating around the target operating point. Based on the feedback from the CC module, it estimates whether 4G-FICC is able to manage the load of an incoming connection. It intelligently admits or rejects an incoming connection based on its QoS requirements and the current load at the core network to prevent QoS degradation of the existing connections.

The idea is for the 4G-FIAC to predict whether, in times of the network congestion, the 4G-FICC can handle load of an incoming connection. If it can, the network is able to operate around the target operating point and hence avoids buffer overflow.

To predict whether 4G-FICC can manage the load, 4G-FIAC employs either of the following two approaches.

a) In the first approach, it validates that in times when the core network is overloaded, 4G-FICC can handle the load of an incoming connection by degrading the rate allocated to the existing connections of lower priority classes to their minimum rate. To perform this estimation it applies Eq. 3.4.

$$minimumrate_{inc} \geq \sum_{i=0}^{N} \sum_{j=0}^{n} f(pr_{inc}, pr_i) Maximumrate_{i,j} - Minimumrate_{i,j} \qquad 3.4$$

In Eq. 3.4, $minimumrate_{inc}$ indicates an incoming connection's minimum rate. The $f(pr_{inc},pr_i)$ returns 1, when the priority of an incoming connection's QoS class is higher than the priority of class 'i'. Otherwise, it returns 0. The equation verifies whether the minimum rate of an incoming connection is greater than the sum of the gap between the maximum and the minimum rates of the existing connections of all QoS classes with the priority lower compared to the priority of an incoming connection. As a result, the load estimation is strict for an incoming connection of lower priority and easy for an incoming connection of higher priority. Thus, in times when the core network is congested, 4G-FIAC encourages admission of high priority classes. Equation

3.4 validates only for the minimum rate of an incoming connection. This is due to the fact that 4G-FICC cannot degrade the minimum rate of a connection to control the load at an output buffer of a base station.

In Eq. 3.4, 4G-FIAC has to perform the steps to check the priority of all connections and also to sum up the rate above their minimum requirements. To formulate load estimation with fewer steps, we offered a second approach.

b) In the second approach, 4G-FIAC determines the network capability for each traffic class 'i'. To determine the network capability for each traffic class 'i', 4G-FIAC estimates the Expected Admission Rate (EAR_i). The EAR_i is the rate that the network can offer to a connection of class 'i' based on the average resource usage of class 'i' and the current network's congestion state. We believe that all flows of one class have the same Service Level Agreements (SLA).

$$EAR_i = MACR_i * f(Q) \quad 3.5$$

In Eq. 3.5, the queue control function indicates the status of the network congestion and hence controls the network capacity for incoming traffic. In case when the core network is congested and the queue at the buffer operates above the target operating point, the queue control function returns a value less than 1. Consequently, the EAR_i reduces and hence discourages incoming connections from entering the network. Otherwise, the EAR_i increases and this encourages new connections in the network. The $MACR_i$ is the mean allowed class rate of class 'i' (Figure 3.2). It reflects the average resource utilization of class 'i' per second. A connection status is set to *ready to admit* only when its requested rate (r_{req}) matches to the current allowed rate of its respective class 'i' (EAR_i) as shown in Eq. 3.6.

$$EAR_i \geq r_{req} \quad 3.6$$

Figure 3.5 summarizes the load estimation procedure of 4G-FIAC.

Load Estimation

```
1. Obtain value of Queue control function from 4G-FICC
//Estimates 4G-FICC can manage load of an incoming connection
2. IF Eq. 3.4 "OR" Eq. 3.6 == TRUE
       Admit_Status = Admit
   Else
       Admit_Status=Reject
   End IF
```

Figure 3.5 Algorithm of Load Estimation of 4G-FIAC

3.3.2 Bandwidth Borrowing

In situations when resources are available in the network to guarantee the QoS requirements of a new connection, 4G-FIAC sets the connection status to *ready to admit*. It admits the connection if it estimates that 4G-FICC can handle the load introduced by the connection (Figure 3.5).

In state of the limited resource availability, 4G-FIAC employs a step-wise degradation scheme to provide differentiation among the flows of different service types. The degradation procedure using a variable size degradation step provides fair resource allocation among the connections at the same and different priority levels. The following subsections discuss the way 4G-FIAC works and provides fair resource allocation among the users in the network.

3.3.2.1 Class level Fair resource allocation

4G-FIAC applies the idea of CS and accepts an incoming connection of any QoS class to maximize the resource utilization. It does not reserve resources for any QoS class with a view to minimize resource underutilization. To differentiate among multiservice users, it employs two strategies. Firstly, in situations when resources are scarce in the network, it applies a step-wise degradation on the rate allocated to existing connections to admit connections of high priority. The lower priority connections are degraded up to their minimum rates.

Secondly, the degradation procedure is applied only until sufficient resources are obtained to grant the rate requested by an incoming connection. The idea is that in times when resources are already scarce, 4G-FIAC tries to maintain the resources with the existing connections to admit incoming connections of high priority by applying stepwise degradation. When the degradation procedure is not applied accurately, resources more than requested by an incoming connection are added back to the available resource pool. A RAC that is based on the CS assigns these resources to an incoming connection of any priority level and often results in high Blocking Probability (BP) of higher priority connections. The degradation procedure of 4G-FIAC scheme degrades the existing connections up to what is required and thus sets aside resources for incoming flows of high priority. In this way, in times of resource scarcity, BP of high priority connections reduces. It results in differentiation and fair resource allocation among the connections of different QoS classes.

3.3.2.2 Flow level Fair resource allocation

4G-FIAC applies a step-wise degradation on all connections of the same QoS class. The rates of all connections at the same priority level are reduced fairly until they reach their minimum rates.

Degradation is applied only when 4G-FIAC predicts that enough resources can be obtained by degrading connections of the low priority to their minimum rates and the 4G-FICC can handle the load introduced by the incoming connection.

The 4G-FIAC first uses Eq. 3.7 to estimate if enough resources can be obtained by degrading the lower priority connections to their minimum rate.

$$Remaining_{Slots} \geq \sum_{i=0}^{N} \sum_{j=0}^{n} f(pr_{inc}, pr_i)\ (Slots\_Maximumrate_{i,j} - Slots\_Minimumrate_{i,j}) \qquad 3.7$$

In Eq. 3.7, $Remaining_{slots}$ refers to the number of slots to be taken from the existing lower priority connections to admit a high priority connection. It is the gap between the slots requested by an incoming connection ($Requested_{slots}$) and the slots in the available resource pool ($Available_{slots}$), ($Remaining_{slots}$: $Requested_{slots} - Available_{slots}$). When Eq. 3.7 is true, it sets the connection status to *ready to admit*. Otherwise, it rejects the connection. When 4G-FIAC

estimates that load estimation also sets connection status to *accept*, it applies degradation if required.

4G-FIAC applies a step-wise degradation starting with the lowest priority QoS class. It keeps on reducing the rate of flows until either enough resources are obtained or the rate of all connections reaches their minimum rates. In case when enough resources are not obtained, it moves to the next lower priority QoS class. It continues moving to the next lower priority levels either until enough resources are obtained, or the priority of the next QoS class is higher compared to the priority of an incoming connection.

The degradation step size varies depending on whether the connections at the same priority level operate at the same or different maximum rates. We define two cases.

a) In case when connections at the same priority level avail different maximum rates, the degradation step size is based on a function of slots f(slots) and the amount of over provisioned resources of a connection. It is estimated as in Eq. 3.8.

$$Step\_Size_{j,i} = f(Slots) * Maximum\_rate_{j,i} \quad 3.8$$

The function of slots in Eq. 3.8 is based on the current usage of network resources and the size of an incoming connection's request. Consequently, it is a function of total and used slots in the network; and the target slots. The target slots are the expected number of the used slots after the degradation procedure completes. It is obtained as the gap between the used slots and the remaining requested slots of an incoming connection ($Used_{slots}$ - $Remaining_{slots}$). The function of slots is defined using Eq. 3.9.

$$f(Slots) = 1 - \frac{Total_{Slots} - Used_{Slots}}{Total_{Slots} - Target_{Slots}} \quad 3.9$$

The use of the function of slots in the degradation step size makes it variable. It is because initially when the degradation procedure starts, the number of used slots in the network is high. As the degradation procedure progresses, the number of used slots reduces. Therefore, the degradation step size gradually reduces to obtain only required resources from the existing connections in the network.

The degradation procedure takes into account the amount of over provisioned resources of each connection. The degradation step size is obtained by applying f(slots) on the maximum rate of connections ($Maximum_{j,i}$) as in Eq. 3.8. As a result, the step size returns a higher value for the connections with high maximum rates compared to the connections with lower maximum rates at the same priority level.

The step size does not take into account the channel conditions of users directly but it is considered automatically during the degradation process. It is because the degradation procedure using Eq. 3.10 converts the step size in Eq. 3.8 to the number of slots, which can be taken from a connection 'j' of class 'i'. The slots size depends on the Modulation and Coding Scheme (MCS) of a user, which in turn depends on its channel conditions. Thus, the degradation procedure, corresponding the step size in Eq. 3.8, takes more resources from the users having lower MCS compared to the users having higher MCS. The step size in terms of slots is defined as a function of the step size in bits (Eq. 3.8) and the MCS of a user, as in equation Eq. 3.10.

$$Step\_Size\_Slots_{j,i} = f\left(Step_{Size\,j,i}, MCS_{j,i}\right) \tag{3.10}$$

When the degradation step size involves over provisioned resources, it needs to be normalized to apply accurate degradation. As discussed, accurate degradation ensures that only the requested resources are taken from the connections of lower priority QoS classes. Eq. 3.11 is used to normalize the step size for each connection.

$$Step\_Size\_Slots_{j,i} = \frac{Step\_Size\_Slots_{j,i}}{\sum_{j=0}^{n} Step\_Size\_Slots_{j,i}} * Remaining_{Slots} \tag{3.11}$$

Briefly, the proposed degradation step size by considering the amount of over provisioned resources ensures that maximum resources are obtained from the connections with higher resources above their minimum requirements, and with lower MCS. These connections normally occupy more resources compared to other connections in the network. In this way, 4G-FIAC ensures fairness among flows with the same QoS class. 4G-FIAC by applying a degradation procedure, in times of resource scarcity, ensures fairness and differentiation among the different QoS classes. Furthermore, the consideration of priority during the degradation procedure provides fairness among the connections with different QoS classes.

b) In case when all connections at the same priority level avail the same maximum rate and have the same channel conditions, the degradation step size can be simplified and is defined using Eq. 3.12.

$$Step\_Size_{i,j} = f(Slots) * Slot_{i,j} \quad 3.12$$

Equation 3.12 shows that the step size is a function of the function of slots (Eq. 3.9) and the slot capacity. As a result, the degradation procedure, at one time, degrades the rate of a connection equal to a fraction of the slot capacity. This way it avoids the step to normalize the contribution of each connection in the degradation procedure.

The degradation procedure of 4G-FIAC is detailed in Figure 3.6.

Degradation Procedure

1. Sort all flows according to the priority of QoS classes.
2. Start With the lowest priority QoS Class.
3. While enough resources are not obtained "AND" priority of QoS class 'i' is lower than the priority of an incoming request

```
Estimate Step size using Eq.3.11 or Eq. 3.12
For all connections 'j' with the same QoS class 'i'
   IF ACR_i,j - Step_size_i,j >= Minimum Rate of class 'i'
      ACR_i,j = ACR_i,j - Step_Size_i,j
   End IF
End For
Estimate updated Used_Slots.
IF ACR of all connections at a priority level 'i'
reaches their minimum rate
        Move to the next lower priority level
End IF
End While
```

Figure 3.6 Algorithm of Degradation Procedure of 4G-FIAC

The detail steps of the 4G-FIAC's admission procedure are presented in Figure 3.7.

Admission Procedure

```
1. IF Available_Slots > Requested_slots
            Admit_status = Admit
            Go to step2
  Else IF Eq. 3.7 == TRUE    //Estimate Remaining Slots can be
  obtained by degrading low priority flows to their minimum
            Admit_status = Admit
            Go to step 2
  Else
           Admit_status = Reject
  End IF
2. IF Admit_status == Admit
          Perform Load Estimation using algorithm of Figure 3.5
   End IF
3. IF Admit_status == Admit
         Apply Degradation if required using Figure 3.6.
   End IF
```

Figure 3.7 Algorithm of Connection Arrival Procedure of 4G-FIAC

3.4 Scheduler

The proposed QoS schemes, 4G-FICC and 4G-FIAC, can operate with any basic underlying scheduler such as Proportional Fair (PF), Round Robin (RR) etc. The scheduler needs to be enhanced to estimate the MACR of each QoS as discussed in Figure 3.2. Furthermore, in order to maintain the network traffic around the target operating point, the scheduler is enhanced to perform the resource allocation in accordance to the feedback from the congestion control module (Figure 3.1).

4G-FICC during its upgradation procedure can raise the expected rate of each QoS class to its maximum rate. The scheduler allocates resources to connections in accordance to the expected rate. Consequently, the connections, which are admitted at their minimum rate also starts getting resources to gain their maximum rate. This can lead in situation when some connections do not get resources to obtain even their minimum guaranteed rate. Therefore, the scheduler must be updated in a manner that it allocates resources to connections separately to gain their minimum and maximum rate requirements. It should not allocate resources to connections above their minimum rate requirements until all connections are getting their minimum rate.

Fairness is a significant performance metric for resource allocation schemes. According to one notation of fairness called index of fairness, the fairness is a function of variations in throughput of users (Jain et al., 1998). When apply the idea of fairness to our proposed 4G-FICC and 4G-FIAC schemes, it is expected that the rate estimated of each QoS during the degradation and upgradation procedures ensures fairness among the users within the same QoS class and with different QoS classes. It is necessary that all users of the same QoS class having the same QoS requirements be treated fairly. Fairness at the same QoS class requires that users are getting the fair share of resources and hence the same throughput in the network. Fairness among different QoS classes necessitates that users be differentiated according to their respective priority levels. Consequently, the resource allocation changes among the users determined by their priority levels. For instance, the users with higher priority are given higher share of network resources compared to other users with lower priority.

To achieve this objective, the 4G-FICC reduces or increases the expected rate of each QoS class depending on its priority and the congestion state in the network. Similarly, 4G-FIAC decreases or increases the rate of connections of each QoS class depending on their priorities. The degradation procedure reduces the rate staring from the lowest priority class. Whereas, the upgradation procedure increases the rate of the connections starting from the highest priority.

The proposed scheduler allocates resources to flows of the same QoS class corresponding the expected rate of the respective QoS class. To attain the same expected rate, the scheduler assigns higher resources to connections having bad channel conditions compared to users with good channel conditions. Initially, scheduler allocates resources to connections to gain their minimum

rate. When all connections gain their minimum guaranteed rate, it allocates resources to connections, which can take them to gain their maximum rate, starting from the highest priority class. In this way, 4G-FICC and 4G-FIAC with the scheduler provides max-min fairness to the flows of the same QoS class and with different QoS classes.

3.5 Summary

This chapter introduced our proposed QoS framework and new control algorithms for the 4G networks. We first defined our proposed QoS framework and further introduced essential features of the new control algorithms, 4G-FICC and 4G-FIAC. These diverse algorithms take charge of satisfying the QoS requirements of existing users in situations when the load and the traffic demand in the network increases. We then introduced the customization of these schemes to match the QoS structure of the 4G networks. Finally, we used the general cases to explain the basic operation of these schemes. Based on the basic congestion control algorithm presented in this chapter, Chapter 4 and Chapter 5 discuss the congestion control algorithms designed to match the specific QoS specifications of WiMAX and LTE networks, respectively. Formulated based on the basic admission control algorithm presented in this chapter, Chapter 6 and Chapter 7 comprehensively present the admission control algorithms designed to address the QoS requirements of WiMAX and LTE networks and to reduce the BP of connections in a fair manner.

Chapter 4 WiMAX Fair Intelligent Congestion Control- (WFICC)

In this chapter, we describe our Congestion Control (CC) mechanism for WiMAX networks, the Fair Intelligent Congestion Control (WFICC). It avoids congestion at a base station. It ensures that the network traffic is scheduled in such a way that a base station output buffer operates close to a target operating point without violating the QoS requirements of connections of different service types. It estimates the fair share of all QoS classes and sends the estimated rates as a feedback to underlying schedulers at a base station making it scalable and flexible. WFICC employs only few parameters. The scheme is robust to various parameter settings.

Section 4.1 focuses on the CC algorithm specifically designed for WiMAX networks. Section 4.2 presents the simulation setup and Section 4.3 provides comprehensive performance analysis. Section 4.4 presents the analysis of the parameters involved in the CC scheme. It evaluates the performance of the system by varying values of the parameters of WFICC. Section 4.5 provides verification of the scheme in terms of scalability and robustness. Finally, Section 4.6 summarizes this chapter.

4.1 Congestion Control Algorithm for WiMAX Networks

In WiMAX architecture a base station lacks the mechanisms to regulate its load, particularly in situations when the core network is congested. Large queue at a base station buffer risks high delays and buffer overflow. This can lead to degradation in the Quality of Service (QoS) experienced by users in the cell. Congestion avoidance is rarely considered to ensure the QoS of wireless connections. Efforts are mainly aimed at reducing the overload once it has occurred or to ensure the provision of QoS under load conditions.

The IEEE 802.16-2005 standard defines specific QoS parameters for each Class of Service (CoS) but it does not provide the mechanisms for ensuring the QoS of the different class of services. The scheduler that allocates resources among service flows of each class of service is not specified by the standard and is left as an open research area. The standard also does not

specify any mechanism for congestion prevention in the network. To provide satisfactory QoS, (Tung et al., 2008) suggested a scheduling algorithm that uses thresholds to achieve QoS of different CoSs. The researchers (Casey et al., 2008) and (Rodrigues and Cavalcanti, 2008) proposed base station initiated handover to deal with the overload once it exists. The authors (E. O. Lucena et al., 2010) proposed an improvement over the scheme that was proposed by (Rodrigues and Cavalcanti, 2008) by adding the delay prediction. All these schemes are based on thresholds, and are active only when the network is already heavily loaded, but do not operate when network is approaching congestion. This mode of operation results in inefficient utilization of network resources.

In this section, we propose a congestion control mechanism for WiMAX networks, the WiMAX Fair Intelligent Congestion Control (WFICC). It allocates the network resources to all Class of Services (CoSs) fairly and efficiently and also ensures their QoS requirements. At the base station, WFICC defines a target operating point and uses a rate allocation scheme to maintain the network traffic around the target. In contrast to all other schemes that operate only when the network is above a specified threshold, WFICC is always operational whether the network is above or below the target operating point. In WFICC, a queue length value is really an indication of the network operation level. Large queue values indicate that the network is overloaded, risking high delays and buffer overflows, whereas small queue values imply light load and resource underutilization. So, WFICC uses a rate allocation scheme that takes this indication into account. It aims for a target operating point where the queue length is acceptable for both throughput and delay.

4.1.1 WiMAX Fair Intelligent Congestion Control (WFICC)

WFICC sets the target operating point at a preset Buffer Utilization Ratio (BUR). WFICC estimates the degree of congestion in the core network by calculating a queue control function (f (Q)) in Figure 3.3. When the value of f(Q) is less than 1, indicating there is congestion in the core network, WFICC applies a degradation procedure.

WFICC is not applied to UGS and ertPS Class of Services (CoSs), as in their case a scheduler allocates fixed amount of the slots periodically in every frame. Therefore, WFICC applies on

rtPS, nrtPS and BE CoSs only. The scheduler at a base station allocates each connection the slots corresponding to the expected rate of its CoS (ER_{CoS}) and updates the $MACR_{CoS}$ as in Figure 3.2.

4.1.1.1 Degradation Procedure

In times of congestion, when the base station's queue length operates beyond the target operating point, WFICC applies a degradation procedure. The degradation procedure reduces the current rate of the existing connections so that the queue length operates around the target operating point and avoids buffer overflow. During the degradation procedure WFICC considers the priority of connections.

The degradation procedure always starts from the lowest priority CoS and degrades its expected rate until either the queue control function indicates that the congestion can be controlled by the applied degradation or the rate of a CoS reaches its minimum class rate. Hence, WFICC while applying degradation considers the QoS requirements of a CoS. Once the expected rate of the lowest priority CoS reaches its minimum rate and if the congestion condition still remains as indicated by the new estimated value of the queue control function, WFICC moves to the next lower priority levels and applies degradation procedure. The detail steps of the degradation procedure to theses classes are as follows.

BE connections

First of all, WFICC reduces the bandwidth allocated to all connections of the BE class. The step size of degradation is controlled by the function of queue as it provides an estimation of how much degradation to apply on this class of traffic. WFICC estimates the updated queue length based on the degradation applied on the BE connections using Eq. 4.1 and recalculates the queue control function. It continues degrading resources allocated to the BE connections until either f(Q) gives a value equal to or greater than 1, or the expected rate of the BE CoS (ER_{BE}) reaches its minimum class rate. As the standard does not specify the minimum rate requirements for the BE CoS so we set minimum rate of the BE CoS ($MRTR_{BE}$) '0'.

$$Q_{len} = Q_{len} - Qlen_{CoS} + (Qlen_{CoS} * Ratio) \quad 4.1$$

In Eq. 4.1, $Qlen_{CoS}$ represents length of the queue, which receives incoming packets of a specific CoS at the base station. Ratio in Eq. 4.1 represents the effect of change in the class expected rate to its current mean allowed class rate ($ER_{CoS}/MACR_{CoS}$).

nrtPS Connections

After degrading the BE connections, WFICC estimates the f(Q) based on the updated queue length. If the value of f(Q) is again less than 1, indicating congestion still exists in the network, WFICC reduces the bandwidth allocated to all connections corresponding to the next priorities class, nrtPS CoS, until either f(Q) is equal to or greater than 1, or the expected rate of nrtPS (ER_{nrtPS}) CoS reaches the class minimum rate ($MRTR_{nrtPS}$). Once again WFICC estimates the updated Qlen based on the degradation applied on the nrtPS connections using Eq. 4.1.

rtPS Connections

After degrading the expected rate of the nrtPS CoS to its minimum class rate (MRTR), WFICC again estimates f(Q). If it is less than 1, WFICC reduces bandwidth allocated to all connections of the next priorities class, rtPS CoS, until either the f(Q) is equal to or greater than 1, or the expected rate of the rtPS CoS (ER_{rtPS}) reaches the class minimum rate ($MRTR_{rtPS}$). When degradation is applied, WFICC is not allowed to degrade the rate of rtPS and nrtPS CoS below the required MRTR of the respective CoS. There is no such restriction for BE CoS.

4.1.1.2 Upgradation Procedure

In times when the network resources are underutilized, WFICC applies an upgradation procedure. The upgradation procedure increases the current rate allocated to the existing connections so that the queue length operates around the target operating point. Consequently, the network resources are utilized maximum and the output link is never idle unnecessarily.

The upgradation procedure always starts from the highest priority CoS. It upgrades its expected rate until either the applied upgradation is sufficient to bring the queue length close the target operating point, or the expected rate of the class reaches the maximum rate (MSTR). Hence, WFICC while applying upgradation considers the QoS requirements of a CoS. Once the rate of the highest priority CoS reaches its maximum rate and the network still has the capacity to

manage further increase in the connections rate, WFICC moves to the next higher priority levels and applies the upgradation procedure. The detail steps of the upgradation procedure are discussed as follows.

rtPS Connections

WFICC raises the bandwidth share of all the connections of the rtPS class. The step size for share increase is still determined by the queue control function but at a different rate indicating how far the network is underutilized. WFICC estimates the updated queue length based on the upgradation applied on the rtPS connections using Eq. 4.1 and recalculates the queue control function. It continues upgrading the expected rate of the rtPS CoS until either the value of f(Q) is equal to or less than 1, or the ER_{rtPS} reaches the maximum class rate ($MSTR_{rtPS}$).

nrtPS Connections

After upgrading the expected rate of the rtPS CoS to its maximum class rate, WFICC estimates the queue control function based on the updated queue length. If again f(Q) is greater than 1, indicating still the network operates below its target point, WFICC raises the bandwidth share of the next higher priority class, nrtPS CoS, until either the f(Q) is equal to or less than 1, or the ER_{nrtPS} reaches the maximum class rate ($MSTR_{rtPS}$). Once again WFICC estimates the updated queue length based on the upgradation applied on the expected rate of the nrtPS CoS.

BE connections

After upgrading the expected rate of the nrtPS CoS to its maximum class rate, WFICC estimates the queue control function based on the updated queue length. If again the value of f(Q) is greater than 1, finally it raises the bandwidth share of the next higher priority class, BE CoS. WFICC is not allowed to raise the bandwidth share beyond the required MSTR of the CoS.

Figure 4.1 provides the detail algorithm of WFICC. The flow chart of WFICC is given in Figure 4.2.

Algorithm of WFICC

1. Obtain values of MACR for rtPS, nrtPS and BE Class of Services calculated by scheduler.
2. Obtain queue length (Qlen) from router interface.
3. Calculate f(Q) as follows.

IF (Qlen > Q_0)

$$f(Q) = \frac{(Buffer\_Size - Qlen)}{(Buffer\_Size - Q_0)}$$

Else

$$f(Q) = \frac{(\alpha - 1) * (Q_0 - Qlen)}{Q_0} + 1$$

End IF

4. Degradation Process

If f(Q) < 1 // Indicates Congestion in the network

For each lowest to highest priority CoS //BE, nrtPS, rtPS

$ER_{CoS}(t) = MACR_{CoS}(t-1) * f(Q)$ // Estimate Expected rate

Do While $ER_{CoS} > MRTR_{CoS}$ AND f(Q) < 1 // Degrade rate of a CoS

$Ratio = \frac{ER_{CoS}(t)}{MACR_{CoS}(t-1)}$ // Estimate ratio of change of ER to its MACR

Update $MACR_{CoS}$(t-1) using following Equation.

If $Qlen > Q_0$

if $ACR_{CoS}(t) < MACR_{CoS}(t-1)$ // Do not track a rate higher than MACR when network is congested.

$$MACR_{CoS}(t) = MACR_{CoS}(t-1) + \beta * \left(ACR_{CoS}(t) - MACR_{CoS}(t-1)\right)$$

Else If $Qlen < Q_0$

$$MACR_{CoS}(t) = MACR_{CoS}(t-1) + \beta * \left(ACR_{CoS}(t) - MACR_{CoS}(t-1)\right)$$

End If

$Qlen = Qlen - Qlen_{CoS}(t-1) + (Qlen_{CoS}(t-1) * Ratio)$//

$Qlen_{CoS}(t) = Qlen_{CoS}(t-1) * Ratio$ // Update estimated queue lengths based on new estimated expected rate.

Algorithm of WFICC (Cont…)

Recalculate queue control function using f(Q) in step 3.

$ER_{CoS}(t) = MACR_{CoS}(t-1) * f(Q)$ // Estimate updated Expected rate

End DO While

End For

5. Upgradation Process

Else If f(Q) > 1 // Indicate the core network is underutilized

For each highest to lowest priority CoS //rtPS, nrtPS, BE

$ER_{CoS}(t) = MACR_{CoS}(t-1) * f(Q)$ // Estimate Expected rate

Do While $ER_{CoS} < MSTR_{CoS}$ AND f(Q) > 1 //upgrade rate of each CoS

$ER_{CoS}(t) = MACR_{CoS}(t-1) * f(Q)$ // Estimate Expected rate

$Ratio = \frac{ER_{CoS}(t)}{MACR_{CoS}(t-1)}$ // Estimate Ratio of change of ER to its MACR

Update $MACR_{CoS}(t-1)$ using following Equation.

If $Qlen > Q_0$

if $ACR_{CoS}(t) < MACR_{CoS}(t-1)$

$MACR_{CoS}(t) = MACR_{CoS}(t-1) + \beta * (ACR_{CoS}(t) - MACR_{CoS}(t-1))$

Else If $Qlen < Q_0$

$MACR_{CoS}(t) = MACR_{CoS}(t-1) + \beta * (ACR_{CoS}(t) - MACR_{CoS}(t-1))$

End If

$Qlen = Qlen - Qlen_{CoS}(t-1) + (Qlen_{CoS}(t-1) * Ratio)$

$Qlen_{CoS}(t) = Qlen_{CoS}(t-1) * Ratio$ // Update estimated queue lengths based on new estimated expected rate.

Recalculate queue control function using f(Q) in step 3.

$ER_{CoS}(t) = MACR_{CoS}(t-1) * f(Q)$ // Estimate updated Expected rate

End DO While

End For

End If

Figure 4.1 Algorithm of WFICC

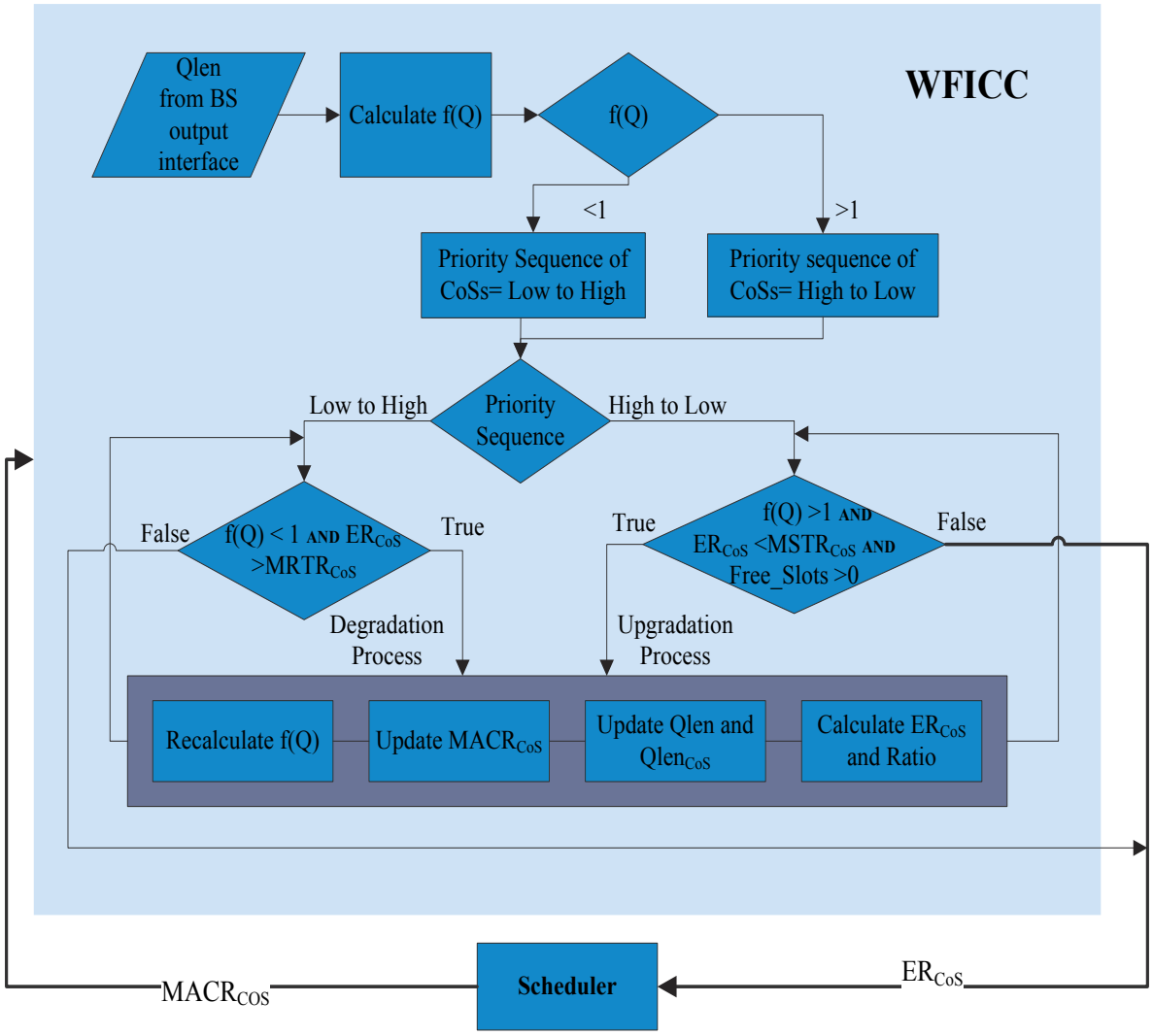


Figure 4.2 Flow Chart of WFICC

4.2 Simulation Setup

The overall goal of the simulation is to investigate the performance of WFICC to avoid congestion and to control congestion when it is inevitable. The simulations thoroughly investigate the impact of WFICC on the performance of the network in terms of throughput, fairness, delay and jitter. The simulations have been performed in ns-2 (ns2, 2010b) using the WiMAX module for OFDMA by (WiMAX Forum and NIST, 2011) namely ns2-wimax-awg.

In the current simulation setup, WiMAX Subscriber Stations (SSs) are connected to a WiMAX BS in IEEE 802.16 PMP mode. The BS is connected to a sink node to reflect the actual

deployment of WiMAX networks. The link capacity between the sink node and the base station is set at 1.3Mbps and the target operating point is initially set at 1/4 of the total buffer capacity.

The 256 kbps VoIP traffic agents are attached to the source nodes. For the video traffic, nodes are generating 256 kbps H.263 data streams. The simulations utilize the trace file for 256 kbps H.263 encoded Jurassic Park movie provided by (Fitzek and Reisslein, 2001). The 256 kbps FTP traffic agents are created and attached to the source nodes. To represent the web traffic, 256 kbps exponential traffic agents are attached to the SSs. On top of the agents, applications are created to initiate UDP and TCP traffics. Each application is generating the same data rate, so MSTR and MRTR of each CoS are set at 80Mbps and 160 Mbps, respectively. The simulation setup is shown in Figure 4.3.

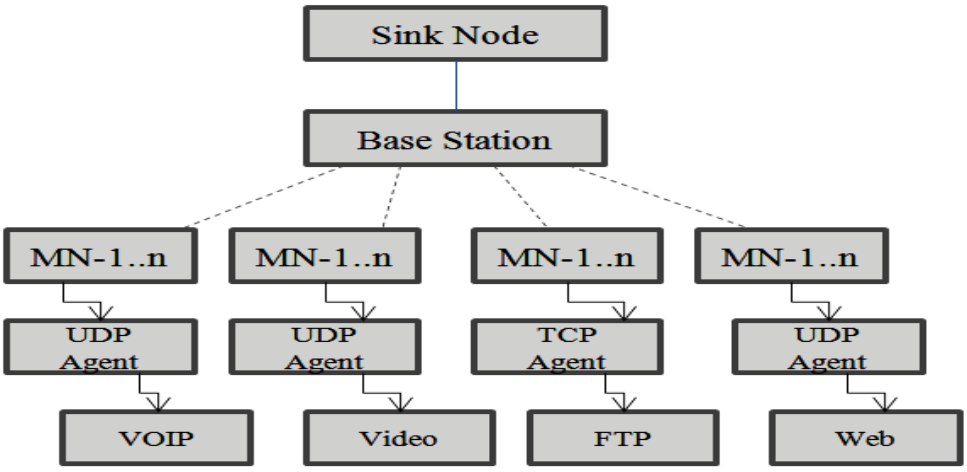


Figure 4.3 Simulation Setup

There is one service flow between each SS and the BS. The UGS service flows are dynamically added to the base station to handle the VoIP traffic. For the video traffic, rtPS flows are added to the base station. The nrtPS flows are created for the FTP traffic and BE service flows are added to manage the web traffic.

An error free channel is assumed during the simulation. The frame duration is set at 5 ms and the ratio of downlink to uplink is set at 7:3. The modulation and coding scheme, QPSK-3/4, with slot capacity of 9 bytes in uplink is assumed for all nodes. Table 4.1 shows the simulation system parameters in detail.

Table 4.1. System Parameters

Frame Duration	5 msec	
Bandwidth (BW)	10MHz	
FFT size for 10 MHz BW	1024	
Total Subcarriers (N_{used})	841	
Modulation and coding Scheme (MCS)	QPSK-3/4	9 B/UL Slot
Symbols in 5 msec frame with Guard Time (Tg)=1/8	48.6	
Symbols used for TTG and RTG	1.6	
Symbols after TTG and RTG	47	

	DL	**UL**
DL:UL	7	3
Symbols per frame	33	14
Symbols per Cluster (PUSC)	2	3
Number of Subcarriers per Cluster or Tile (N_{Sub})	14	4
Clusters or Tiles per Slot (Nn)	2	6
Number of subcarriers per Cluster/Tile ($N_{Total\_Sub}= N_{Sub}* N_n$)	28	24
Total Clusters/Tiles in 10 MHz bandwidth (N_{used}/N_{Total_Sub})	30	35
Slots in 5 ms Frame (Symbols*Clusters/Tiles)	480	140

In the current implementation of WFICC, we assume there is no admission control scheme at the BS. Scheduler used in the current implementation is Proportional Fair (PF), which does not consider the QoS requirements of each CoS and the load in the network while allocating resources to connections. The proportional fair scheduler provided in ns-2 grants resources to connections until their queues are empty.

4.3 Simulation Results

This section presents results of the proposed WFICC scheme. For comparison, the system performance with WFICC is compared with system performance without WFICC.

4.3.1 Queue Length (Qlen)

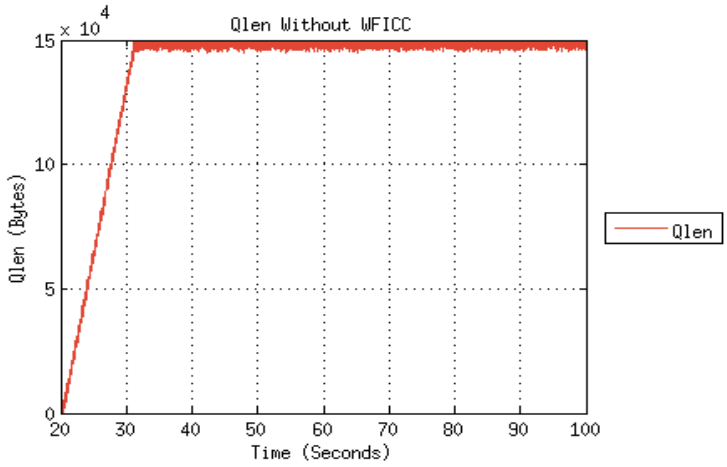


Figure 4.4 Queue length (Bytes) without WFICC

Figure 4.4 shows the queue length at the base station without WFICC. The simulations are performed with only one output buffer for data of all class of services. In times when the packet core is overloaded, the base station cannot transmit data at the same rate at which it receives from the users in the network. As a result, the queue at the base station starts building up.

Figure 4.4 proves that in such situations if the scheduler keeps on allocating resources to the existing connections without taking into account the capacity of the output buffer, a point comes when the buffer gets overflowed and packets drop starts.

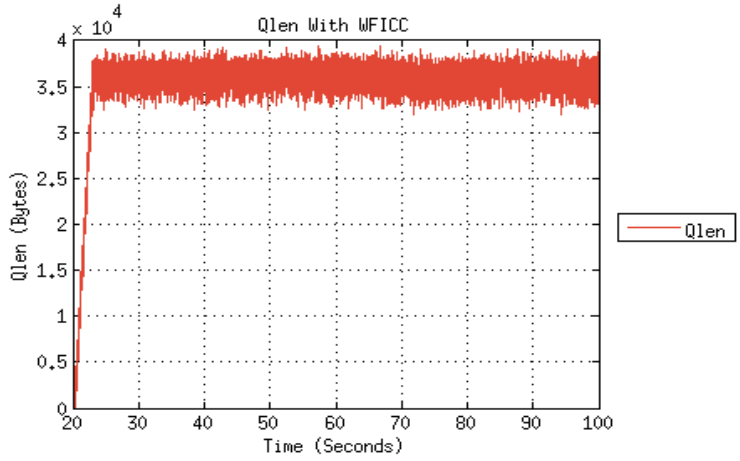


Figure 4.5 Queue length (Bytes) with WFICC

Figure 4.5 shows the queue length at the base station output buffer with WFICC. As designed, the queue is maintained around the desired target operating point. Since, as soon as the queue length reaches the target operating point, WFICC's congestion handling mode starts its operation and discourages incoming traffic until the queue length drops below the target point. When the queue length is operating below the target operating point, WFICC operates in oversell mode and encourages incoming traffic to maintain the network operation around the target operating point.

4.3.2 Fair resource allocation among CoSs and within a CoS

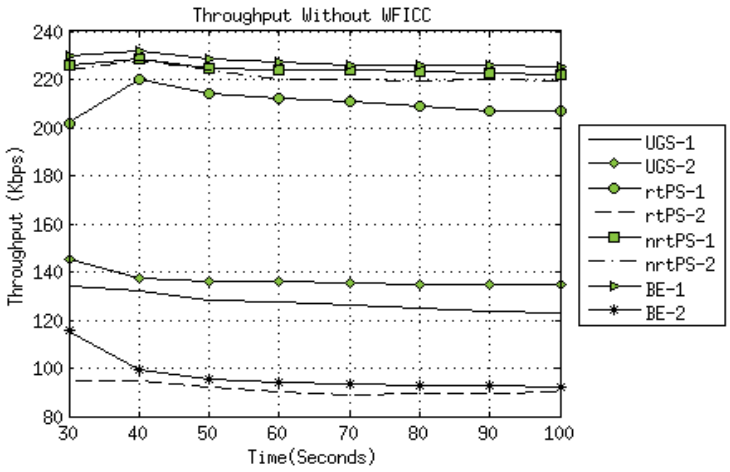


Figure 4.6 Throughputs (kbps) of Service flows without WFICC

Figure 4.6 demonstrates that when the PF scheduler allocates resources without taking into consideration the QoS parameters of each CoS, the system could not deliver fair resource allocation among the aggregates of different CoSs. It demonstrates that the throughput of UGS, rtPS and nrtPS connections is even less than the MRTR of each respective class. It is because, there is no admission control in the network, and so the resource requirement of the network connections to gain their MSTR is higher than the resources available in the network. The proportional fair scheduler provided in ns-2 doesn't take into account MRTR and MSTR of any CoS during resource allocation and grants resources to connections until their queues are empty. As a result, the PF scheduler is not able to provide the minimum QoS to the connections.

It is clearly shown from Figure 4.6 that there is even no fairness among the aggregates of the same CoS. The two service flows from the same CoS (such as rtPS) and with the same data rate at the application layer are not getting the same throughput.

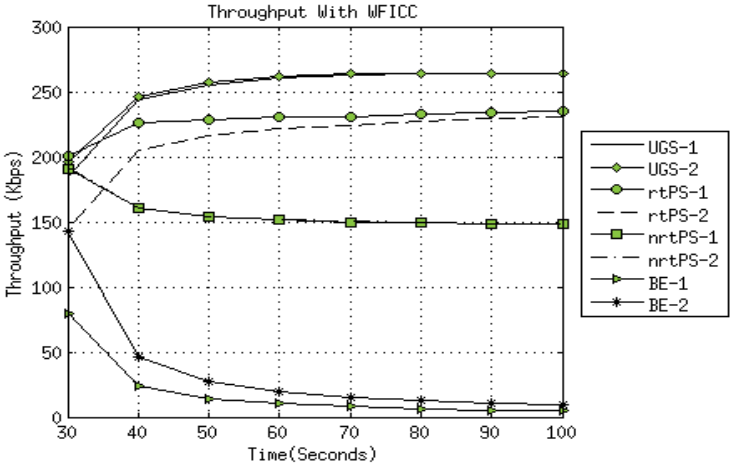


Figure 4.7 Throughputs (kbps) of Service flows with WFICC

Figure 4.7 shows the throughput of each CoS with WFICC. Initially with only the PF scheduler operating, the throughput of even high priority classes such as UGS and rtPS is less than or equal to the throughput of low priority classes such as nrtPS and BE. As soon as the queue length approaches the target queue length and congestion is detected, WFICC starts degrading the expected rate of the service flows starting with the lowest priority class of service. The graph shows that resources taken from the less priority classes (such as BE) are assigned to the high priority classes (such as UGS, rtPS, nrtPS) so that the system can provide the MRTR of these classes.

When the queue length becomes less than the target operating point, WFICC starts the upgradation process. During the upgradation it first upgrades the expected rate of rtPS CoS. Based on the new estimated f(Q), it distributes resources among the nrtPS connections and finally among the BE connections. So in both the degradation and the upgradation processes, it gives priority to the high priority class of services. As a result, in Figure 4.7 the throughput of the aggregates of all CoSs is achieved in the order of precedence given in the WiMAX standard

that's UGS > rtPS > nrtPS > BE (IEEE 802.16-2009). Hence, WFICC ensures fair resource allocation among different CoSs.

WFICC also provides fairness among the service flows of the same CoS. Figure 4.7 indicates that the throughput of the two flows of any CoS (such as UGS) are the same.

4.3.3 Average Delay

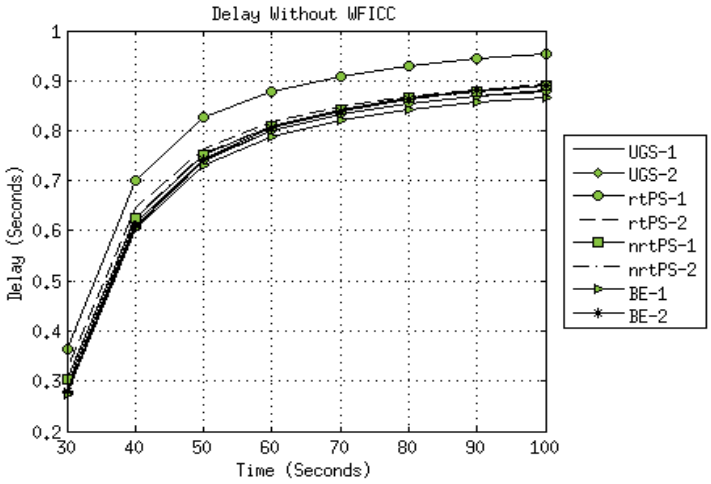


Figure 4.8 Average Delay (sec) without WFICC

Under the heavy load conditions, when the scheduler keeps on allocating resources without considering the current status of the output buffer, the queue at the buffer builds up and this results in very high delay in the network as illustrated in Figure 4.8.

Figure 4.9 shows the average delay for each class of service. Initially as there is less amount of data in the queue, so the delay is less. When the traffic increases, the amount of data in the queue increases resulting in an increase of the delay. WFICC maintains the delay of each class at a steady level. It is because it estimates the expected rate of each CoS and passes it to the scheduler to ensure that the network traffic is scheduled in such a way that the queue length is always around the target point.

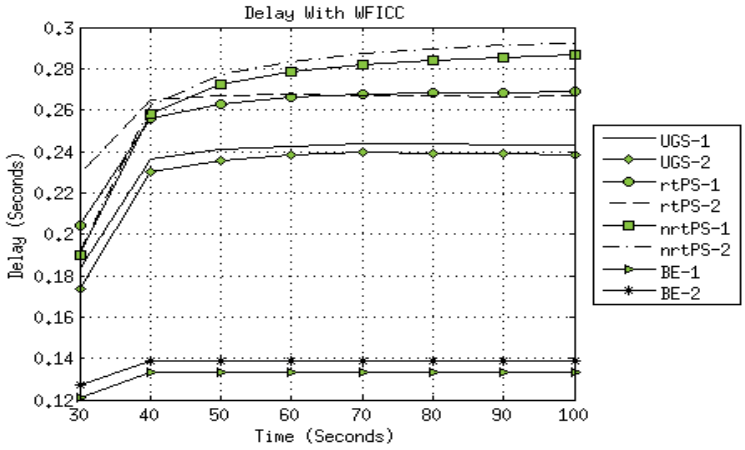


Figure 4.9 Average Delay (sec) with WFICC

4.4 Parameter Settings

To offer the desired level of control, WFICC uses three parameters: a target operating point (Q_0), an over selling factor (α) and an averaging factor (β). The target operating point and the over selling factor are used in the queue control function and the parameter 'β' is used for updating the $MACR_{CoS}$. This section demonstrates that WFICC is robust with respect to changes in these parameters.

Effect of the Averaging Factor (β)

Let $MACR_{CoS}(t-1)$ denote the old value of MACR of each CoS calculated in time (t-1). The scheduler at a base station updates $MACR_{CoS}(t-1)$ to a new value represented as $MACR_{CoS}(t)$ using Eq. 4.2.

$$MACR_{CoS}(t) = (1-\beta)MACR_{CoS}(t-1) + \beta * ACR_{CoS}(t) \qquad 4.2$$

In Eq. 4.2, $ACR_{CoS}(t)$ denotes the actual rate allocated to the service flows of each CoS in the current time (t). The parameter β determines how fast the $MACR_{CoS}(t-1)$ converges to the $ACR_{CoS}(t)$.

`

Effect of the Target Operating Point (Q_0)

It determines whether to restrict or encourage the expected rate (ER) of each class of service. It also determines the slope of f(Q), the rate at which f(Q) decreases or increases the ER of each CoS.

The network condition when the parameter Q_0 is more critical is as follows.

State When Total input > Bandwidth/Link Capacity

When the total input to a base station (as shown in Eq. 4.3) is greater than its output link capacity, congestion is experienced at the base station and the queue length starts increasing. In this situation, only the parameter Q_0 determines when WFICC starts handling the congestion.

$$Total\ input = \sum_{i=1}^{n_{UGS}} ACR_{UGS+} \sum_{i=1}^{n_{rtPS}} ACR_{rtPS+} \sum_{i=1}^{n_{nrtPS}} ACR_{nrtPS+} \sum_{i=1}^{n_{BE}} ACR_{BE} \quad 4.3$$

In Eq. 4.3, n_{CoS} shows the total number of service flows per each CoS.

Effect of the Oversell Factor (α)

The oversell factor (α) is used in f(Q) when the network is in the non congested state and most likely the network is underutilized. In this situation, the network may want to overstate its present capacity in the hope that the network users take the available bandwidth and maximize the network utilization. It also determines the slope and size of the queue control function and hence the degree at which the network can oversell its capacity.

The network state when the parameter α is critical one is discussed as follows.

State When Total input < Bandwidth/Link Capacity

When the total input of the network to a base station (Eq. 4.3) is less than the output link capacity and the queue length is below the target operating point, the parameter 'α' determines the rate at which users are encouraged to inject their traffic inside the network.

Constraint on parameter 'α'

The value of the parameter α must be larger than 1.0 to increase the f(Q) and hence the expected rate to avoid the link underutilization. The maximum that a queue length grows depends on the target queue length. However, the height of variations above the target point is determined by 'α'. So, the parameter 'α' should not be set very high that it causes large variations in the rate allocation and the resulting queue length. Therefore, a value slightly greater than 1 is desirable to bring the network around the target operating point. WFICC oversells resources and increases the queue length until it reaches the target operating point and maintains it around this point.

To validate our analysis of the parameters of WFICC, we perform simulations and compare simulation results with the expected results from our analysis.

This section presents and discusses the results of our investigation of the effect of different parameter settings of WFICC. As the BE CoS does not have any MRTR or delay requirements so results for BE CoS is not shown in the following subsections.

4.4.1 Impact of Target Operating Point (Q_0)

In this subsection the effect of the Q_0 on the performance of the system is investigated. To investigate the Q_0, the oversell factor 'α' is set at 1.3 and the parameter 'β' is set at 0.3.

Figure 4.10 and Figure 4.11 show the queue lengths at the BS output buffer for the BURs of 1/8 and 1/16, respectively. As designed, the queue length is clearly maintained around the target.

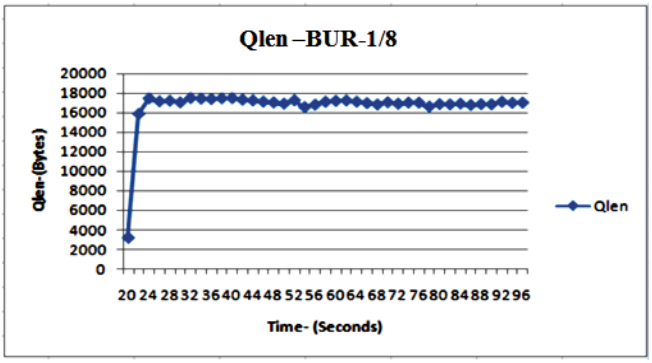


Figure 4.10 Queue length (Bytes) with BUR– 1/8

Figure 4.10 indicates that when the Q_0 is set at a higher level, the queue length fluctuates lightly around an average value, resulting in low jitters.

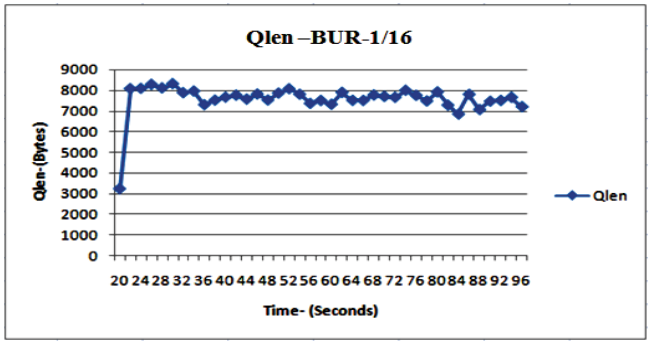


Figure 4.11 Queue length (Bytes) with BUR- 1/16

When the Q_0 is set at a lower level, the queue length fluctuates noticeably around an average value resulting in higher jitter as shown in Figure 4.11.

The effect of the target operating point on the system performance in terms of free slots, throughput, fairness, delay and jitter is discussed below.

Free –Slots

As discussed before, in WiMAX networks users are allocated resources in terms of slots. Free slots referred to the number of slots that are left unused after allocating resources to users.

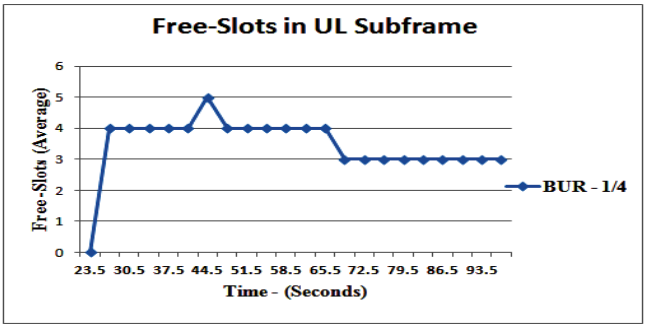


Figure 4.12 Average numbers of free slots with BUR-1/4

Figure 4.12 shows the average number of free slots in the uplink frame with the target operating point set at 1/4 of the buffer size. The PF scheduler allocates resources to the

connections without taking into consideration the congestion at the BS output buffer. As a result, the average number of free slots is zero. As soon as the queue length goes beyond the target operating point, WFICC starts its operation and reduces the bandwidth allocated to the existing connections to ensure that the network operates around the target operating point and therefore results in more free slots.

When the network starts it operation below the target operating point, WFICC oversells the bandwidth and results in again full utilization of resources, so an average number of free slots is ranging between 1 and 5 per frame.

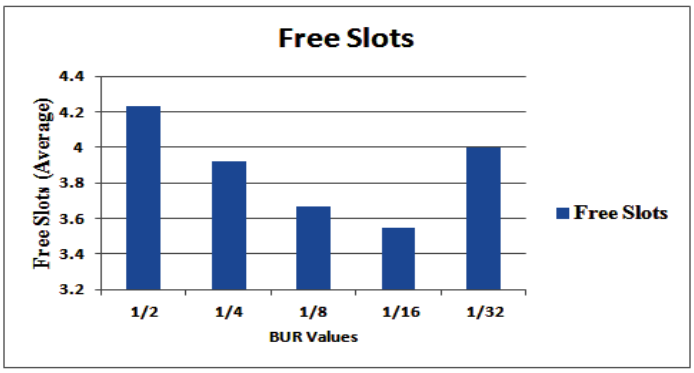


Figure 4.13 Average Number of free Slots with different values of BUR

Figure 4.13 shows the effect of the target operating point on the number of free slots in the system. It shows that as the target operating point reduces from 1/2 to 1/16 of the buffer size, the average number of free slots reduces. Consequently, the throughput increases and the delay reduces. Whereas, when the target point is reduced to a very small value such as 1/32 of the buffer capacity, the number of average free slots increases. It is because to keep the queue length around a very low target point, WFICC reduces the expected rate of each CoS to a very low value approaching to the MRTR of the respective CoS.

Throughput

Figure 4.14 and Figure 4.15 show throughput of the network for BUR of 1/2 and 1/16, respectively.

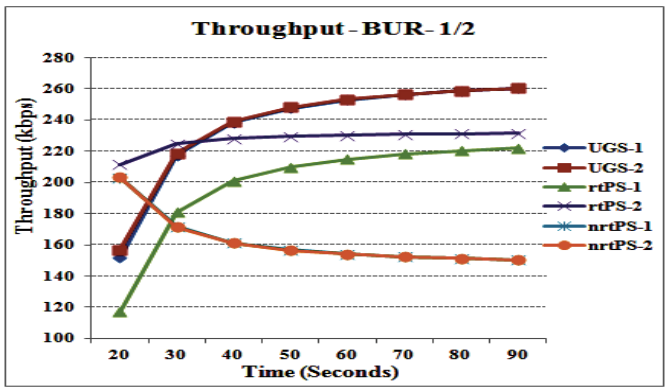


Figure 4.14 Throughputs (kbps) of Two-Flows per CoS with BUR-1/2

The PF scheduler allocates resources without taking into consideration the QoS parameters of each CoS. As a result, when the Q_0 is set at a higher level such as 1/2 of the buffer capacity, the network does not provide fairness for a longer period as clearly indicated in Figure 4.14. However, when the queue length goes beyond the target and WFICC executes, it ensures fairness among the aggregates of different CoSs as well as with the same CoS.

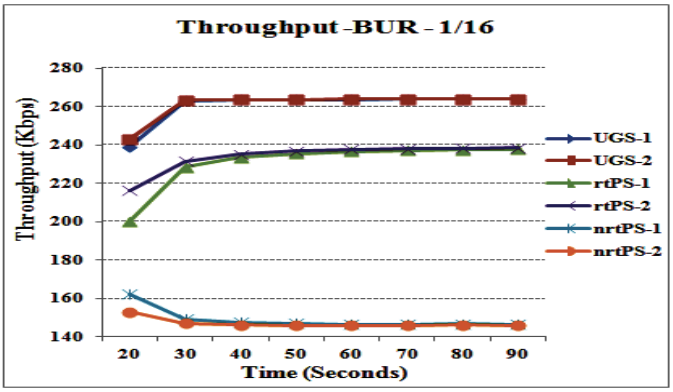


Figure 4.15 Throughputs (kbps) of Two-Flows per CoS with BUR-1/16

Figure 4.15 shows when Q_0 is set at a small level such as 1/16 of the buffer capacity, WFICC starts its operation relatively early and ensures that resources are allocated fairly among the aggregates of different CoSs according to their QoS constraints and priorities in the network. Consequently, the network starts providing fairness among the connections relatively early compared to scenario discussed in Figure 4.13.

Table 4.2 shows the throughput of two rtPS SFs for the Q_0 ranging from 1/2 to 1/32 of the total buffer size. The table shows when the Q_0 is set at a small value not less than 1/32 of the buffer

size, the throughput of rtPS connections increase at a small percentage. Additionally, as the Q_0 reduces from 1/2 to 1/32 of the buffer size, the gap between the throughput of the two SFs of the rtPS CoS reduces. So, better fairness is achieved among the SFs of the same CoS when the Q_0 is set at a small level as also shown in Figure 4.15.

Table 4.2. Throughputs (kbps) of Two-Flows of rtPS with various values of BUR

Time	BUR - 1/2		BUR - 1/4		BUR -1/8		BUR - 1/16		BUR-1/32	
Sec	rtPS-1	rtPS-2	rtPS-1	rtPS-2	rtPS-1	rtPS-2	rtPS-1	rtPS-2	rtPS-1	rtPS-2
20	117	211	145	210	179	214	200	216	213	215
30	181	225	205	227	219	229	228	231	230	231
40	201	228	216	230	226	231	233	235	233	233
50	210	230	221	231	230	234	235	237	234	234
60	215	230	224	231	232	235	237	238	235	235
70	218	231	227	233	234	236	237	238	235	235
80	220	231	229	234	235	237	237	238	235	235
90	222	232	230	235	236	238	238	238	235	235

Table 4.2 shows if the target operating point is set to a very small value like 1/32 of the buffer size, the throughput of the network reduces for rtPS. Since, to keep the queue length around a very low target operating point, WFICC reduces the expected rate of each CoS to a very low value followed by a decrease in the throughput and an increase in free slots (**Figure 4.13**).

Average Delay

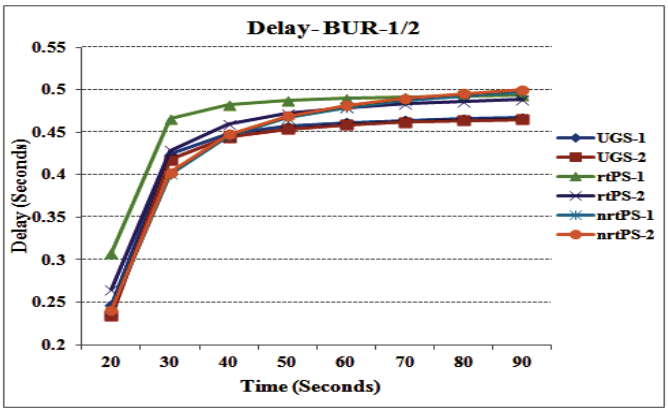


Figure 4.16 Average Delay (sec) of Two-Flows per CoS with BUR-1/2

Figure 4.16 and Figure 4.17 clearly demonstrate that the average delay for each class of service reduces as we decrease the target operating point from 1/2 to 1/16 of the buffer size. It is

because when the target point is set at a low level, WFICC starts its operation relatively early. It takes resources from the low priority CoSs such as BE and assigns them to the high priority CoSs to meet their respective MRTR. In this way, packet delay for high priority services reduces. When the target point is set at lower level, the packets queuing delay also reduces. Consequently, these factors contribute to the reduced delay with lower values of the target operating point.

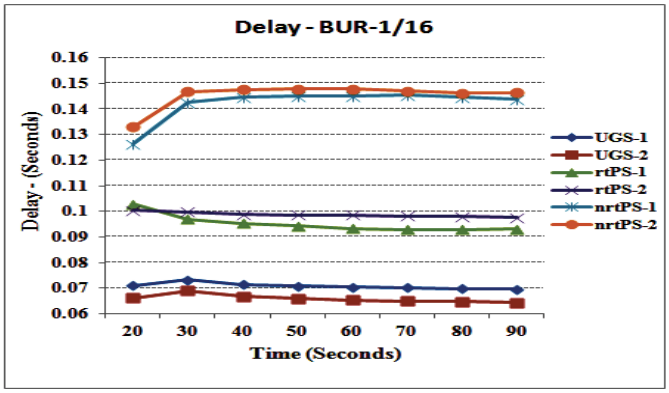


Figure 4.17 Average Delay (sec) of Two-Flows per CoS with BUR–1/16

The average delay for one service flow of rtPS and nrtPS CoSs at different levels of Q_0 is given in Table 4.3.

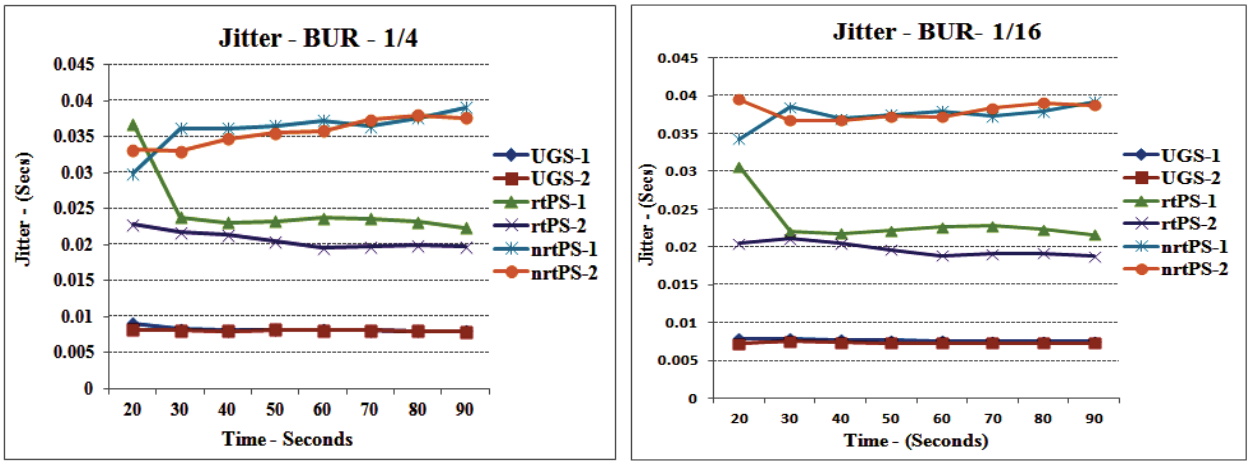
Table 4.3. Average Delay (sec) of rtPS and nrtPS Service Flows with various values of BUR

Time	BUR - 1/2		BUR - 1/4		BUR -1/8		BUR - 1/16		BUR-1/32	
Secs	rtPS-1	nrtPS-1	rtPS-1	nrtPS-1	rtPS-1	nrtPS-1	rtPS-1	nrtPS-1	rtPS-1	nrtPS-1
20	0.31	0.24	0.23	0.20	0.15	0.16	0.10	0.13	0.07	0.11
30	0.47	0.40	0.26	0.28	0.15	0.19	0.10	0.14	0.07	0.11
40	0.48	0.45	0.27	0.29	0.15	0.20	0.10	0.14	0.07	0.12
50	0.49	0.47	0.27	0.30	0.15	0.20	0.09	0.14	0.06	0.12
60	0.49	0.48	0.27	0.30	0.15	0.20	0.09	0.14	0.06	0.12
70	0.49	0.49	0.27	0.31	0.15	0.20	0.09	0.15	0.06	0.12
80	0.49	0.49	0.27	0.31	0.15	0.20	0.09	0.14	0.06	0.12
90	0.49	0.50	0.27	0.31	0.15	0.20	0.09	0.14	0.06	0.12

Table 4.3 shows that with larger values of the target operating point, the network suffers longer delays, as it takes longer to drain packets from the queue. While with a lower Q_0, system can achieve smaller delays for all CoSs.

Jitter

Figure 4.18 shows the jitter of the system for Q_0 set at 1/4 and 1/16 of the buffer size, respectively.



(a) (b)

Figure 4.18 Jitter (sec) of Two-Flows per CoS (a) with BUR-1/4 (b) with BUR-1/16

Figure 4.18 indicates that the jitter for the UGS CoS is unaffected because WFICC does not operate on UGS. The jitter for rtPS and especially nrtPS CoS is slightly higher when the target point is set at lower levels (Figure 4.18 (a)). As discussed earlier (Figure 4.11), when the target point is set at lower levels, the queue length fluctuates highly around the target point and results in higher jitter.

Hence, when the network is congested, the queue at the BS buffer increases and if at that time the scheduler allocates resources without taking into consideration the remaining buffer capacity, scheduled traffic is dropped and results in low throughput (Figure 4.4). So, zero free slot does not truly indicate that the network is utilized efficiently and the throughput is satisfactory.

An effective congestion control scheme ensures the scheduler schedules resources taking into account the load in the network and achieve better throughput, delay and fairness even with some free slots existed in the network.

4.4.2 Impact of Over Sell Factor (α)

Secondly, we investigate the effect of the oversell factor (α) on the system performance. To investigate 'α', Q_0 is set at the 1/4 of buffer size and the parameter β is set at 0.3. Figure 4.19 shows queue lengths for different values of 'α'.

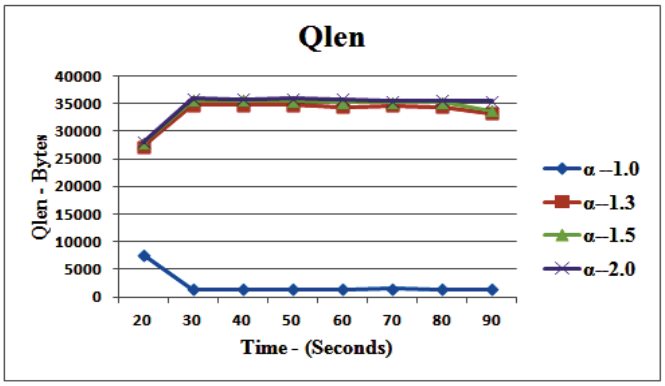


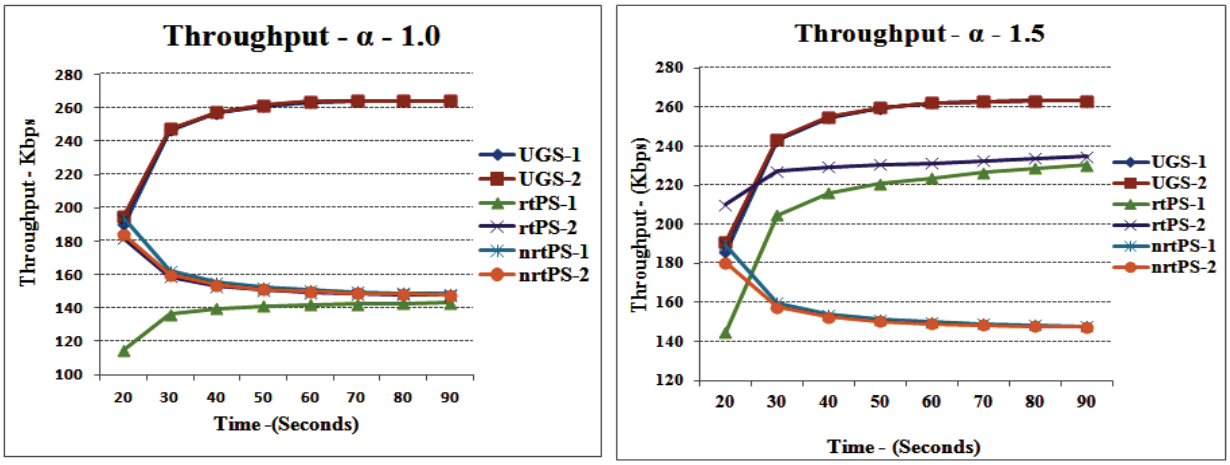
Figure 4.19 Queue lengths (Bytes) with various values of α

This critical parameter is used by WFICC to oversell resources to service flows that have data to send when the queue length is less than the target operating point. In this way, the system resources can be better utilized. When the network is operating below the target operating point and α is set at 1.0, the system does not oversell the resources and the resource utilization remains at a very low level. Consequently, there is a very low level of queue at the BS buffer as illustrated in Figure 4.19. The results show that if we increase the value of α above 1.0, the system allocates more resources to connections that are willing to take them. This results in better utilization of resources and also the network operates around the target operating point.

Figure 4.19 demonstrates that as we increase α larger than 1.1, there is also an increase in the queue length but this increase is not very high even for α equals to 2.0. It is because WFICC oversells resources and hence pushes the queue length only until it reaches the target operating point and maintains it around this point. The effect of oversell factor on the system performance is discussed in terms of throughput and delay in the following section.

Throughput

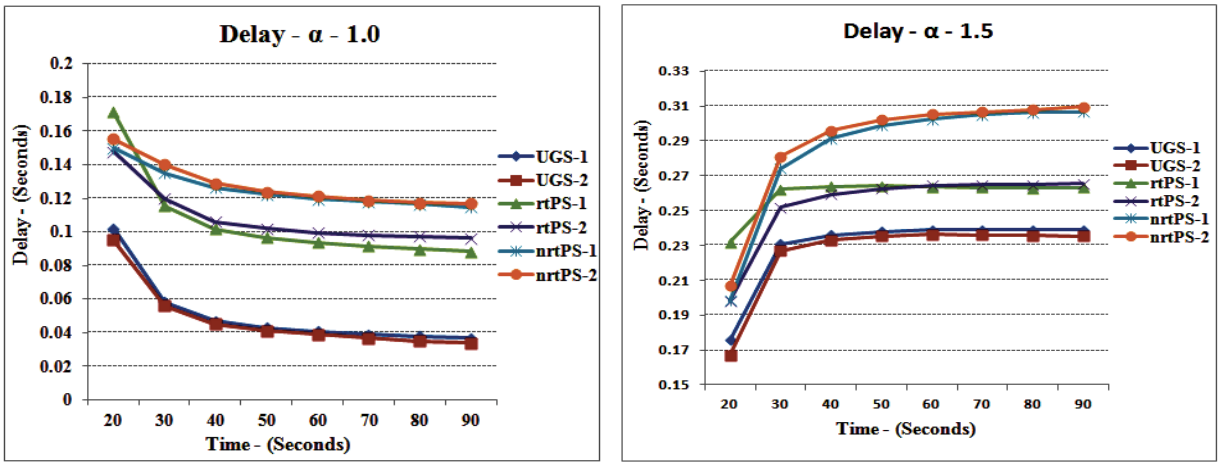
Figure 4.20 shows the throughput of the network at α equal to 1.0 and 1.5, respectively.



(a) (b)

Figure 4.20 Throughputs (Kbps) of Two-Flows per CoS (a) with α -1.0 (b) with α- 1.5

Figure 4.20 illustrates that when α is 1.0, mean no overselling of bandwidth even when the network operates below the target point, the throughput of the rtPS is very low (around 150 kbps) compared to the throughput with α equal to 1.5 (around 230 kbps). When the value of oversell factor 'α' is set at 1.0 and 1.5, respectively, we don't see any effect on the throughput of UGS CoS, as WFICC does not restrict or encourage the bandwidth allocated to UGS connections.



(a) (b)

Figure 4.21 Average Delay (sec) of Two-Flows per CoS (a) with α -1.0 (b) with α- 1.5

Figure 4.21 shows the average delay of various CoSs at α equal to 1.0 and 1.5, respectively. When α is set at 1.0, the network does not oversell the network resources even when the network operates below the target operating point. This results in very low level of queue at the BS as shown in Figure 4.19. Hence, the average delay for all CoSs is very low compared to the delay with α equal to 1.5 at cost of very low throughput.

Table 4.4 provides the throughput of the service flows of rtPS and nrtPS CoSs for various values of α.

Table 4.4. Throughputs (Kbps) with Various Levels of α

	CBR α-1.3	CBR α-1.5	CBR α-2.0	rtPS α-1.3	rtPS-α-1.5	rtPS-α-2.0	nrtPS-α-1.3	nrtPS-α-1.5	nrtPS-α-2.0
20	181.1	185.8	185.8	139.2	144.6	144.5	192.2	189.7	189.6
30	240.9	243.0	242.9	202.3	204.7	204.7	161.1	159.9	159.9
40	254.0	254.6	254.6	215.4	216.2	216.3	154.3	153.9	153.9
50	259.3	259.6	259.6	220.7	221.0	221.1	151.5	151.3	151.3
60	262.1	262.3	262.3	223.5	223.8	223.8	150.0	149.9	149.9
70	263.0	263.0	263.0	226.3	226.6	226.6	149.1	149.0	149.0
80	263.2	263.1	263.1	228.6	228.8	228.8	148.4	148.4	148.4
90	263.3	263.3	263.2	230.3	230.5	230.5	148.0	147.9	147.9

Table 4.4 shows that as we increase α larger than 1.1, there is also an increase in the throughput but the increase is not very high even for α equal to 2.0 because WFICC encourages the rate until the queue length reaches the target operating point and keeps it around this point (Figure 4.19).

4.4.3 Impact of Exponential Average Factor (β)

The exponential Average Factor (β) is used in the calculation of the $MACR_{CoS}$ and determines how fast the $MACR_{CoS}$ converges to the ACR_{CoS}**.** To investigate the parameter β, Q_0 is set at 1/4 of the buffer size and the parameter α is set at 1.3.

Figure 4.22 shows the queue length is the same for β values set at 0.1, 0.3 and 0.6, respectively. It reflects that change in the value of β does not have high impact on the queue length and hence the overall throughput, delay and jitter of the network. However, the value of β

impacts the computational performance of WFICC. A small value of β results in long computational time for the $MACR_{CoS}$ to reach the ACR_{CoS}.

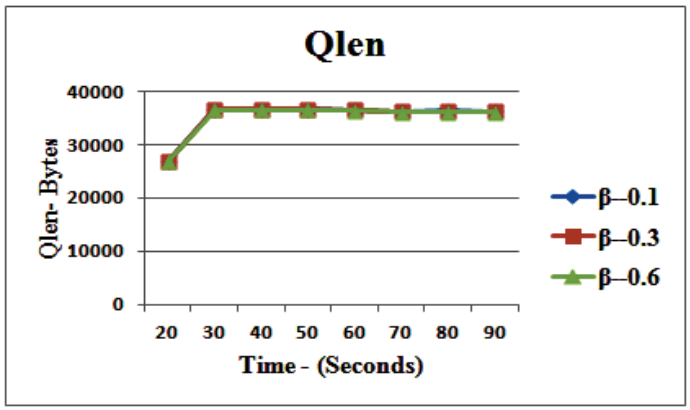


Figure 4.22 Queue lengths (Bytes) with various values of β

Figure 4.23 shows the total number of iterations used in the process of upgradation and degradation, respectively, throughout the simulation.

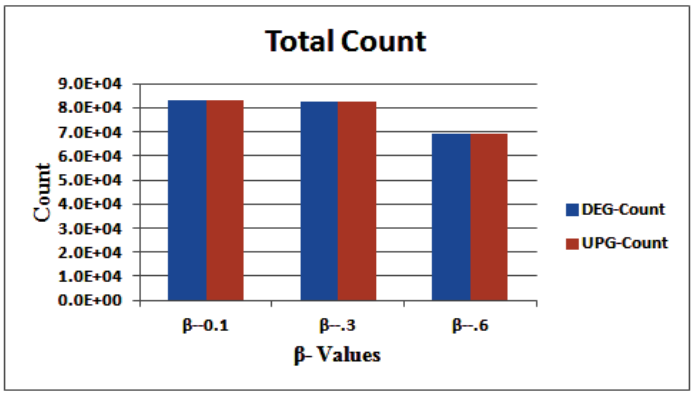


Figure 4.23 Total count of Upgradation and Degradation

Figure 4.23 clearly demonstrates that as the value of β increases, the step size for conversion of the $MACR_{CoS}$ to the ACR_{CoS} increases, which results in less number of iterations compared to β with a small value.

4.5 Discussion on Results

The simulation study focused on the performance of the network and hence the effectiveness of WFICC for various settings of the parameters including Q_0, α and β. The results showed that WFICC is robust in a way that changes in the parameters do not affect the performance of the network widely. It achieves high network performance without violating the QoS requirements of all class of services. Most importantly, under WFICC, fairness is achieved among all class of services.

Consistency

Simulations results demonstrate that WFICC is consistent in obtaining the network performance in terms of high throughput, high link utilization, small queue lengths, small average packet delays and small packet delay variations.

Fairness

All simulations results clearly demonstrate that WFICC estimates accurately and consistently the fair share of the bandwidth of all aggregates of different CoSs according to their respective priority without much deviation from their true fair share under all network conditions. WFICC also ensures that connections of different CoSs get a fair share in accordance to their QoS constraints such as priority, MSTR and MRTR. As a result, a connection is given a share neither less than the MRTR, nor more than the MSTR of its respective CoS. The oversell aspect of WFICC lets the unconstrained aggregates to use the residual bandwidth that cannot be used by the constrained connections. It also ensures the bandwidth is fairly shared among the unconstrained connections. As a result, it ensures that packets that have already taken buffer represent a good mix in terms of fair share of aggregates of different CoSs.

Bounded queue length

A distinguishing feature of WFICC is that it keeps a queue length around a target level and ensures that output link is always busy. WFICC is able to achieve this even when the number of

connections of different CoSs that are using the buffer increases. Hence, WFICC is more scalable in this aspect.

Parameter Setting Sensitivity

Target operating point (Q_0)

The Q_0 was chosen within the range of 1/2 to 1/32 of the buffer size. The results have shown that when BUR is set at small level not less than 1/32, the network experiences relatively less delay, better fairness among the different CoSs and also among the service flows of the same CoS.

When Q_0 is set at high level like 1/2 of the buffer size, throughput is relatively high at the cost of high delay and less fairness among the service flows. The overall throughput of the network reduces when Q_0 is set at a low level but as we keep reducing Q_0 from 1/4 to 1/16 of the buffer size, the throughput of high priority CoS such as rtPS increases while the throughput of low priority CoS such as nrtPS decreases. Since, an early execution of WFICC ensures that connections of each CoS are assigned bandwidth according to their respective priority and MRTR. It assigns left over resources fairly first to connections of the high priority CoSs and then to the connections of low priority CoSs.

So as discussed in the analysis the target operating point is critical as it determines when WFICC starts its operation. The changes in the Q_0 reflect some changes in the throughput, delay and fairness, but minor changes in Q_0 produced minor changes in the QoS of the network as illustrated by the simulation results. Thus, WFICC performed well under a wide range of variations in Q_0 parameter.

Oversell factor (α)

The parameter α is selected in the range of 1.0 to 2. In conformity with the theoretical analysis, simulations results clearly indicate no overselling of resources when α is set at 1.0. Hence, α needs to be set at a value larger than 1.0. With an increase of α, the maximum queue length increases only up to the defined target operating point and remains around this level. Different

values of α above 1.1 do not cause any big difference in the maximum queue length and hence in the throughput and the average packet delay.

Exponential average factor (β)

The averaging factor is selected in the range of 0.1 to 0.6. In compliance with the theoretical analysis, simulation results illustrate that the parameter β influences how quickly $MACR_{CoS}$ converges to ACR_{CoS}. So, if the parameter β is set at small value the network takes more time to reach to a value expected by the network.

4.6 Summary

In this chapter, a new CC algorithm is proposed for WiMAX networks. It is demonstrated to perform effectively and efficiently to avoid and control load at the core network. Instead of using thresholds to reduce the network congestion, it employs a target operating point. It estimates the accurate level of fair share for all class of services, with the aim to maintain the desired target operating point. It maintains the network traffic around the target point, hence avoids congestion and loss at the base station output buffer. In this chapter, we have discussed the impact of various settings of the parameters of WFICC on the network performance. We also discuss the value selection of these parameters for the efficient load control and optimal performance. It employs only few parameters, which do not require complex fine-tuning. The simulations results demonstrated that WFICC is robust because it is relatively insensitive to parameter settings.

Chapter 5 Fair Intelligent Congestion Control for LTE Networks (LTE-FICC)

This chapter introduces a mechanism for controlling congestion at an eNodeB, the *LTE Fair Intelligent Congestion Control (LTE-FICC)*. LTE-FICC deals with both, the unfair bandwidth allocation and the congestion issues encountered in the LTE networks. It maintains the network traffic around a target operating point, hence avoiding congestion and loss at an eNodeB's output buffer. LTE-FICC uses a rate allocation scheme, which takes into account the degree of congestion at an eNodeB's output buffer. It estimates the fair share of all QoS classes and sends the estimated rates as a feedback to the underlying schedulers at an eNodeB.

Section 5.1 presents the overall system architecture designed to implement LTE-FICC effectively in LTE networks. Section 5.2 presents the modified scheduler proposed to operate with LTE-FICC. Section 5.3 comprehensively describes the CC algorithm. Section 5.4 presents the simulation model and discusses the issues in the current simulator. Section 5.5 provides the simulation setup and Section 5.6 presents the evaluation of the proposed scheme. Section 5.7 discusses the results. Finally, Section 5.8 summarizes this chapter.

5.1 Overall System architecture

As discussed earlier (Section 2.3), in this thesis the QoS classes of LTE networks are grouped into Class of Bearers (CoBs) according to their resource types; GBR and non-GBR. The GBR CoB includes priority levels ranging from 2 to 5 and the non-GBR CoB consists of priority levels ranging from 6 to 9. These two groups are defined for the sake of simplicity. LTE-FICC estimates the expected rate for each CoB only. The scheduler estimates the rate for each QCI by applying a scheduling weight to the expected rate. The scheduling weight is based on the priority of a QCI. Hence, a higher scheduling weight is assigned to QCIs with higher priority. Furthermore, the bandwidth allocated to GBR bearers above their GBR requirements is hereafter referred to as GBR_Ad.

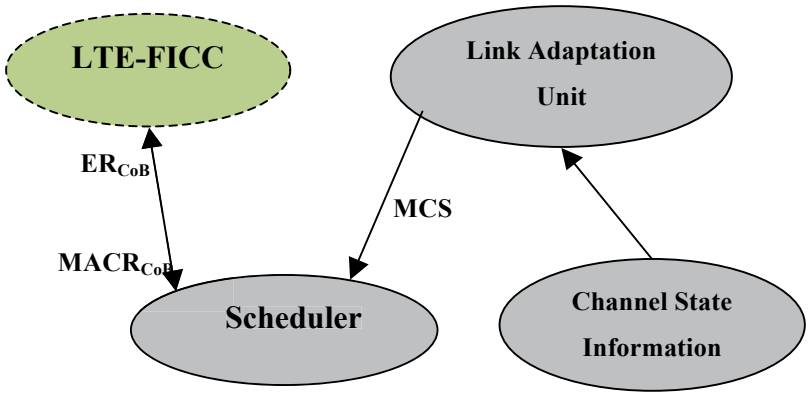


Figure 5.1 Coordination between LTE-FICC, Scheduler and Link adaptation unit

Figure 5.1 indicates that LTE-FICC passes the new estimated expected rate of each CoB to the scheduler. The scheduler estimates the rate of a connection with a specific priority by applying a scheduling weight to the expected rate of its CoB. It obtains the Modulation and Coding Scheme (MCS) of a User Equipment (UE), determined by an Adaptive Modulation and Coding (AMC) function of a Link Adaptation Unit (LAU), based on its channel conditions. Based on the MCS, the scheduler calculates the number of Physical Resource Blocks (PRBs) required to provide a connection the estimated rate for its QCI's priority level. LTE-FICC is simple because it employs a small number of parameters. It operates with any basic underlying scheduling algorithm to minimize the delay and maximize the throughput.

To evaluate the effectiveness of the proposed algorithm, simulations are performed in OPNET (OPNET, 2012) using LTE module.

5.2 Modified Round Robin (MRR)

In the LTE module of OPNET an uplink frame generator is used to generate a subframe. The frame generator uses two rounds to allocate resources to connections. In the first round, it allocates resources first to GBR bearers to meet their respective GBR and later to non-GBR bearers. The LTE module of OPNET uses a Proportional Fair (PF) scheduler to allocate resources to GBR connections to meet their GBR. The PF scheduler in OPNET allocates resources to GBR bearers based on their GBR and delay requirements. In the second round, the frame generator allocates remaining PRBs to GBR bearers to gain their MBR. It uses either an

equal capacity sharing algorithm, or a Round Robin (RR) algorithm, to allocate resources to non-GBR connections in the round one and to GBR connections above their GBR requirements in the round two of the frame generation.

The equal capacity queue selection algorithm given in LTE module of OPNET simulator always starts with the highest priority QCI in each CoB (GBR or non-GBR). In a subframe, it shares resources equally among 3 connections, while remaining connections are given resources in the next subframe. So, for an instance, when there are 4 connections with a particular QCI, there will be unfair resource allocation among the connections. It is because the equal capacity scheduler shares resources equally among 3 connections in one subframe and the fourth connection is given the whole next subframe. Furthermore, the equal capacity queue selection algorithm can lead to the situation when connections of low priority QCIs in each CoB are not granted resources for the transmission of data. As in every subframe the algorithm starts with the highest priority QCI in each CoB (such as QCI 1 in GBR CoB and QCI 6 in non-GBR CoB) and when all queues at this priority level are empty, it then moves to the next lower priority level.

The RR algorithm given in LTE module of OPNET simulator provides fairness among the connections at the same priority level in each CoB but similar to the equal capacity queue selection algorithm, the RR scheduler in every subframe starts with the highest priority QCI in each CoB and schedules connections from this level until all queues at this level are empty. Consequently, these basic schedulers result in imbalanced bandwidth allocation in the network. So, we modified the RR algorithm to ensure that connections at all priority levels are served.

In the modified RR algorithm (M-RR), once a queue from a specific priority level is served, the RR algorithm moves to the next lower priority level. To maintain differentiation among different QCIs of a specific CoB, the modified round robin scheduling algorithm assigns weight to each QCI based on its priority level. So, the scheduler assigns resources to connections at any priority level equal to ER_{CoB} **X** weight[QCI] or the maximum class rate, whichever is the minimum.

The modified round robin algorithm always starts from the highest priority QCI in each CoB and in a RR manner scans to find a non-empty queue for scheduling. Once a queue is scheduled

or there is no queue with data at that level, the algorithm moves to the next lower priority QCI. The algorithm continues to the next lower priority levels and scan all queues until a non-empty queue is found. In this way, the M-RR scheduler ensures that queues from all priority levels are scheduled and also maintains service differentiation according to the assigned weights.

5.3 Congestion Control Algorithm for LTE

In the LTE architecture, the basic schedulers allocate resources without taking into consideration congestion at an Evolved NodeB's (eNodeB) output buffer. This leads to buffer overflows and deterioration in the overall QoS of connections. Congestion avoidance and fair bandwidth allocation is hardly considered in existing research for LTE uplink connections.

The authors (Vulkan. and Heder., 2011) presented congestion control (CC) scheme to provide fairness in heterogeneous radio access networks based on Radio Link control (RLC) packet discard mechanism. A congestion control mechanism is proposed by (Kwan. et al., 2010) that mitigates load in the network by removing the low priority bearers until the system load reaches to a predefined target value. They do not clearly discuss how the target load can be defined for a network. The authors (Qiu. et al., 2011) proposed a congestion control mechanism that protects an eNodeB's output buffer from overflow by controlling the TCP advertisement window. The RLC layer in the eNodeB monitors the buffer utilization and sets the congestion flag once it reaches to the threshold. So the proposed scheme violates the protocol layer design principles. The researchers (Zolfaghari. and Taheri., 2012) proposed queue aware scheduling technique. They presented the performance of various queue aware scheduling schemes with an end-to-end congestion control scheme that controls the rate of elastic traffic and consequently affects the buffer status. The scheme is proposed for downlink only. The end-to-end congestion scheme used in their work involves the overhead of congestion field that is added in every packet. The above mentioned schemes to perform load balancing are either based on thresholds or are applicable to a specific protocol, such as TCP only. Furthermore, they merely discuss the fair bandwidth allocation among flows of the same and different CoBs.

This section introduces our proposed mechanism, the LTE Fair Intelligent Congestion Control (LTE-FICC), to control congestion at an eNodeB. LTE-FICC jointly exists with a scheduler at an

eNodeB to guarantee efficient traffic scheduling. It also overcomes the problem of unfair bandwidth allocation among the flows that share the same eNodeB interface. LTE-FICC defines a target operating point at a preset Buffer Utilization Ratio (BUR). In order to make an output buffer operate around the target operating point, it estimates the current mean bandwidth allocated to connections of each CoB (GBR and non-GBR) and calculates the expected rate for each CoB using the queue control function.

5.3.1 Queue Control function (f (Q))

The output queue status of an eNodeB depends on the output link capacity and the capacity of an Evolved Packet Core (EPC). So, a large queue length at the eNodeB buffer serves as an indication that the EPC is congested. To detect congestion at the EPC, LTE-FICC uses the same queue control function as given in Figure 3.3. It obtains the value of the queue length from the eNodeB's output interface.

5.3.2 Mean Allowed Class Rate of Each Class of Bearer ($MACR_{CoB}$)

To estimate the average bandwidth allocated to the flows of each CoB, LTE-FICC similar to FICC, uses a variable named MACR but one for each CoB ($MACR_{CoB}$). We suggest two MACRs corresponding to the two classes of bearers, $MACR_{GBR\_Total}$ and $MACR_{non\_GBR}$. $MACR_{GBR\_Total}$ maintains the average value of the rate allocated to all active connections of GBR CoB. It is obtained as the sum of $MACR_{GBR}$ and $MACR_{GBR\_Ad}$ and is estimated as follows.

$$MACR_{GBR\_Total} = MACR_{GBR} + MACR_{GBR\_Ad} \quad 5.1$$

In Eq. 5.1, the $MACR_{GBR}$ is an estimate of the average bandwidth allocated to all data flows of the GBR CoB to meet their respective GBR. The $MACR_{GBR\_Ad}$ maintains the average value of the additional bandwidth allocated to all connections of the GBR CoB above their individual GBR requirements to gain their respective MBR. The GBR bearers always obtain GBR even when the network is congested. Therefore, LTE-FICC does not need to operate on $MACR_{GBR}$. It operates only on $MACR_{GBR\_Ad}$, as it is the only part of resource allocation to GBR bearers that can be controlled by LTE-FICC. The $MACR_{non\_GBR}$ maintains the mean value of the bandwidth allocated to all active connections of the non-GBR CoB. The scheduler at the eNodeB updates

the values of $MACR_{GBR}$ same as in Figure 3.2. However, the scheduler allocates resources to the GBR bearers above their GBR requirements and the non-GBR bearers using the assigned weights. So, the scheduler updates the values of $MACR_{GBR\_Ad}$ and $MACR_{non\_GBR}$ using Eq. 5.2.

In Eq. 5.1, the $MACR_{GBR}$ is an estimate of the average bandwidth allocated to all data flows of the GBR CoB to meet their respective GBR. The $MACR_{GBR\_Ad}$ maintains the average value of the additional bandwidth allocated to all connections of the GBR CoB above their individual GBR requirements to gain their respective MBR. The GBR bearers always obtain GBR even when the network is congested. Therefore, LTE-FICC does not need to operate on $MACR_{GBR}$. It operates only on $MACR_{GBR\_Ad}$, as it is the only part of resource allocation to GBR bearers that can be controlled by LTE-FICC. The $MACR_{non\_GBR}$ maintains the mean value of the bandwidth allocated to all active connections of the non-GBR CoB. The scheduler at the eNodeB updates the values of $MACR_{GBR}$ same as in Figure 3.2. However, the scheduler allocates resources to the GBR bearers above their GBR requirements and the non-GBR bearers using the assigned weights. So, the scheduler updates the values of $MACR_{GBR\_Ad}$ and $MACR_{non\_GBR}$ using Eq. 5.2.

$$if\ Qlen > Q_0$$

$$if\ \frac{ACR_{QCI}\,(t)}{Weight[QCI]} < MACR_{CoB}(t-1)$$

$$MACR_{CoB}(t) = MACR_{CoB}(t-1) + \beta * \left(\frac{ACR_{QCI}(t)}{Weight[QCI]} - MACR_{CoB}(t-1)\right)$$

$$else\ if\ Qlen < Q_0$$

$$MACR_{CoB}(t) = MACR_{CoB}(t-1) + \beta * \left(\frac{ACR_{QCI}(t)}{Weight[QCI]} - MACR_{CoB}(t-1)\right) \quad 5.2$$

In Eq. 5.2, $MACR_{CoB}(t-1)$ and $MACR_{CoB}(t)$ reflects the previous and the new values of the mean of the Allowed Class Rate (ACR) of all active connections of each CoB. When the network load exceeds the target operating point that is the queue length is more than the target operating point, LTE-FICC does not allow $MACR_{CoB}(t)$ to increase further. Therefore, $MACR_{CoB}(t)$ does

not track any value of the current allocation ($ACR_{QCI}(t)$/weight[QCI]) larger than the previous mean allowed class rate ($MACR_{CoB}(t-1)$).

In Eq. 5.2, $ACR_{QCI}(t)$ represents the actual bandwidth allocated to the active connections of each priority level stated by its QCI. $ACR_{QCI}(t)$ is estimated as follows.

$$ACR_{QCI}(t) = \quad ER_{CoB}(t) * Weight[QCI] \tag{5.3}$$

In Eq. 5.3, Weight [QCI] is the weight assigned by the scheduler to each QCI based on its priority level as discussed in section 5.2. In this chapter QCIs 1, 2, 3 and 4 from GBR CoB and QCIs 6, 7, 8 and 9 from non-GBR CoB are assigned weights of 4, 3, 2 and 1, respectively. The $ER_{CoB}(t)$ is the expected rate of each CoB and is estimated by LTE-FICC as follows.

$$ER_{CoB}(t) = \quad MACR_{CoB}(t-1) * f(Q) \tag{5.4}$$

LTE-FICC obtains the estimates of $MACR_{CoB}(t-1)$ from the scheduler (Figure 5.1). It then estimates the degree of network congestion by calculating the queue control function (f(Q)).

In LTE networks, a scheduler grants resources in every subframe 'n' for the $n+4^{th}$ uplink subframe. LTE-FICC is executed in every subframe to provide a feedback to the scheduler about the rate to allocate to each connection based on its priority, the QoS requirements and the load in the network.

To keep the network traffic around the target operating point, LTE-FICC estimates the fair share of bandwidth of each CoB. In situations when LTE-FICC estimates that the core network is congested, as indicated by the value of queue control function, it applies a degradation procedure to discourage traffic in the network. As soon as, it estimates that the core network is non-congested, it applies an upgradation procedure to encourage the network traffic.

5.3.3 Degradation Procedure

When the queue length at an output buffer of an eNodeB operates beyond the target operating point, it risks buffer overflow and high delays in the network. In this scenario, LTE-FICC applies a degradation procedure on the existing connections to control their current rate. LTE-FICC

initially degrades the rate allocated to non-GBR CoB. In case when LTE-FICC reduces the rate allocated to non-GBR CoB to its minimum and it still estimates congestion in the network, it reduces the rate of GBR connections above their GBR requirements (GBR_Ad). The detail of each step of the degradation procedure of LTE-FICC is as follows.

Non-GBR connection

When f(Q) is less than 1, it indicates that there is congestion in the network. In this situation, LTE-FICC reduces the expected rate of non-GBR CoB and requires the scheduler to reduce the bandwidth allocation to all non-GBR bearers. The step size of degradation is controlled by f(Q) as it provides an estimation of how much degradation is required to apply on this class of traffic. Based on the degradation applied on non-GBR bearers, it updates the estimates of queue length, queue control function, $MACR_{non\_GBR}$ and ER_{non_GBR}. LTE-FICC continues reducing the ER_{non_GBR} until either the value of f(Q) is equal to or greater than 1, or the ER_{non_GBR} reaches the minimum rate that must be allocated to a connections in UL (MIN_UL).

GBR Connections

After degrading the expected rate of non-GBR CoB, LTE-FICC again estimates the queue control function with the updated queue length. If f(Q) is again less than 1 indicating the network is still operating above the target operating point, LTE-FICC reduces the expected rate of GBR_Ad of GBR bearers. In this way, it ensures that the scheduler reduces the rate allocated, above the GBR requirements (GBR_Ad), to all connections of GBR CoB. The GBR bearers always obtain GBR even when the network is congested. Therefore, LTE-FICC degrades only the rate allocated above the GBR requirements. LTE-FICC based on the degradation applied on GBR bearers, updates the estimates of the queue length, queue control function, $MACR_{GBR\_Ad}$ and ER_{GBR_Ad}. LTE-FICC continues reducing the ER_{GBR_Ad} until either the value of f(Q) is equal to or greater than 1, or the ER_{GBR_Ad} reaches the minimum rate that must be allocated to a bearer in UL (MIN_UL).

When the value of the queue control function becomes equal to or greater than 1, it indicates that with the applied degradation, the network will manage to bring the queue length close to the target operating point. Consequently, the network avoids buffer overflow and increasing delay.

5.3.4 Upgradation Procedure

When the queue length at an output buffer of an eNodeB operates below the target level, it risks network resource underutilization. LTE-FICC in this situation applies an upgradation procedure on the expected rate of each CoB. Consequently, the scheduler by allocating more resources to connections encourages them to send more traffic to the network. LTE-FICC initially upgrades the rate allocated to GBR connections above their GBR requirements. In case when LTE-FICC increases the rate allocated to GBR CoB to its maximum and still estimates that the core network can take more traffic, it increases the rate of non-GBR connections. The detail of each step of the upgradation procedure of LTE-FICC is as follows.

GBR Connections

If f(Q) is greater than 1, indicating the network is operating below its target level, LTE-FICC raises the expected rate of GBR_Ad of GBR bearers. In this way, LTE-FICC ensures that the scheduler increases the additional bandwidth share above the GBR requirements of all GBR bearers. The step size for the share increase is determined by the function of queue but at a different rate indicating how far the network is being underutilized. LTE-FICC based on the upgradation applied on the GBR bearers, updates the estimates of queue length, f(Q), $MACR_{GBR\_Ad}$ and ER_{GBR_Ad}. LTE-FICC keeps on increasing the ER_{GBR_Ad} until either f(Q) is equal to or less than 1, or the ER_{GBR_Ad} reaches the maximum GBR_Ad of the GBR CoB.

Non-GBR Connections

After upgrading the expected rate of GBR bearers, LTE-FICC recalculates the f(Q). If again f(Q) is greater than 1, indicating the network is still operating below its target level, LTE-FICC raises the bandwidth share of non-GBR CoB. LTE-FICC based on the upgradation applied on the expected rate of non-GBR bearers, updates the estimates of queue length, f(Q), $MACR_{non\_GBR}$ and ER_{non_GBR}. LTE-FICC continues increasing the ER_{non_GBR} until either f(Q) is equal to or less than 1, or the ER_{non_GBR} reaches the MBR of non-GBR CoB.

In the upgradation procedure, when the value of the queue control function becomes equal to or less than 1, it indicates that with the applied upgradation the network will manage to bring the

queue length close to the target operating point. Consequently, the network in times of non-congested state achieves stable throughput and minimum delay.

To simplify the information maintained by LTE-FICC, it does not maintain queue lengths of each CoB as in WFICC (refer to section 4.1). It approximates an impact of the change in the expected rate with respect to the mean allowed class rate ($MACR_{CoB} - ER_{CoB}$) on the actual queue length. To perform the approximation, it only considers the effect of the new expected rate on the maximum resource blocks, which can be allocated per subframe in the uplink with the current system bandwidth (RB_UL_subframe). This is due to the fact that the network actually allocates only these resource blocks to connections in every subframe and hence the effect of new expected rate is enforced on RB_UL_Subframe. Consequently, LTE-FICC updates the queue length using the following equation.

$$Qlen = Qlen - RB\_UL\_Subframe * (MACR_{CoB} - ER_{CoB}) \quad 5.5$$

Once the expected rate of non_GBR and GBR_Ad is determined for the next n+4th uplink subframe, LTE-FICC passes the values of ER_{non_GBR} and ER_{GBR_Ad} to the scheduler (Figure 5.1).

Algorithm of LTE-FICC

$MAX\_UL_{GBR\_Ad}$:= MBR-GBR, $MAX\_UL_{non\_GBR}$:= MBR

1. Obtain Value of queue length from an eNodeB output interface.
2. Obtain value of $MACR_{GBR\_Ad}$ and $MACR_{non-GBR}$ from scheduler.
3. Calculate f(Q) as follows.

IF (Qlen > Q_0)

$$f(Q) = \frac{(Buffer\_Size - Qlen)}{(Buffer\_Size - Q_0)}$$

Else

$$f(Q) = \frac{(\alpha - 1) * (Q_0 - Qlen)}{Q_0} + 1$$

End IF

Degradation Procedure

```
4. IF f(Q) < 1
     For each lowest to Highest priority CoB //Non_GBR and GBR_Ad
        ER_CoB := MACR_CoB *f(Q)            // Estimate Expected rate
        Do While ER_CoB > MIN_UL_CoB OR f(Q) < 1
           Qlen  := Qlen – RB_UL_Subframe *(MACR_CoB – ER_CoB)
           ACR_CoB := ER_CoB                // Update ACR
           Recalculate f(Q) given in step 3.
           Update MACR_CoB using Eq. 5.2.

           ER_CoB := MACR_CoB *f(Q)

        End Do While
     End For
  End IF
```

Upgradation Procedure

```
4. IF f(Q) > 1
     For each Highest to lowest priority CoB // GBR_Ad and Non_GBR
        ER_CoB := MACR_CoB *f(Q)            // Estimate Expected rate
        Do While ER_CoB < MAX_UL_CoB OR f(Q) > 1
           Qlen  := Qlen + RB_UL_Subframe *(ER_CoB - MACR_CoB)
           ACR_CoB := ER_CoB                // Update ACR
           Recalculate f(Q) given in step 3.
           Update MACR_CoB using Eq. 5.2.

           ER_CoB := MACR_CoB *f(Q)
        End Do While
     End For
  End IF
```

Figure 5.2 Algorithm of LTE-FICC

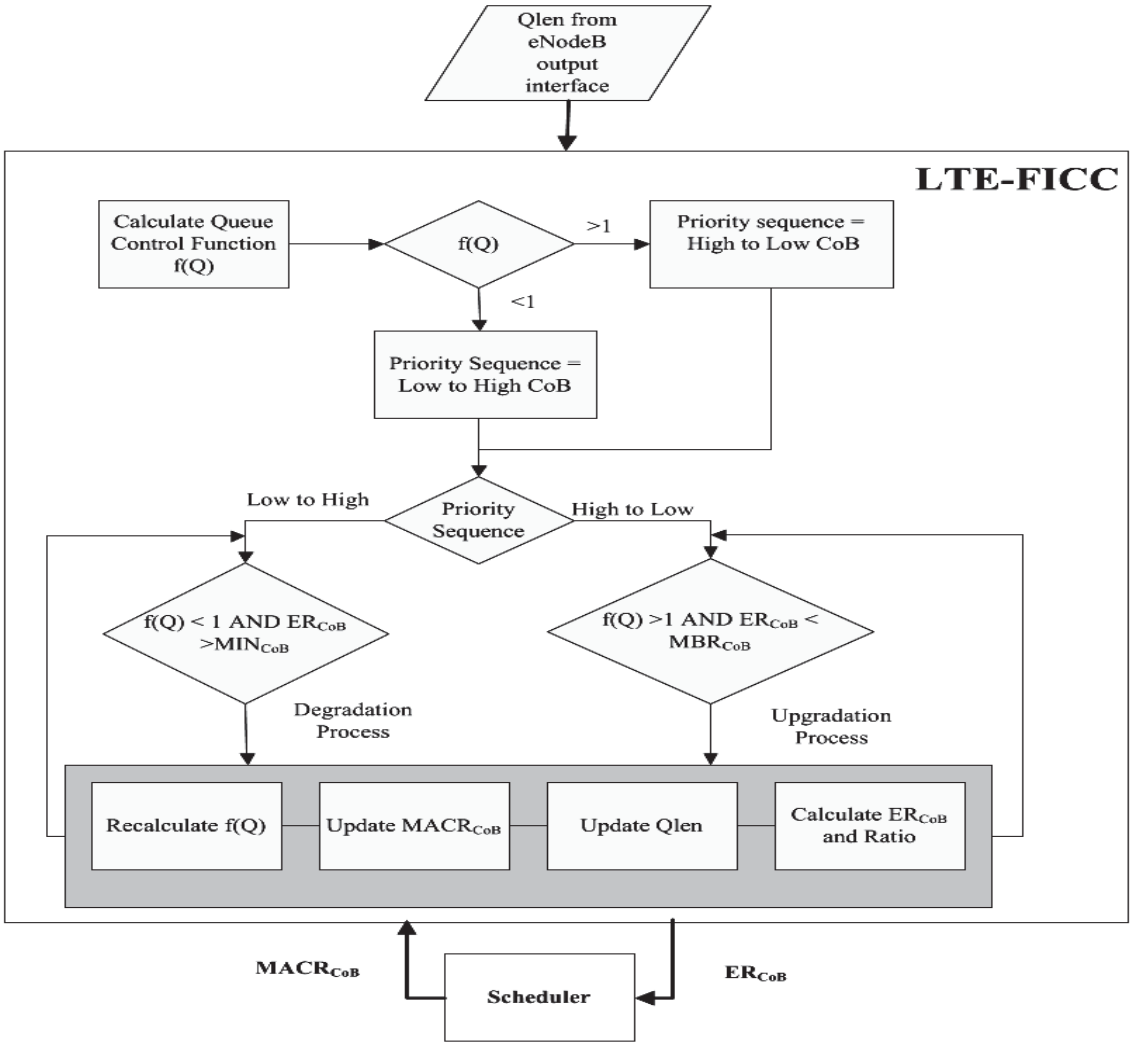


Figure 5.3 Flow chart of LTE-FICC

LTE-FICC:

1. Estimates congestion in the network using f(Q).

2. If f(Q) < 1, indicating network congested, degrade expected rate of each CoB.
   Apply degradation on expected rate of low to high priority CoBs, non-GBR and GBR_Ad, until estimated f(Q) is greater than or equal to 1 or expected rate has reached to minimum rate of CoB as shown in Figure 5.3.

3. If f(Q) > 1, indicating network underutilized, upgrade expected rate of each CoB.
   Apply upgradation on expected rate of high to low priority CoBs, GBR_Ad and non-GBR, until estimated f(Q) is less than or equal to 1 or MBR_{CoB} has reached as shown in Figure 5.3.

5.4 Simulation Model

The overall goal of the simulation is to analyze the performance of the proposed algorithm to meet the QoS requirements of each class of bearers in terms of fairness, throughput and delay in a congested scenario. The simulations have been performed in the system level simulator, Optimized Network Engineering Tool (OPNET) release 17.1.A (OPNET, 2012) using LTE module.

The OPNET modeler has a hierarchal structure. It consists of three models; LTE network model, node model and process model. To perform the simulation, all of these models are required to be configured. The LTE network model is given in Figure 5.4. The network model consists of UEs, an eNodeB and an EPC. The EPC entity combines the features of MME, S-GW and PDN-GW. The LTE Config node is used to define the configuration parameters of LTE networks such as EPS bearer definitions and LTE physical profile including bandwidth and frequency configurations. All LTE nodes in the network use these configurations.

Long Term Evolution (LTE) Architecture in OPNET

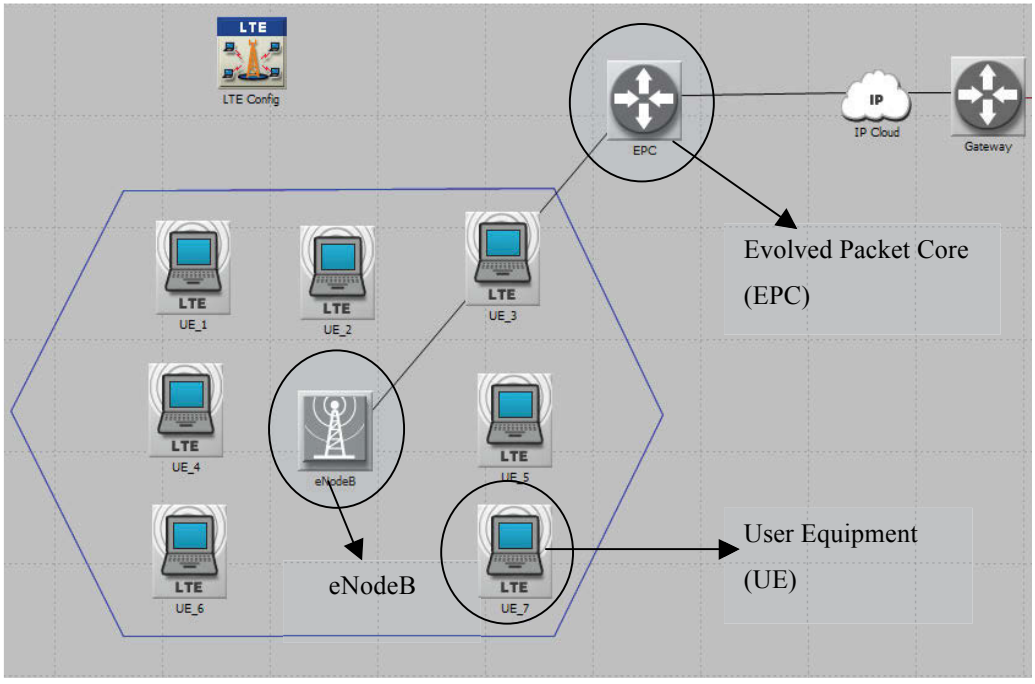


Figure 5.4 LTE Architecture in OPNET

Simulation Model entities:

Simulation model entities can be obtained from the ***Object palette*** in the ***project Editor*** window of OPNET modeler.

- Click on the ***Topology*** Menu and select the ***open Object palette*** from the menu.
- The ***Object palette*** window will appear.
- From Search box, find ***lte_adv tree*** and select desired node models.

Following figures show the configurable attributes of different LTE entities available in OPNET.

LTE EPC Node Models include

- lte_enodeb_atm8_ethernet8_slip8

Figure 5.5 EPC Configurable Attributes

LTE configuration node (LTE Config) Models include

- lte_ attribute_definer_adv

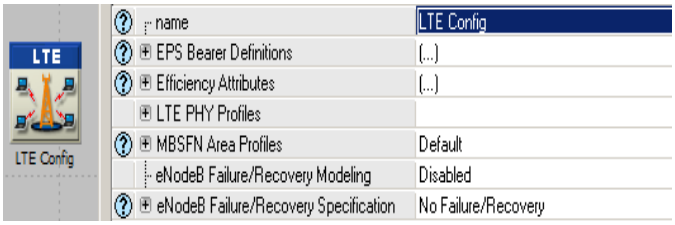


Figure 5.6 LTE Config node ConfigurableAttributes

LTE eNodeB Node Models include

- lte_enodeb_atm4_ethernet_slip4
- lte_enodeb_ ethernet4
- lte_enodeb_ slip4
- lte_enodeb_3sector_slip4
- lte_enodeb_6sector_slip4

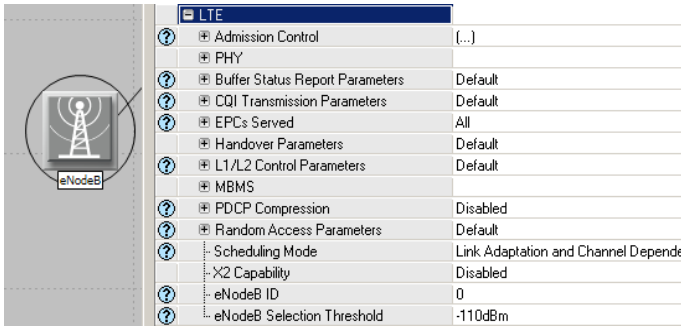


Figure 5.7 LTE eNodeB Configurable Attributes

Issues with LTE model

The OPNET release 17.1.A supports most features of the LTE release 8.0. The LTE module in OPNET assumes the same value for GBR and MBR. It admits the GBR connections through an admission control. The non-GBR bearers are admitted by default and their GBR is not guaranteed.

The LTE model of OPNET does not include any congestion control mechanism. The scheduler at an eNodeB keeps on scheduling connections without taking into consideration the load at the core network. As a result, when the core is overloaded, the queue at an output buffer of an eNodeB builds up. When the queue reaches the maximum buffer capacity, it drops packets in FIFO order. Consequently, the QoS of existing connections degrades. Additionally, schedulers at an eNodeB do not provide fair resource allocation when allocating resources to non-GBR and GBR bearers above their GBR requirements (Section 5.2).

Changes in LTE Model

- LTE module is updated to support the feature of LTE-Advanced to set the MBR attribute higher than the GBR of an EPS bearer. To implement these changes the lte_attribute_definer process model in OPNET is updated and is discussed in Appendix B.
- To handle the issues of the congestion and the unfair resource allocation, we added a CC module, LTE-FICC, at an eNodeB. We also enhanced the existing scheduler, the RR, so that during resource allocation it takes into account the feedback from the CC module to provide stable throughput, minimum delay and fair resource allocation in the network. To implement these changes, an eNodeB's node model needs to be updated. The following subsections focus on an eNodeB's node model and process models.

5.4.1 LTE eNodeB Node Model

Figure 5.8 shows the node model of an eNodeB in OPNET. Its gives an overview of all layers of 3GPP LTE protocol stack provided at an eNodeB by OPNET.

The node model of an eNodeB includes several process models. In Figure 5.8 we highlighted the two process models, *lte_s1* and *lte_enb_as* process models.

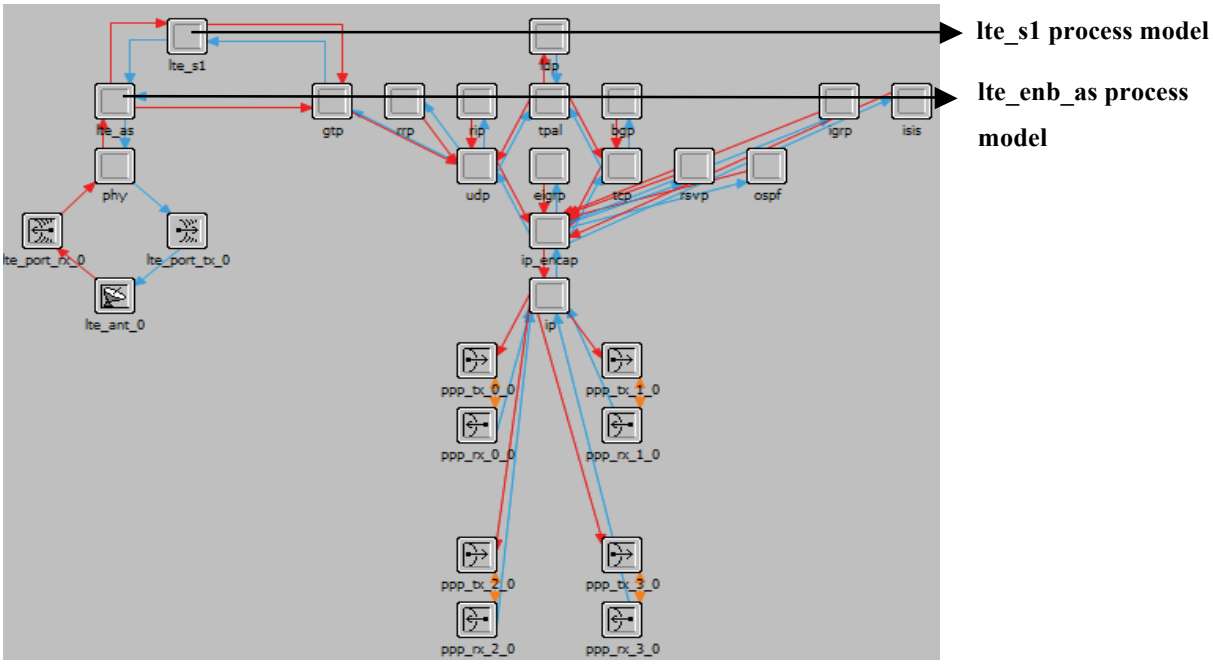


Figure 5.8 LTE eNodeB's Node Model

5.4.2 LTE eNodeB Process Models

5.4.2.1 S1 process Model of an eNodeB

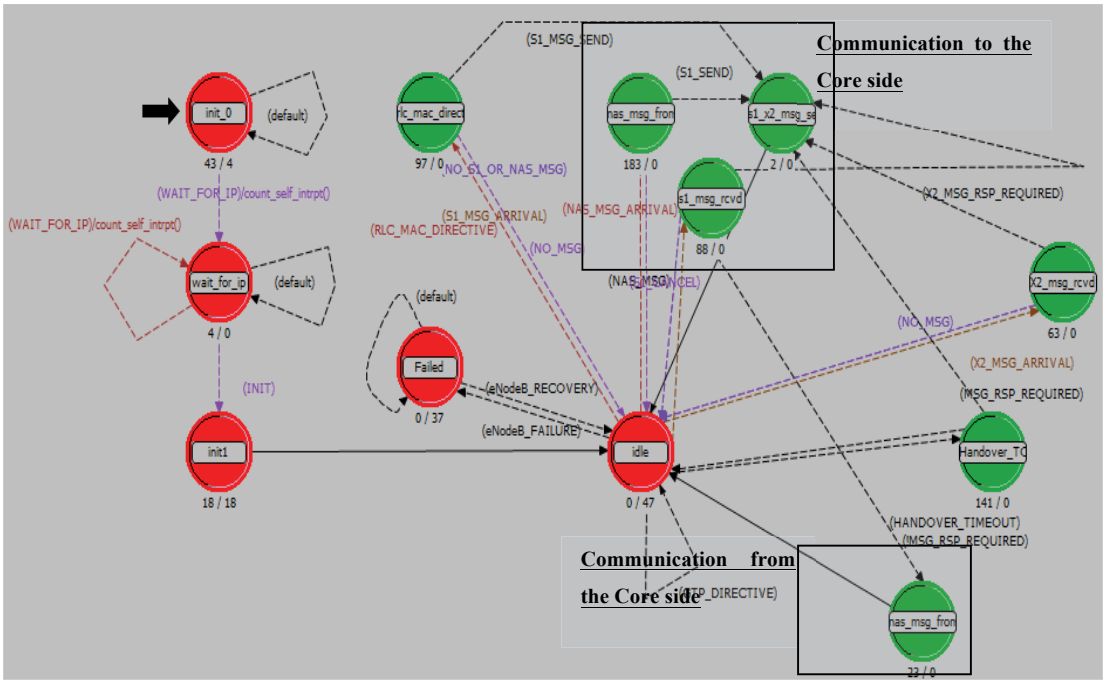


Figure 5.9 lte_s1 Process Model

- Figure 5.9 presents an eNodeB's S1 process model provided in OPNET.
- This process model acts as a translator between the core (EPC) and the radio (E-UTRAN) domains of LTE networks.
- It communicates a UE's Non Access Stratum (NAS) messages to the core side.
- It also translates the core NAS messages for the radio side, such as bearer activation and deactivation messages.

S1_Msg_rcvd State

In Figure 5.9, the S1_msg_rcvd represents a state in which an eNodeB received S1 messages from an EPC. The S1 message may include an EPS bearer setup request, UE context release, handover preparation request, handover cancel request, MME status transfer and bearer release request. When the message is a new bearer setup request, this state commands an access stratum layer to run the functions to establish a dedicated radio bearer. It also commands the GTP layer in an eNodeB to create a tunnel for this bearer towards the EPC in the uplink direction.

5.4.2.2 AS Process Model of an eNodeB

It performs function of an admission control to manage the scarce radio resources. It communicates with S1 process model for this purpose. It keeps record of all bearers admitted in the network.

- It creates uplink and downlink subframes to send and receive the traffic on wireless medium. To accomplish this it
  - Performs scheduling of traffic on radio resources.
  - Manages uplink and downlink HARQ transmissions.
- It receives uplink MAC PDUs (MPDUs) and transmits downlink MPDUs.
- It performs HARQ and RLC retransmission for downlink MPDUs in error.

The process model lte_enb_as is modified to add a congestion control module at an eNodeB. Furthermore, the scheduler employed by the frame generator in the lte_enb_as process model needs to be enhanced as discussed in section 5.2 to provide fair resource allocation in the network.

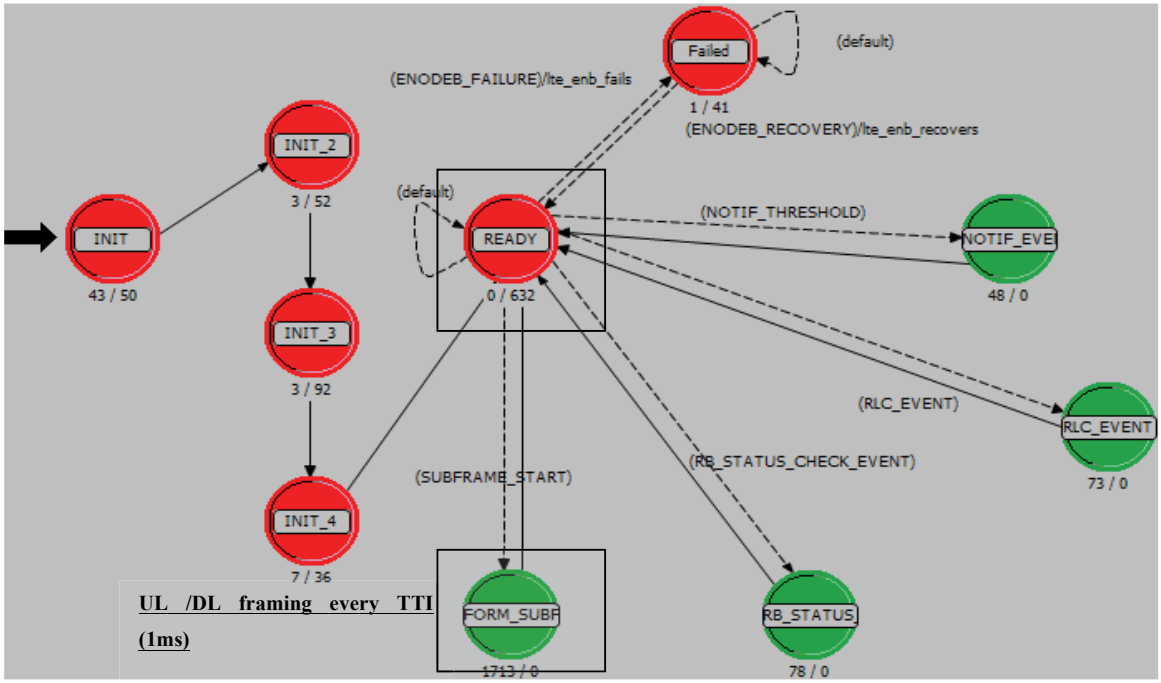


Figure 5.10 lte_enb_as Process Model

Figure 5.10 presents an eNodeB as process model provided in OPNET.

FORM_SUBFRAME State

Figure 5.10 indicates this state. It performs UL and DL framing in every Transmission Time Interval (TTI) - 1 ms. To perform framing, it executes DL and UL schedulers in every TTI. To apply the proposed congestion control scheme we need to change this state. This state is updated so that it first executes the congestion control algorithm, which estimates the fair share of each service type and passes it as a feedback to the scheduler. Later, it executes the scheduler, which allocates resources considering the feedback from the congestion control algorithm.

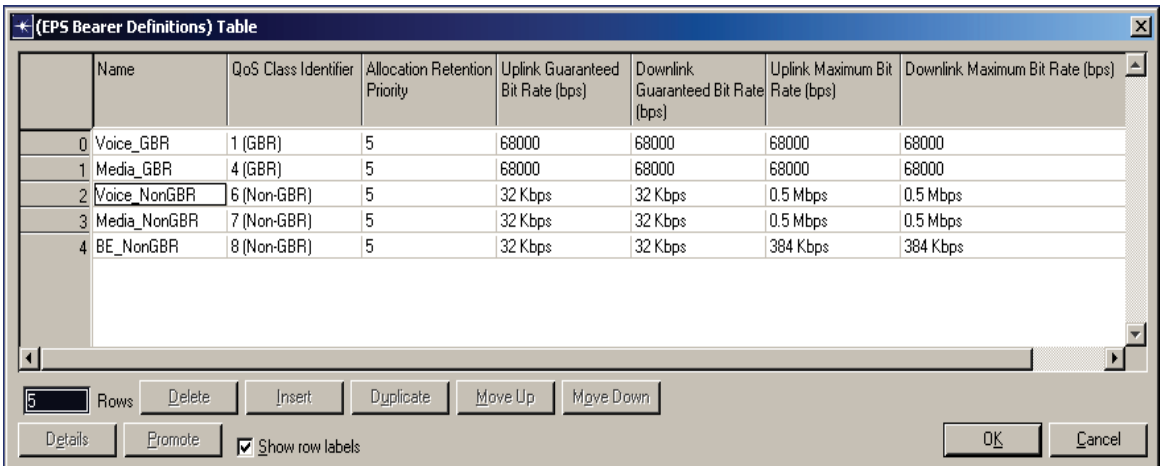
READY State

In this state, it receives packets coming from S1 layer through interrupts streams and tries to decode them. When the message is S1 command to setup a bearer, it calls function to admit a dedicated EPS bearer. To update the admission control algorithm we need to modify function of lte_admit_control_support_radio_bearer_admit of the lte_enb_as process model, which admits an EPS bearer.

5.5 Simulation Setup

In the current simulation setup, LTE UEs are connected to an eNodeB that in turn is connected to an EPC. The EPC is connected to a server through the Internet to reflect an actual deployment of an end-to-end network. The eNodeB is set to operate in FDD mode and uses physical profile of 3 MHz bandwidth. The target operating point is set at 1/16 of the total buffer capacity of 1.5 Mbps. The link capacity between the eNodeB and the EPC is set at 1.3 Mbps to depict a high congestion scenario. The EPS bearer configuration attribute defines five bearers. Each bearer is assigned a Traffic Flow Template (TFT) packet filter, which in this case is Type of Service (ToS) value. Table 4.1 shows each bearer's name, QCI, ARP, GBR and MBR for both UL and DL, respectively.

Table 5.1. EPS bearer Configuration

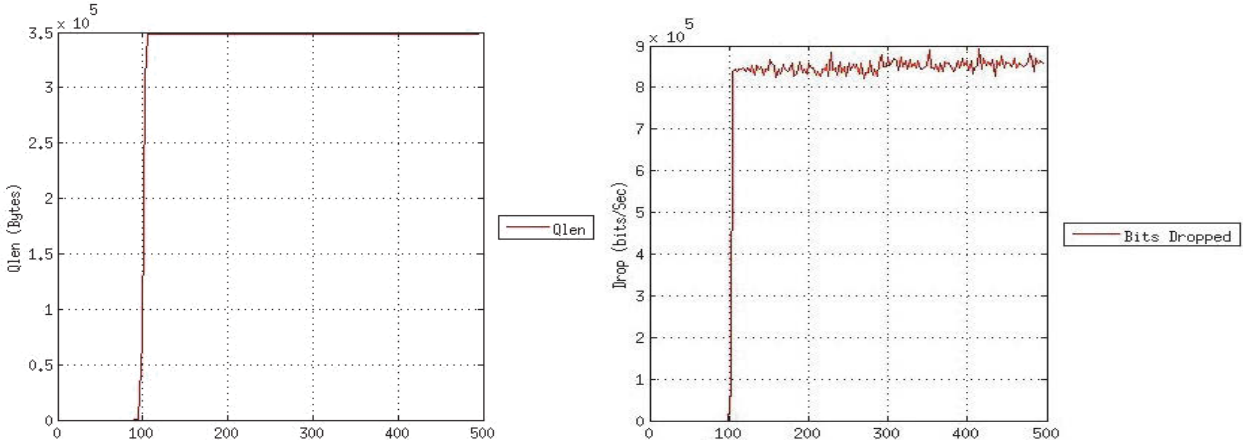


In the current simulation setup, a cell is taken with the GBR and the non-GBR UEs. In the simulation, the maximum MCS (Max _MCS) of all UEs is set at 15. A GBR_UE has two GBR bearers each with GBR of 64 kbps. The GBR UE transmits VoIP G.711 and 64 kbps H.263 video streams using GBR bearers with QCI-1 and QCI-4, respectively. The non-GBR UEs transmit VoIP G.711, 256 kbps H.263 video streams and 64 kbps web traffic using bearers with QCI-6, QCI-7 and QCI-8, respectively. The simulation uses the trace file for 64 kbps and 256 kbps H.263 encoded Jurassic Park movie provided by (Fitzek and Reisslein, 2001). All connections start transmission at around 100 ms of the simulation and stop at end of the simulation. The total simulation time is 500 seconds.

5.6 Simulation Results

This section presents and discusses the results of our proposed congestion control scheme for the LTE networks with the modified round robin algorithm. For comparison a reference scenario is taken without the proposed LTE-FICC and the modified round robin algorithms.

5.6.1 Queue Length (Qlen) and Traffic Dropped

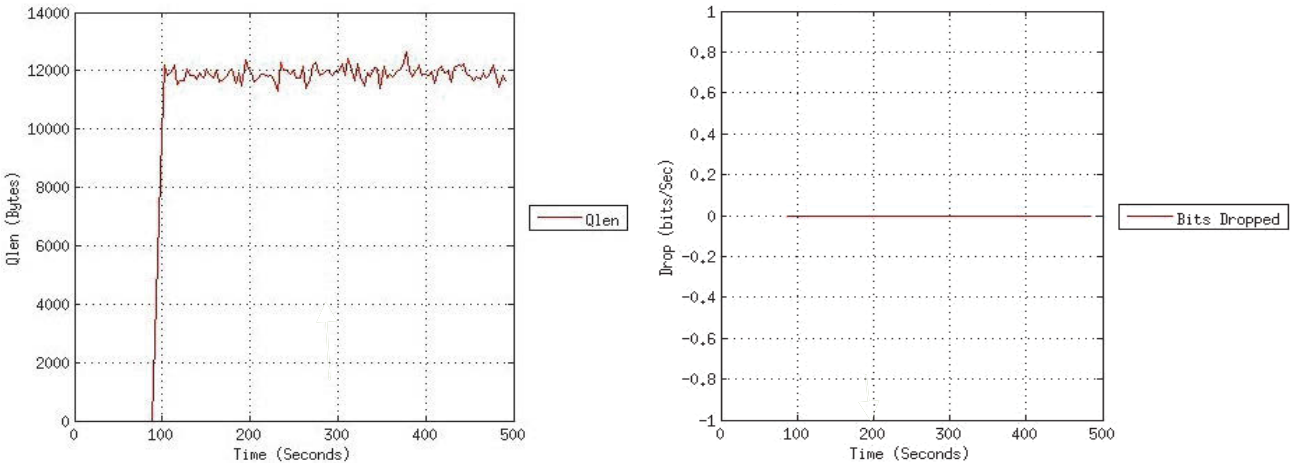


(a) (b)

Figure 5.11 (a) Queue Length (Bytes) (b) traffic Dropped at an eNodeB, without LTE-FICC

Figure 5.11 (a) shows the queue length at an eNodeB output buffer. It confirms that if during the congestion periods, the scheduler allocates resources without taking into account the capacity of an eNodeB's output buffer, a point comes when the buffer overflows and packets-drop starts as shown in Figure 5.11 (b). The buffer drops packets in FIFO order. As a result, the QoS of all connections in the network degrades.

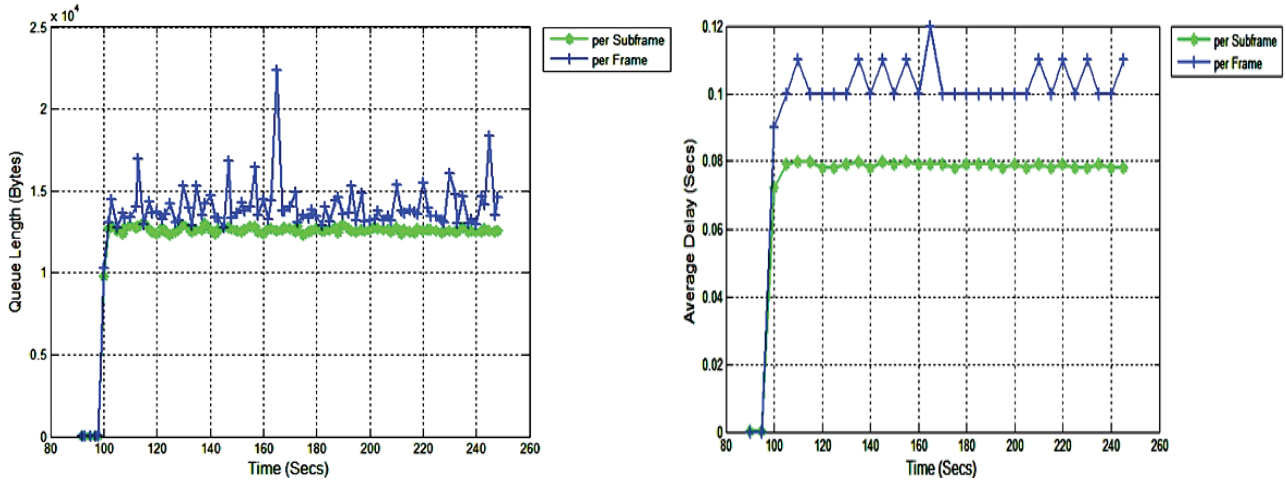
Figure 5.12 (a) shows the queue length at an eNodeB output buffer with the proposed LTE-FICC implemented on an eNodeB. With LTE-FICC, as soon as the queue length reaches the target point, LTE-FICC starts the degradation procedure. It controls the rate offered to connections of different QoS classes and maintains the queue length close to the target operating point. Consequently, there is no loss at an eNodeB output interface as indicated by Figure 5.12 (b).



(a) (b)

Figure 5.12 (a) Queue length (Bytes) (b) Traffic Dropped at an eNodeB, with LTE-FICC

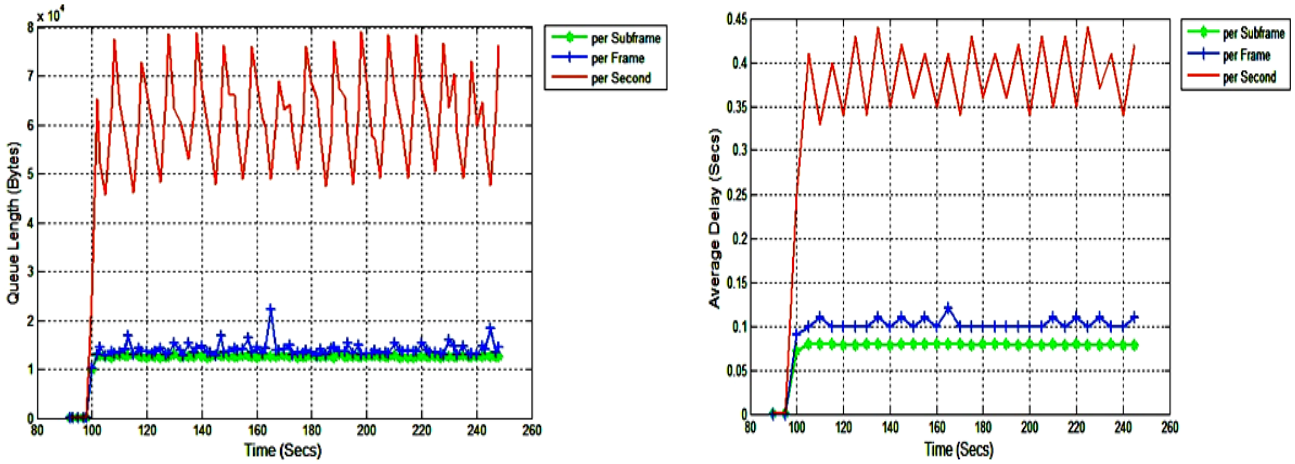
In the current simulation setup, LTE-FICC is executed per subframe before the execution of the scheduler. Simulations are also performed to execute LTE-FICC per frame and per second.



(a) (b)

Figure 5.13 (a) Queue lengths (Bytes) (b) Queuing delays (sec) at an eNodeB, with LTE-FICC executed per subframe and per frame

Figure 5.13 (a) indicates that the execution of LTE-FICC per subframe provides better control on the queue length, and maintains it around the target operating point. It results in reduced and stable delay as shown in Figure 5.13 (b). However, when LTE-FICC is executed per frame, the queue length operates above the target operating point and results in increasing the queuing delay as indicated in Figure 5.13 (b).



(a) (b)

Figure 5.14 (a) Queue lengths (Bytes) (b) Queuing delays at an eNodeB, with LTE-FICC executed per subframe, per frame and per second

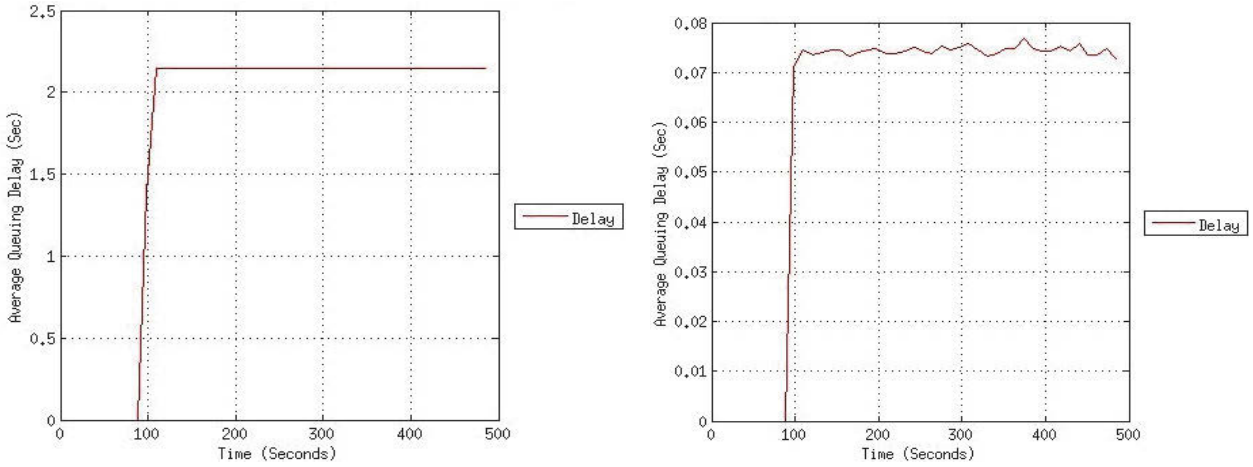
Furthermore, when LTE-FICC is executed per second, the queue length does not maintain around the target. It fluctuates highly close to the buffer capacity as indicated in Figure 5.14 (a). This results in a very high and unstable delay as demonstrated in Figure 5.14 (b). It is due to the fact that the scheduler for the duration of 1 second utilizes the same rate as estimated by LTE-FICC in the first subframe at the start of the second. For the remaining subframes (999 subframes), it schedules connections without any feedback from the LTE-FICC.

The rest of the chapter, therefore, presents results with LTE-FICC executed in every subframe.

5.6.2 Average Queuing Delay

Figure 5.15 (a) and (b) illustrates the average queuing delay at the eNodeB output buffer. Initially, as there is less amount of data in the queue, so the delay is less. As the transmission starts, amount of data in the queue increases and hence the delay increases.

Figure 5.15 (a) clearly indicates that without LTE-FICC the queuing delay is very high. It can be attributed to the fact that when LTE-FICC is not applied, the queue reaches the maximum buffer capacity (Figure 5.11 (a)) and results in very high queuing delay. Figure 5.15 (b) illustrates that with LTE-FICC the queuing delay is very low. It is because LTE-FICC maintains the queue length close the target point (Figure 5.12 (a)).



(a) (b)

Figure 5.15 Queuing Delay (sec) (a) without LTE_FICC (b) with LTE-FICC

5.6.3 Throughput of GBR Bearers

Figure 5.16 shows the throughput of GBR bearers without the application of LTE-FICC.

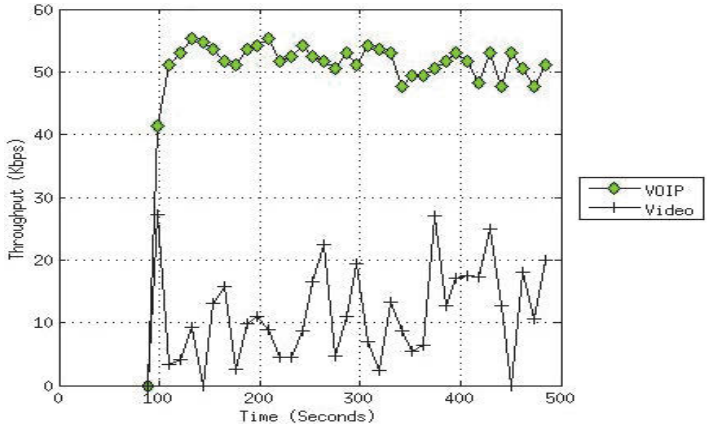


Figure 5.16 Throughput (kbps) of GBR Bearers without LTE-FICC

It clearly illustrates that the GBR bearers are getting less than their GBR value of 64 kbps. This is due to the fact that when the queue length at the buffer reaches the maximum capacity, it starts dropping packets in FIFO order (Figure 5.11 (b)). The video traffic has packets of large sizes, hence the drop of packets significantly affects its throughput as illustrated in Figure 5.16. LTE-FICC maintains the queue length around the target operating point. To control the queue

length, it degrades or upgrades the rate allocated to non-GBR bearers and the rate allocated to GBR bearers above their GBR. In the current simulation, the GBR bearers have the same value of GBR and MBR as shown in Table 5.1. Hence, LTE-FICC does not apply on these GBR bearers. Therefore, flows of GBR CoB get the requested GBR as shown in Figure 5.17.

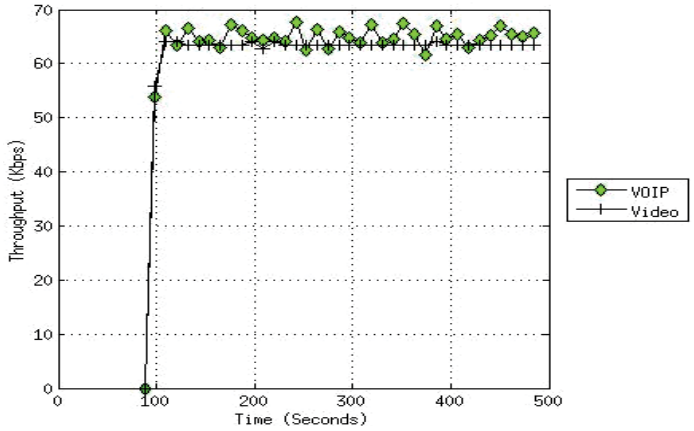


Figure 5.17 Throughput (kbps) of GBR Bearers with LTE-FICC

5.6.4 Fair Resource Allocation

Fair Resource Allocation among the QCIs of non-GBR CoB

Figure 5.18 shows the cumulative throughput of different QCIs of non-GBR CoB when the equal capacity sharing algorithm or the RR algorithm allocates resources.

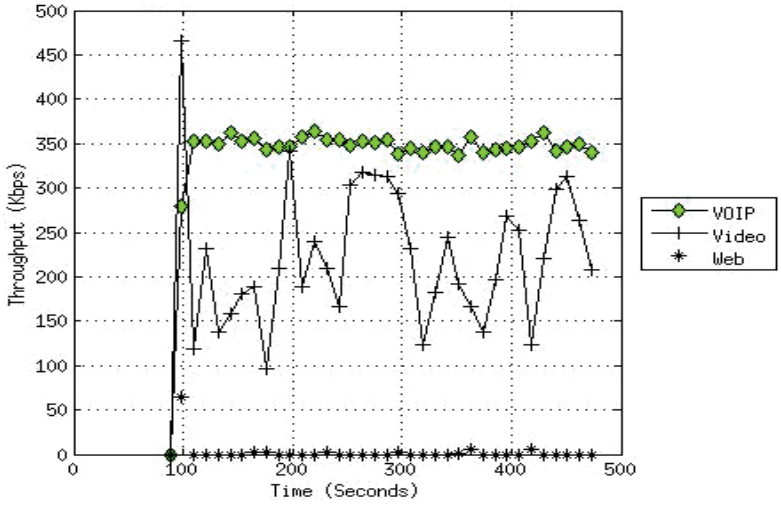


Figure 5.18 Total Throughput (kbps) of non-GBR bearers without LTE-FICC

Figure 5.18 indicates that the throughput of web application with the lowest priority QCI is almost zero. It is due to the fact that the two algorithms always start scheduling with the highest priority QCI and serve it until all queues at that priority level are empty. In this way, it results in unfairness to connections with low priority QCI. In the simulation, queues of high priority voice and video non-GBR bearers always have data to send. Consequently, these schedulers do not allocate resources to low priority web bearers as illustrated in Figure 5.18.

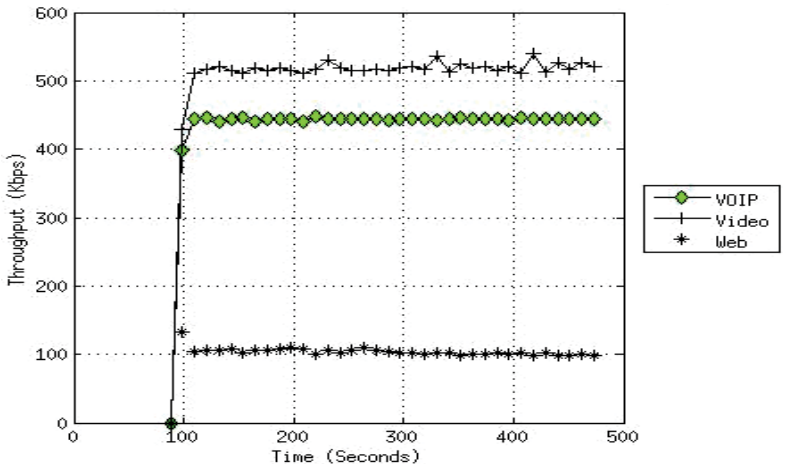


Figure 5.19 Total Throughput (kbps) of non-GBR bearers with LTE-FICC

Figure 5.19 demonstrates that the modified round robin with LTE-FICC ensures that queues at all priority levels within the non-GBR CoB are served. To provide differentiation, the modified RR allocates resources to connections according to the assigned weights of QCIs. Consequently, the throughput of each service is in the order of precedence stated by their respective QCIs. Thus, modified round robin provides fairness among QCIs of non-GBR CoB. In Figure 5.19, the throughput of video traffic with QCI-9 is higher than the throughput of voice traffic with QCI-8. This is because in the simulation, video sources have more traffic to send and thus take additional share of bandwidth when resources are available in the network.

Fair Resource Allocation among the Flows of the same QCI of non-GBR CoB

Figure 5.20 shows the throughput of connections with the equal capacity sharing algorithm provided in OPNET. It clearly indicates that some flows such as VoIP-1 and VoIP-5 are getting

higher throughput around 64 kbps compared to other flows such as VoIP-4 and VoIP-6, which are getting the throughput less than 40 kbps.

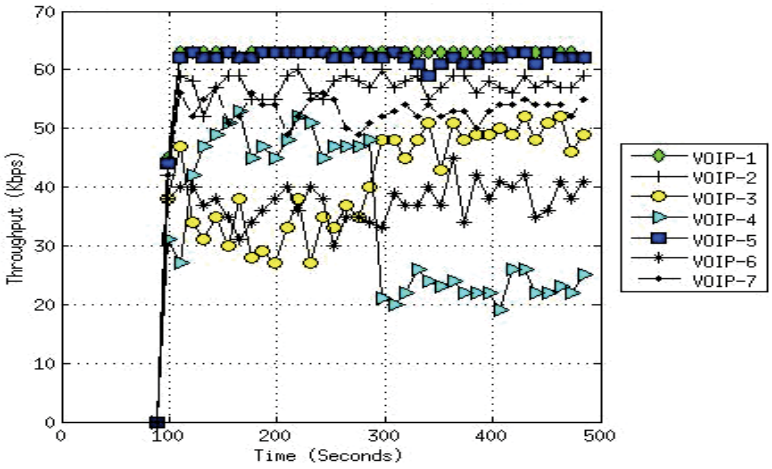


Figure 5.20 Throughput (kbps) of non-GBR flows without LTE-FICC

Thus, when scheduling is performed using the equal capacity sharing algorithm, the network cannot provide fair resource allocation among the flows of the same QCIs as discussed in section 5.2.

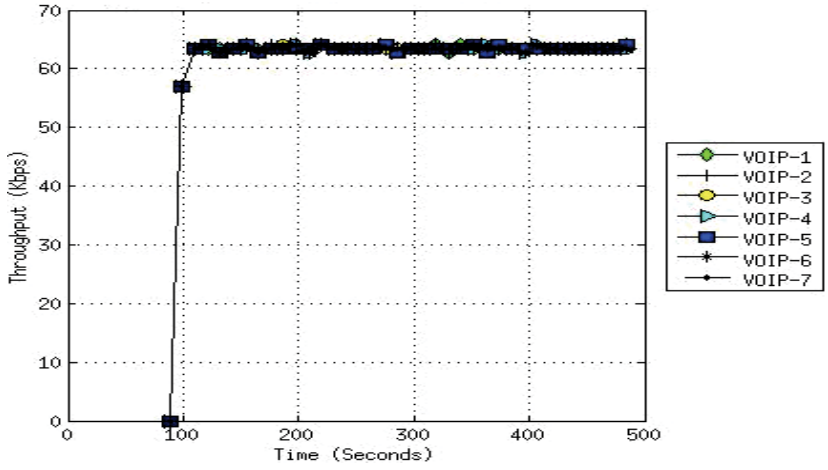


Figure 5.21 Throughput (kbps) of non-GBR flows with LTE-FICC

Figure 5.21 demonstrates that the modified round robin algorithm with LTE-FICC provide fairness within the QCI, as all flows at the same priority level are getting the same throughput and hence the fair amount of network resources.

5.7 Discussion on Results

Consistency

WFICC is implemented in ns-2, which employs the proportional fair scheduler to allocate resource among the connections of all CoSs to gain their maximum rate as discussed in **0**. LTE-FICC is implemented in OPNET, which employs different schedulers to allocate resources to GBR connections to gain their GBR, and above their GBR requirements. Furthermore, it applies separate scheduler to allocate resources to non-GBR connections. All of these schedulers in ns-2 and OPNET do not ensure fair resource allocation. It is because they allocate resources without taking into account the QoS requirements of each QoS class and load at the core network.

The congestion control schemes, WFICC and LTE-FICC, estimate the fair share of resources of each QoS class in accordance to their QoS requirements and the load at the network. They pass this rate as a feedback to the underlying schedulers and enable them to provide fair resource allocation among the connections with the same QoS requirements and also with different priority levels. The PF scheduler in ns-2 scans all queues in a round robin manner. So, it is updated only to consider the feedback from the CC module. However, the schedulers in OPNET, which allocates resources to GBR connections above their GBR requirements and to non-GBR connections, do not ensure fair resource allocation among different QCIs even when LTE-FICC estimates the accurate fair share of each QoS class. It is due to their defined way of operation in the simulator as discussed in section 5.2. In our implementation, therefore, we enhanced these schedulers to ensure provision of the fair resource allocation among the data flows with different priorities.

Extensive simulations demonstrated that LTE-FICC with the modified round robin performed similar to WFICC. It is consistent in obtaining network performance in terms of fair resource allocation, high throughput and high link utilization. It successfully maintains the queue length around the target operating point and results in small deviations in the queue length at an eNodeB's output buffer and hence in the average queuing delay. Hence, the proposed congestion control schemes effectively and consistently provide fair resource allocation in any architecture (WiMAX or LTE) that defines various QoS classes to handle different type of services. They can

operate with any underlying scheduler provided the scheduler considers all QoS classes while allocating resources to connections.

Fair Bandwidth Allocation

LTE-FICC accurately and consistently estimates the fair share of each CoB. To manage the load in the network, LTE-FICC adjusts the rate allocated to GBR bearers above their GBR requirements (GBR_Ad) and to non-GBR bearers. It cannot change the rate, which is allocated to GBR connections to gain their GBR requirements. In the implementation of WFICC in ns-2, the proportional fair scheduler allocates resources directly to gain MSTR of connections. So, separate expected rate for each QoS class is maintained by WFICC. However, in the LTE module of OPNET connections are given resources separately to gain their GBR and above GBR requirements by applying separate schedulers. So, for the sake of simplicity, LTE-FICC maintains the same expected rate for connections of all QoS classes (QCI 1 to 4) to allocate in addition to their GBR requirements (ER_{GBR_Ad}). It also estimates one value of expected rate for non-GBR bearers (ER_{non_GBR}) with different priorities (QCI 6 to 9). Now to differentiate among the connections of different QoS classes, scheduler assigns weight to the expected rate estimated by LTE-FICC in accordance to the priority of each QoS class. Consequently, it simplifies the operation of LTE-FICC, as now it sends only ER_{GBR_Ad} and ER_{non_GBR} as a feedback to the scheduler.

Thus, LTE-FICC with the modified round robin ensures that the packets that occupy buffer represent the fair share of connections with different QCIs within each CoB.

Bounded Queue Length

In times when the network is in non-congested state and the queue length operates below the target point, LTE-FICC encourages the expected rate of each CoB. The overselling feature of LTE-FICC allows the unconstrained connections to take up the resources that cannot be utilized by the constrained connections. This feature allows LTE-FICC to successfully maintain the queue length around the target operating point and ensure the output link is always fully utilized.

Parameter Setting Sensitivity

The basic concept of WFICC and LTE-FICC is the same, although the implementation varies due to the changes in the technology and the implementation environments. Both WFICC and LTE-FICC estimates the expected rate for connections based on their QoS requirements and load at the network. LTE-FICC involves the same parameters as in WFICC including a target operating point, an oversell factor (α) and an exponential average factor (β).

The basic applications of the parameters are also identical in both the schemes. Similar to WFICC, LTE-FICC upgrades or degrades the expected rate depending on whether the queue length varies above or below the target point. Consequently, the target operating point impacts when LTE-FICC starts its operation. Similar to WFICC, in LTE-FICC when the queue operates below the target point, the amount of overselling depends on the value of oversell parameter (α). In order to keep the queue length fluctuating closely around the target operating point, the oversell parameter must be set a little above 1.0. Furthermore, similar to WFICC, in LTE-FICC the exponential average parameter (β) only impacts how fast the value of the $MACR_{CoB}$ reaches the ACR_{CoB}. So the impact of the parameters is the same on both the schemes as discussed in section 4.4.

5.8 Summary

In this chapter, a new CC algorithm for LTE networks is proposed. It is demonstrated to perform extremely well to provide stable throughput and fair resource allocation in the network. Instead of using thresholds to reduce the network congestion, it employs a target operating point. It maintains the network traffic around the target point, hence avoiding congestion and loss at an eNodeB output buffer.

It estimates the fair share of bandwidth for connections with different priority levels. In this way, it enables the schedulers to provide fair resource allocation among the connections based on their QoS constraints and the load in the network. The scheme is simple and easy to implement at an eNodeB of LTE networks.

Chapter 6 Fair Intelligent Admission Control - WiMAX

Chapter 6 introduces a predictive Radio Admission Control (RAC), the Wireless Fair Intelligent Admission Control (WFIAC). WFIAC operates at a BS in Grant per Connection (GPC) mode. It admits or rejects an incoming connection base on the resource availability and the load in the core network. The proposed RAC is based on the bandwidth borrowing. It applies a degradation on the over provisioned connections in order to minimize the blocking probability and to maximize the resource utilization in the network. WFIAC employs a variable size degradation step, which ensures fair resource allocation among the connections at the same priority level. The degradation is applied in such a manner that the proposed RAC provides fairness and service differentiation among the connections at different priority levels. WFIAC operates with a load control module to determine the load in the network. WFIAC along with the load control module ensures that the network operates around the target operating point, guaranteeing QoS to end users in terms of stable throughput and delay.

The structure of this chapter is as follows: Section 6.1 discusses the fundamentals of the proposed WFIAC scheme mainly involving the estimations of the rate to be allocated to a new connection of a specific service type and the variable size degradation step. Section 6.2 comprehensively describes the admission procedures of WFIAC scheme. Section 6.3 presents the simulation setup and Section 6.4 discusses the simulation results with the WFIAC scheme. Section 6.5 provides discussion on the simulation results to determine the effectiveness of the proposed scheme.

6.1 Fair Intelligent Admission Control for WiMAX Networks

The IEEE 802.16-2005 standard (802.16-2005, October 2004) defines specific QoS parameters for each class of service but it does not provide mechanisms to ensure QoS of connections of different Class of Services (CoSs). The Admission Control (AC) mechanism is not specified by the standard and is left as an open research area. The AC plays a crucial role in admitting

connections of different traffic types with different QoS constraints and priorities. An efficient AC ensures QoS provisioning to new and existing connections in the network. IEEE 802.16 is a connection oriented MAC and any Mobile Station (MS) before sending the data establishes a logical connection with the BS by sending a Dynamic Service Addition (DSA) request that contains all the QoS parameters of the incoming connection. The AC module admits/rejects a new connection based on the QoS requirements of incoming connection and the QoS constraints of existing connections.

The authors (Tung et al., 2008) suggested a CAC scheme based on thresholds. A CAC scheme that reserves resources for high priority UGS CoS in busy hours only is proposed by (Antonopoulos and Verikoukis, 2010). To prioritize handoff connections the researchers (Chaudhry and Guha, 2007) proposed to reserve adaptive temporal channel bandwidth instead of fixed guard channels. CAC schemes proposed by (Tung et al., 2008), (Antonopoulos and Verikoukis, 2010) and (Chaudhry and Guha, 2007) accept an incoming connection only if sufficient resources are available in network. To provide service differentiation in fixed WiMAX, a CAC scheme based on degradation is proposed by (Murawwat et al., 2009). CAC schemes based on both bandwidth reservation and degradation are proposed by (Wang et al., 2005), (Hou et al., 2006), (Jiang and Tsai, 2006) and (Luo et al., 2009). To prioritize handoff connections in mobile WiMAX, a degradation based strategy is proposed by (Suresh. et al., 2008) and (Wang et al., 2007). A CAC scheme that uses a combination of bandwidth reservation and a degradation based strategy is proposed by (Chandra and Sahoo, 2007).

In literature all CAC schemes that are based on degradation strategy to admit a new/handoff connection use a fixed size degradation step. The same size degradation step is applied to connections of all CoSs without taking into account priority of different CoSs. The rationale behind the selection of the degradation step size is also not discussed.

In this section, we describe the proposed WFIAC. For simplicity we assume that connections of the same CoS have the same QoS requirements including MRTR, MSTR, delay and jitter. We also assume that all users in a cell use the same modulation and coding scheme. The terminology used in this chapter is given below.

NOTATIONS

C : Total amount of bandwidth available at the BS for uplink in terms of slots.

C_{used} : Total number of slots in use.

S_{CAP} : Slot Capacity in bytes.

C_{REQ} : Slots requested by an incoming connection. It is equal to $MSTR_{UGS}/S_{CAP}$ for UGS and ER_{CoS}/S_{CAP} for rtPS, nrtPS and BE CoSs.

C_{rem} : It is the gap between slots requested by an incoming connection and slots available in the available resource pool. C_{REQ} - (C – C_{used}).

C_0 : Target total number of slots in use after WFIAC completes degradation (C_{used}-C_{rem}).

B_{UGS}: MSTR of UGS.

B_{rtPS}^{min}, B_{nrtPS}^{min}: MRTR of rtPS and nrtPS, respectively.

B_{rtPS}^{max}, B_{nrtPS}^{max}: MSTR of rtPS and nrtPS, respectively.

ER_{CoS}, CC_ER_{CoS}: Expected rate of each CoS maintained by WFIAC and WFICC, respectively.

Qlen : Current Queue length at a base station output buffer.

Q_0 : Target Queue length.

f(Q) : function of Queue used in WFICC.

f(S): function of slots used in WFIAC.

N_{UGS} , N_{rtPS} , N_{nrtPS} : Number of connections already admitted for UGS, rtPS and nrtPS CoSs, respectively.

Pr_{REQ} : Priority of the requesting connection's CoS.

Pr_i : Priority of ith CoS

In our proposed WFIAC, when a new connection request arrives with the resources requested, C_{REQ}, WFIAC first verifies the resource availability. To estimate the available resources in the network, WFIAC obtains the expected rate of each CoS maintained by WFICC (Chapter 4), hereafter referred to as CC_ER_{CoS}. WFIAC updates the expected rate for each CoS (ER_{CoS}) maintained by it as follows.

$$ER_{CoS} = \min(ER_{CoS}, CC\_ER_{CoS}) \quad 6.1$$

In Eq. 6.1, CoS represents a class of service of the WiMAX networks, including rtPS, nrtPS and BE CoSs. The ER_{CoS} refers to the expected rate of each CoS maintained by WFIAC. The CC_ER_{CoS} indicates the expected rate provided by WFICC to a scheduler with a view to keep the network operation close to a target operating point. It is estimated based on the load in the network. The aim of Eq. 6.1 is to determine the rate that the scheduler actually allocates to the connections of each CoS and utilize it to estimate the total slots used (C_{used}) in the network as in Eq. 6.2. Furthermore, WFIAC assigns an incoming connection of a particular CoS the same rate as used by the existing connections of the respective CoS. The current resource utilization in network in terms of slots used (C_{used}) is calculated as follows.

$$C_{used}=ceil(ER_{UGS}/S_{CAP})*N_{UGS}+ceil(ER_{rtPS}/S_{CAP})*N_{rtPS}+ceil(ER_{nrtPS}/S_{CAP})*N_{nrtPS}+ceil(ER_{BE}/S_{CAP})*N_{BE} \quad 6.2$$

In Eq. 6.2, the ER_{CoS} (determined in Eq. 6.1) is divided it by the slot capacity (S_{CAP}) to obtain the number of slots used by a connection of a particular CoS. Equation 6.2 estimates the slots used (C_{used}) as the sum of the total slots used by all existing connections of each CoS (N_{CoS}). When WFIAC estimates that the network has sufficient available resources ($C-C_{used}$), it updates the queue length based on the data rate assigned to an incoming connection. Otherwise, it applies a step-wise degradation on all connections of lower priority CoSs to obtain the remaining slots (C_{rem}).

The step size of the proposed WFIAC employs a function of slots, which is defined as follows.

$$f(S) = 1 - \frac{C - C_{used}}{C_0} \quad 6.3$$

In Eq. 6.3, C_0 is the target number of used slots after the degradation process completes (C_{used} - C_{rem}). The step size is estimated as follows.

$$\S = f(S) * S_{CAP} \quad 6.4$$

In Eq. 6.4, § indicates the step size. In WFIAC, a connection at a certain priority level is admitted with the same expected rate as used by the existing connections of the same CoS. So, to keep the degradation procedure simple, WFIAC applies the function of slots to a slot capacity.

As a result, WFIAC degrades the rate allocated to connections of a CoS equal to a fraction of a slot at a time. The proposed step size maintains priority of connections of different CoSs. Since, when the degradation is applied on connections of lower priority CoSs, the number of used slots C_{used} decreases, the degradation step size decreases. This results in less amount of degradation applied on connections of higher priority CoSs. Consequently, it provides differentiation and fair resource allocation among the different priority CoSs.

At Wimax Base Station Admission Control (AC) receives a DSA Request.

a. Obtain ER for each CoS from WFICC namely CC_ER_{rtPS}, CC_ER_{nrtPS} and CC_ER_{BE}.
b. Update ER of each CoS (ER_{CoS}) using Eq. 6.1.
c. Calculate C_{used} using Eq. 6.2.
d. $C_{REQ} = ER_{CoS}/S_{CAP}$ // Slots Requested

1. IF $C_{REQ} < (C\text{-}C_{used})$ // Found Enough Resources

   Go to step 2 to perform load estimation.

   Else

   $C_{rem} = C_{REQ}\text{-}C$

   Step-wise degrade all connections of lower priority CoSs.

   IF $C_{REQ} < (C\text{-}C_{used})$ // Found enough Slots after degradation

   Go to step 2 to perform the load estimation.

   Else

   Reject Connection Request.

   End IF

   End IF

2. IF $\sum_{i=0}^{j}\sum_{n=0}^{N_{CoS}} ER_{CoS} - MRTR_{CoS} > MRTR_{REQ\_CoS}$ // 'i' indicates CoS with priority lower than the priority of requesting connection's CoS.

   Accept Connection

   Else

   Reject Connection

   End If

Figure 6.1 Algorithm of WFIAC

In case when the requested resources are not obtained after the degradation process completion, the requested connection is rejected as indicated in the algorithm of WFIAC in Figure 6.1. Otherwise, WFIAC updates the current queue length and the expected rate of each CoS based on the degradation applied. It then estimates whether WFICC can manage the load of an incoming connection as indicated in step 2 of Figure 6.1. When WFIAC estimates that WFICC cannot manage the load of an incoming connection, it rejects a connection. Otherwise, it accepts the new connection request.

6.2 Description of WFIAC

To reduce the blocking probability and to provide service differentiation in the network, WFIAC applies the degradation procedure as indicated in algorithm of WFIAC in Figure 6.1.

Degradation Procedure

```
Sort connections according to priority.
i= Lowest priority CoS
While C_REQ > (C-C_used)
  Estimate f(Slots) given in Eq. 6.3
  Estimate Step size using Eq. 6.4
  IF  ER_i= ER_i - § > MRTR_i
        ER_i= ER_i - §
  Else
      i=i+1 // Move to next lowest priority CoS
      IF Pr_i > Pr_REQ   //   If priority of CoS 'i' (Pr_i) is greater than the priority of requesting connection's CoS (Pr_REQ), stop degradation.
         Break
      End If
  END IF
  Update C_used using Eq. 6.2.
End While
```

Figure 6.2 Degradation Procedure of WFIAC

Figure 6.2 provides detailed steps involved in the degradation procedure of WFIAC. The degradation is applied to obtain the gap between the resources requested by an incoming connection and the available resources in the network (C_{REQ} - C). The degradation procedure attempts to obtain resources from CoSs with lower priority compared to the priority of the requesting connection.

Figure 6.2 indicates that the degradation procedure starts with the lowest priority CoS. When its expected rate reaches the minimum rate, it moves to the next lower priority CoS. The degradation is applied to the next lower priority CoS only when its priority is less than the priority of the incoming connection. Otherwise, the degradation procedure stops. The degradation procedure, based on the degradation applied, continues updating the estimate of the used slots (C_{used}) using Eq. 6.2. The degradation procedure terminates when it manages to obtain the remaining slots (C_{rem}) from the existing lower priority connections. WFIAC deals with connection of each CoS in a different manner as discussed below.

6.2.1 UGS connection

When a BS receives a Dynamic Service Addition (DSA) request for a UGS connection with the bandwidth request of B_{UGS}, the admission control module checks the availability of resources. WFIAC sets the connection status *to be accepted*, if the system has sufficient available resources. Otherwise, it applies a step-wise degradation on the over provisioned connections of lower priority CoSs for the remaining slots (C_{rem}). For UGS connections, it applies the step-wise degradation first on the lowest priority CoS, the BE. In case when resources are still not sufficient after degrading the rate of BE connections to their minimum, it step wise degrades connections of the next lower priority CoS, the nrtPS, until either enough resources are obtained, or their expected rate reduces to $MRTR_{nrtPS}$. If still enough resources cannot be obtained, WFIAC step-wise degrades connections of the next lower priority CoS, the rtPS, until either enough resources are obtained, or their expected rate reduces to $MRTR_{rtPS.}$

WFIAC rejects the connection request, if adequate resources are not obtained after the degradation process completes. Otherwise, using the updated reduced expected rate of each CoS,

it verifies for the load in network. For a UGS connection, the network is in load if the following statement is true.

$$\sum_{n=0}^{N_{BE}} ER_{BE} - 0 + \sum_{n=0}^{N_{nrtPS}} ER_{nrtPS} - MRTR_{nrtPS} + \sum_{n=0}^{N_{rtPS}} ER_{rtPS} - MRTR_{rtPS} < B_{\text{UGS}} \quad 6.5$$

WFIAC, using Eq. 6.5, determines whether the sum of the rate above the minimum rate of all connections of lower priority CoSs (rtPS, nrtPS, BE) is less than the bandwidth requested by an incoming UGS connection. If true, WFIAC rejects the connection. This is because, Eq. 6.5 indicates that WFICC will not be able to manage the load introduced by the new connection even after degrading the rate of lower priority connections to their minimum. So, WFIAC does not accept the connection to maintain the QoS of existing connections.

6.2.2 rtPS connection

When a BS receives a DSA request for an rtPS connection with the bandwidth request of B_{rtPS}^{max} and B_{rtPS}^{min}, WFIAC assigns an incoming connection the rate equal to ER_{rtPS} as determined using Eq. 6.1. The aim is to assign an incoming rtPS connection the same rate as used by the existing connections of rtPS CoS. WFIAC sets the connection status *to be admitted*, if the system has enough available resources. Otherwise, it applies a step-wise degradation on the over provisioned connections of lower priority CoSs for the remaining slots (C_{rem}). For rtPS connection, WFIAC applies the step-wise degradation first on the lowest priority BE CoS. If after degrading BE connections resources are still not adequate, it step wise reduces the rate of the next lower priority nrtPS CoS until either enough resources are obtained, or their expected rate reduces to $MRTR_{nrtPS}$. WFIAC rejects the connection request if enough resources are not obtained after the degradation procedure completes. Otherwise, WFIAC using the new expected rate of each CoS, updated after the degradation, updates the queue length and check*s* for the load in network. For an rtPS connection, the network is in load if following statement is true.

$$\sum_{n=0}^{N_{BE}} ER_{BE} - 0 + \sum_{n=0}^{N_{nrtPS}} ER_{nrtPS} - MRTR_{nrtPS} < MRTR\text{rtPS} \quad 6.6$$

Equation 6.6 describes the situation when the total rate above the minimum rate of all connections of lower priority CoSs is less than the minimum rate of the requesting rtPS connection. In this situation, WFIAC rejects the connection. It is due to the fact that Eq. 6.6 indicates that WFICC will not be able to manage the load introduced by the requesting rtPS connection even after degrading connections of lower priority nrtPS and BE CoSs to their minimum rate. So, WFIAC does not accept the connection to maintain the QoS of existing connections. To manage the load at the core, WFICC can degrade the rate of an rtPS connection above its minimum reserve rate ($MRTR_{rtPS}$). Therefore, WFIAC validates only the load for $MRTR_{rtPS}$ ( ), which WFICC is not allowed to degrade.

6.2.3 nrtPS connection

When a BS receives a DSA request for an nrtPS connection with the bandwidth request of $B_{nrtPS}^{max}$and B_{nrtPS}^{min}, WFIAC assigns an incoming connection the rate of ER_{nrtPS} as calculated in Eq. 6.1. In case when sufficient resources are available in the network, WFIAC sets the connection status *to be accepted*. Otherwise, it executes the step-wise degradation on connections of lower priority CoSs for the remaining slots (C_{rem}). For nrtPS connections, WFIAC applies step wise degradation only on the lowest priority BE CoS. It rejects the connection request, if enough resources are not obtained after the degradation process completes. Otherwise, WFIAC using the updated expected rate of BE CoS checks for the load in network. For an nrtPS connection, the network is in a congested state if following statement is true.

$$\sum_{n=0}^{N_{BE}} ER_{BE} - 0 < MRTR_{nrtPS} \qquad 6.7$$

Equation 6.7 indicates the situation when the sum of the rate above the minimum rate of all connections of the lowest priority BE CoS is less than the MRTR requested by an incoming nrtPS connection. In this situation, WFIAC rejects the connection. Equation 6.7 presents the scenario when WFICC will not be able to manage the load introduced by an incoming nrtPS connection even after degrading the lowest priority BE CoS connections to their minimum rate. So, WFIAC does not accept the connection to maintain the QoS of existing connections.

6.2.4 BE connection

When a BS receives a DSA request for a BE connection with the bandwidth request of B_{BE}, WFIAC always admits the connection. It is because a BE connection does not require the QoS guarantee. After allocating resources to connections of high priority CoS, remaining free slots (C- C_{used}) are equally shared among the BE connections. So, the expected rate of BE CoS (ER_{BE}) is calculated as follows.

$$ER_{BE} = \frac{C - C_{used}}{N_{BE}} \quad 6.8$$

In Figure 6.1, WFIAC algorithm does not discuss any step to forward the new expected rate to WFICC in scenarios when degradation is applied. It is because WFICC uses the $MACR_{CoS}$ to determine the expected rate of a CoS. The $MACR_{CoS}$ maintains the average rate allocated to connections of each CoS. So, when new connection is admitted through degradation, existing connections achieves lower rate and hence $MACR_{CoS}$ reduces. Consequently, WFICC automatically adjusts the expected rate of each CoS.

Furthermore, in Figure 6.1, WFIAC algorithm does not discuss any step to increase the expected rates of CoSs when a connection is rejected. It is because, WFIAC operates on the expected rate maintained by it. In every frame, WFICC estimates the expected rate of each CoS and passes it to the scheduler. Consequently, in case when a connection is not admitted, resource allocation in the network is not affected by the estimations performed by WFIAC.

6.3 Simulation Setup

The overall goal of the simulation is to investigate the effectiveness of the proposed WFIAC scheme in terms of blocking probability of incoming connections, and the guaranteed QoS of existing connections in terms of throughput and delay. Simulations have been performed in ns-2 (ns2, 2010b) using the WiMAX module for OFDMA by (WiMAX Forum and NIST, 2011) namely ns2-wimax-awg.

In the current simulation setup, WiMAX SSs are connected to a WiMAX BS in IEEE 802.16 PMP mode. The BS is connected to a sink node to reflect the actual deployment of WiMAX

network. The link capacity between the sink node and the base station is initially set at 1.4 Mbps. The target queue length is defined at 1/4 of the total buffer capacity of 1.2 Mbps. In the simulations, 5 ms frame duration is assumed. The ratio of downlink to uplink is set at 7:3. The MCS, QPSK-3/4, with a slot capacity of 9 bytes is assumed for all nodes.

The UGS service flows are dynamically added to handle the VoIP traffic. For the video traffic, rtPS flows are added and nrtPS flows are created for the FTP traffic.

The arrival rate of rtPS and nrtPS follows a Poisson distribution with an average rate of 1 connection per second. The arrival rate of UGS is a Poisson distribution with an average rate that is increased from 1 connection per second to 7 connections per second. The data rate used by the connections of each CoS is as follows.

Table 6.1. QoS Parameters of each Class of Service

Service	**Rate**
B_{UGS}	64 kbps
B_{rtPS}^{max} , B_{nrtPS}^{max}	256 kbps
B_{rtPS}^{min} , B_{nrtPS}^{min}	88 kbps

6.4 Simulation Results

This section presents and discusses the results of our simulations to show the effectiveness of the proposed WFIAC. First, we show the comparison of WFIAC in terms of the connection blocking probability of UGS and non-UGS (rtPS and nrtPS) connections with a Strict Admission Control (SAC) scheme. The SAC scheme admits an incoming connection only if enough resources are available in the network.

Later, we present the investigation of the effect of the load estimation performed by WFIAC on the QoS of existing connections in terms of throughput and delay. The BE CoS does not have

any MRTR or delay requirements, therefore, results for BE CoS are not discussed in the following sections.

6.4.1 Blocking Probability (BP)

The blocking probability of connections is calculated as follows.

$$BP = \frac{Total\ Number\ of\ Connections\ Rejected}{Total\ Number\ of\ Connection\ Requests} \quad 6.9$$

Figure 6.3 and Figure 6.4 show the BP of UGS and non-UGS connections for the networks in congested (CON) and non-congested (NCON) states, respectively. To simulate light congestion (LCON) and high congestion (HCON) in the core network, the link capacity between the BS and the sink node is reduced from 1.4 Mbps to 1.1 Mbps and 0.9 Mbps, respectively. The load estimation is enabled in WFIAC to perform the simulations in order to estimate the BP of connections.

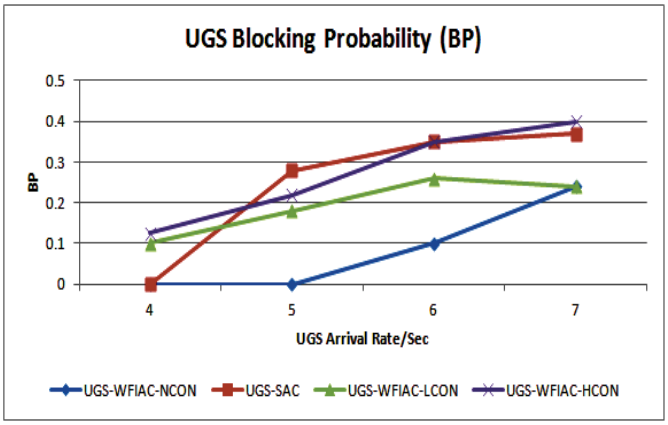


Figure 6.3 BP of UGS connections

Figure 6.3 shows the BP of UGS CoS starting with the arrival rate of 4 connections per second as at this rate the BP for UGS CoS is already zero. The SAC scheme admits connection only if enough resources are available to allocate to an incoming connection. As a result, the BP of UGS connections with the SAC scheme is the highest. Furthermore, it does not consider the load in the network. Consequently, the BP with the SAC scheme is the same for congested and non-congested scenarios as illustrated in Figure 6.3.

Figure 6.3 clearly indicates that when the network is in non-congested state (NCON), the BP of UGS is the lowest with WFIAC scheme. This is because when the connections arrival rate increases and resources become scarce in the network, WFIAC applies a bandwidth degradation procedure on connections of lower priority to admit the highest priority UGS connections. When the core network is lightly congested and the queue length grows above the target operating point, WFIAC admits less number of connections to maintain the QoS of existing connections. Therefore, the BP of UGS increases with WFIAC scheme in light congestion (WFIAC-LCON) but it is still less than the BP of UGS with the SAC scheme as illustrated in Figure 6.3.

Figure 6.3 also demonstrates the BP of UGS in scenario when the network is highly congested. When the arrival rate of UGS connections increases, the BP of connections with WFIAC scheme in heavy congestion (WFIAC-HCON) increases and exceeds the BP of UGS connections with SAC scheme (Figure 6.3). It is because when the core is heavily loaded, WFIAC rejects more incoming connections to maintain the queue length operating around the target point. Hence, it trades off the BP of UGS with the QoS of already admitted connections in the network.

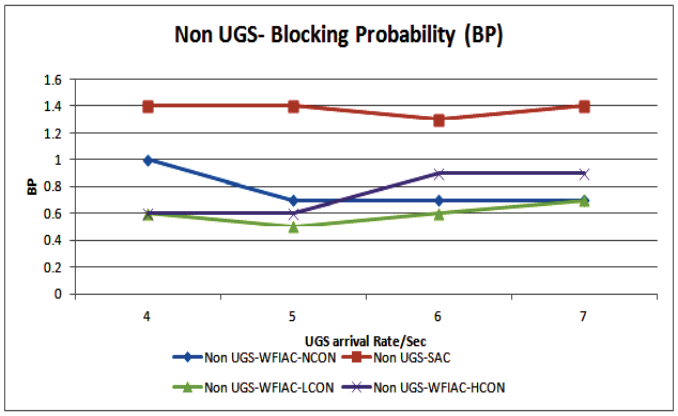


Figure 6.4 BP of non-UGS connections

The combined BP of non-UGS (rtPS and nrtPS) connections is shown in Figure 6.4. It clearly illustrates that similar to UGS CoS, the BP of non-UGS CoSs is the highest with the SAC scheme. In times of resource scarcity, WFIAC degrades connections of nrtPS CoS to accommodate high priority rtPS connections. As a result, with WFIAC the BP of non-UGS in non-congested (NCON) scenario is lower as compared to the SAC scheme.

Figure 6.4 shows in the non-congested scenario (NCON) when arrival rate of UGS increases, the BP of non-UGS connections decreases. This is because when UGS connections arrival rate increases, WFIAC degrades the expected rate of non-UGS connections (both nrtPS and rtPS) to accommodate more UGS connections. Therefore, WFIAC can admit new incoming connections of non-UGS CoSs with the new reduced expected rate even when limited resources are available in the network.

Figure 6.4 illustrates that when the network is lightly congested (LCON), BP of non-UGS reduces even more. It is because when the network is in load state, WFICC in order to control the load degrades the rate allocated to non-UGS connections. Consequently, the expected rate of respective non-UGS CoS (rtPS and nrtPS) maintained by WFICC reduces even when UGS arrival rate is low. WFIAC assigns an incoming connection the minimum of the rate maintained by it (ER_{CoS}) and the rate maintained by WFICC (CC_ER_{CoS}) using Eq. 6.1. Hence, more connections of non-UGS CoSs are admitted with relatively lower resource demand.

Figure 6.4 also demonstrates that when the core network is highly congested (HCON) and the arrival rate of UGS increases, the BP of non-UGS connections with WFIAC scheme increases. It exceeds the BP of non-UGS connections with WFIAC in non-congested (WFIAC-NCON) scenario. It is because when the core network is heavily loaded, WFIAC rejects more incoming connections even with relatively reduced expected rate to maintain the queue length close to the target operating point. Consequently, WFIAC trades off the BP of non-UGS CoSs to maintain The QoS of existing connections in terms of stable throughput and minimum delay.

6.4.2 Effect of Load Estimation on QoS

To show the effect of the load estimation on the QoS of existing connections, the simulations have been performed with a link capacity of 0.9 Mbps. The arrival rate of rtPS and nrtPS is set at 1 connection per second. For UGS CoS, the arrival rate is set at 7 connections per second.

a. WFIAC without Load Estimation

Figure 6.5 shows the queue length at a BS output buffer when WFIAC admits connections only on the base of resource availability.

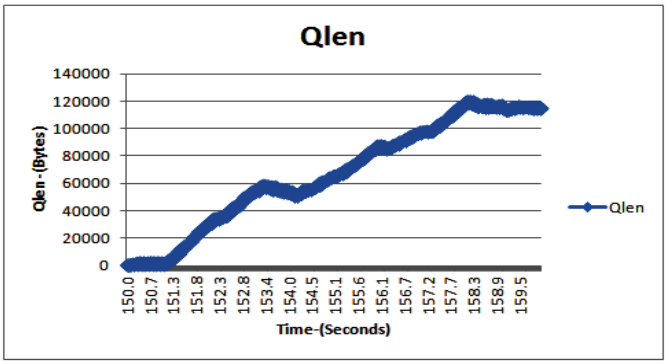


Figure 6.5 Queue length (Bytes) without load estimation

When the core network is congested, and the base station cannot transmit traffic at the rate at which it receives from SSs, the BS's output buffer gets congested indicated by the increasing queue length. When WFIAC further admits connections in the network without taking into account the current load at the output buffer, WFICC will not able to manage the load of the new admitted connections. This is because WFICC can reduce the rate allocated to any connection only to its MRTR. Whereas, when WFIAC admits more and more high priority UGS connections by degrading low priority rtPS and nrtPS connections to their MRTR, WFICC is left with no rate control in the network. Consequently, the queue length at the base station output buffer reaches the maximum buffer limit.

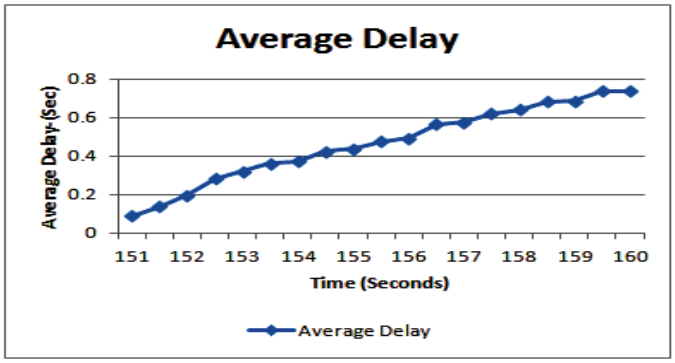


Figure 6.6 Average Delay (sec) without load estimation

Figure 6.6 shows the effect of the increasing queue length on the average delay of the network traffic. It demonstrates that when the core network is congested and WFIAC admits new

connections without considering the load, the output buffer at the base station overflows (Figure 6.5). It results in an increase in the packet average delay for all network connections.

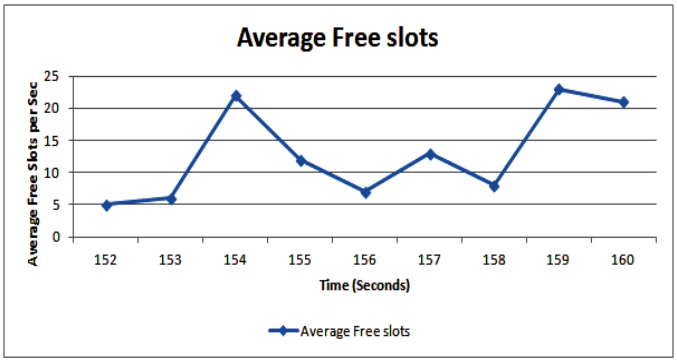


Figure 6.7 Average Free Slots without load estimation

Figure 6.7 shows the average number of free slots in the network. In the current scenario, when a new request arrives, WFIAC admits it if enough resources are available or obtained after degrading connections of lower priority CoSs. Consequently, overall the average number of free slots is low except few peaks, which exist due to the departure of connections.

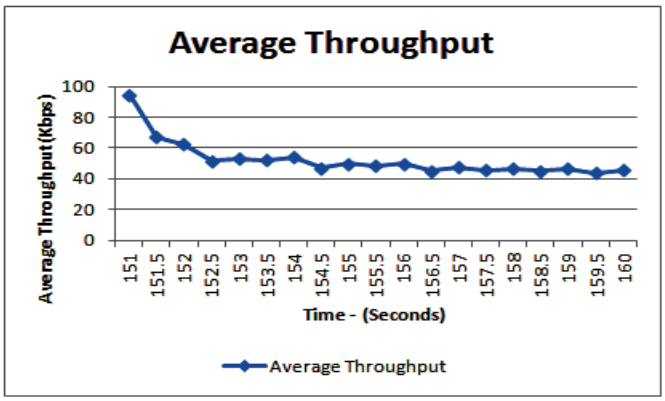


Figure 6.8 Average Throughput (kbps) without load estimation

Figure 6.8 illustrates the effect of the increasing queue length on the average throughput of all connections in the network. Without the load estimation, the queue length at the base station output buffer and hence the average delay of all connections become very high. It results in decreasing average throughput of the network even though maximum resources (in terms of slots) are utilized in the uplink to send user's data (Figure 6.7).

So, if WFIAC admits connections on the basis of the resource availability, without taking into account the network load status, the QoS of existing connections degrade in terms of delay and throughput (Figure 6.6 and Figure 6.8).

b. WFIAC with Load Estimation

The results below present the scenario when WFIAC admits or rejects connections based on both the resource availability and the estimation of load in the core network.

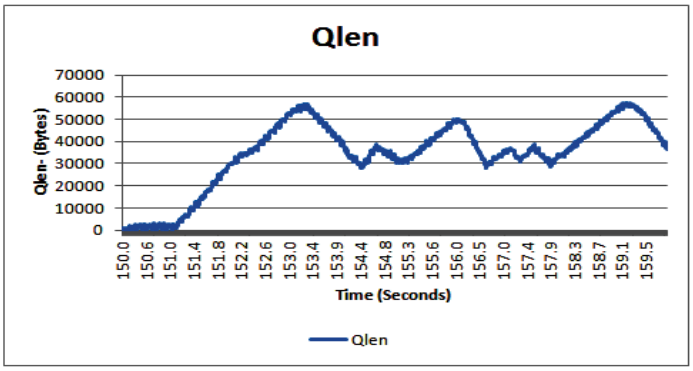


Figure 6.9 Queue Length (Bytes) with load estimation

Figure 6.9 clearly illustrates that with the load estimation the queue length at the base station operates around the target operating point. It is due to the fact that in this scenario, WFIAC admits connections only if it estimates that WFICC will be able to manage the load introduced by an incoming connection. Otherwise, it rejects a new connection request.

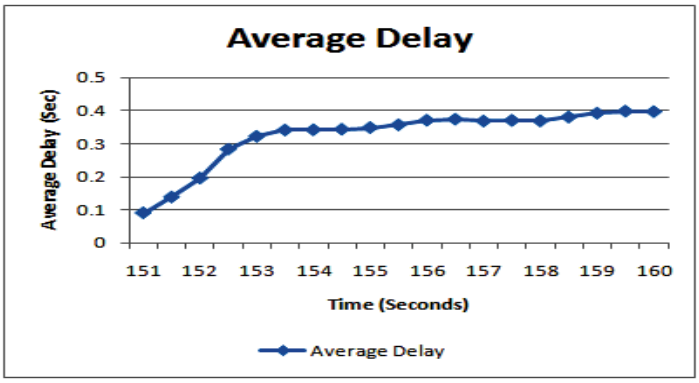


Figure 6.10 Average Delay (sec) with load estimation

Figure 6.10 shows the average delay of the network traffic when WFIAC admits connections with the load estimation. It illustrates that when the traffic in the network increases, the average delay at the base station output buffer increases. It shows that when the queue length increases beyond the target operating point, the average delay of connections become high but as soon as the queue length reduces and remains around the target operating point, the average delay becomes stable. Hence, WFIAC by admitting connections with the load estimation maintains the queue length around the target operating point (Figure 6.9). It results in stable and reduced delay in the network compared to the situation when WFIAC admits connections without the load estimation.

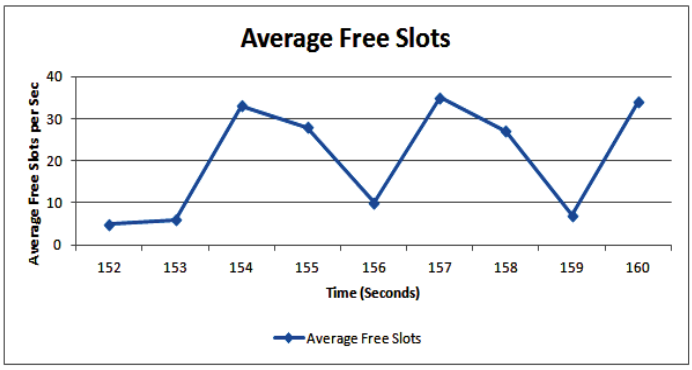


Figure 6.11 Average Free Slots with load estimation

Figure 6.11 shows the average number of free slots in the network. In the current scenario, when an incoming connection request arrives and resources are available in the network, WFIAC admits connection if it estimates that the connection will not overload the buffer. Otherwise, it rejects the request. Consequently, the average number of free slots is overall a higher value compared the scenario when WFIAC admits connection without the load estimation (Figure 6.7).

Figure 6.12 indicates that when congestion in the core network increases, the average throughput of all connections in the network decreases. This is due to the fact that when the queue length increases above the target operating point; WFICC starts its degradation procedure and reduces the rate allocated to existing connections with the view to bring the queue length close to the target point. Furthermore, it can be contributed to the increasing delay (Figure 6.10). As soon as the queue length reduces and remains around the target operating point, the

throughput becomes stable. So, even if the average number of free slots is higher (Figure 6.11), the overall average throughput of the network is higher as compared to the scenario when WFIAC admits connections without the load estimation (Figure 6.8).

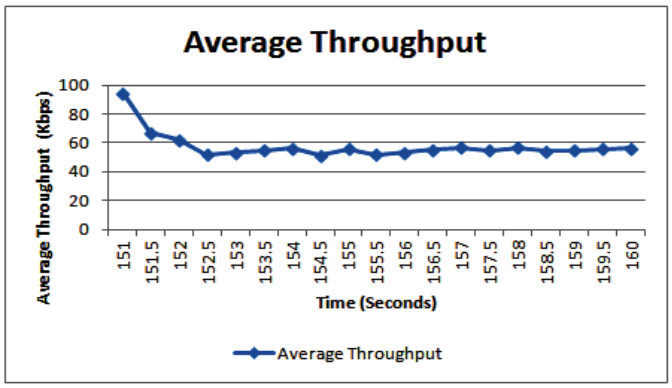


Figure 6.12 Average Throughput (kbps) with load estimation

So, when WFIAC admits or rejects incoming connections based on the load estimation in the network, it can maintain the queue length at a base station output buffer around the target operating point. This results in smaller average delay, stable and slightly better average throughput in the network.

6.5 Discussion on Results

This section provides discussion on the results of WFIAC scheme provided in above section.

Bounded Queue Length

WFIAC with the load estimation ensures that the queue length at an output buffer of a base station operates around the target operating point. Consequently, the network achieves stable throughput and reduced delay in the network.

Fair Resource Allocation among CoSs

The degradation procedure of WFIAC ensures fair resource allocation among the different CoSs, as it always starts the degradation with the lowest priority CoS. When the rate of the lowest priority level reaches the minimum rate and still enough resources are not obtained, it then

moves to the next lower priority levels. Consequently, more degradation is applied on low priority and the over provisioned CoSs.

Fair Resource Allocation among Flows of same CoS

WFIAC provides fair resource allocation among the connections of the same CoS. It is because of the two main reasons. Firstly, it assigns an incoming connection the same expected rate as used by the existing connections of its CoS, resulting in all connections of the same CoS using the same rate. Secondly, WFIAC degradation procedure reduces the expected rate of each CoS rather than reducing the individual connection's rate. Therefore, all connections of the same CoS employ the same reduced expected rate. In this way, WFIAC ensures that resources are allocated fairly among the service flows of the same CoS.

6.6 Summary

In this chapter, we presented an intelligent admission control scheme, WFIAC. By using different network scenarios we presented a performance evaluation of the proposed WFIAC scheme. The results illustrated that when the network is lightly loaded, WFIAC performs extremely well in terms of lower blocking probability and preserving the QoS of existing connections. When the network is congested there is a tradeoff between the blocking probability of connections of high priority CoSs such as UGS, and the QoS of existing connections. WFIAC ensures the QoS of existing connections in terms of bandwidth and delay guarantee by rejecting incoming connections until the queue length at a base station output buffer reduces to the desired target operating point. As soon as the queue length reaches the target operating point, WFIAC starts admitting connections. The results also indicate that when the network is lightly loaded, higher bandwidth is allocated to existing connections. This causes the extra allocation increase the efficiency of the network, and yet the network is kept stable around the target operating point. So, the proposed admission scheme is well suited to the networks, which aim to ensure the QoS of existing connections in terms of throughput and delay.

Chapter 7 Radio Admission Control for LTE

Radio Admission control (RAC) is a key function of the Radio Resource Management (RRM) at layer 3 of an eNodeB. However, 3GPP standard does not specify RAC and left it as an eNodeB vendor specific. Most admission control schemes proposed to date do not provide fairness among the users.

In this chapter, we introduce a novel RAC scheme for Long Term Evolution (LTE) and LTE-Advanced networks. To guarantee maximum resource utilization, it avoids employing thresholds and resource reservation for any specific type of service. In state of limited resource availability, it employs a step-wise degradation scheme to obtain resources for an incoming high priority bearer. It also considers the effect of channel fluctuations on user's resource demand and reserves extra resources to enable connections to maintain minimum requested QoS. Moreover, it includes a congestion control module. Based on the feedback from CC module, it infers the current network capability for each service type. LTE-FIAC intelligently admits or rejects an incoming connection based on its QoS requirements and the current network capability, to preserve the QoS of existing traffic flows in the network.

To analyze the performance of the proposed RAC scheme, comprehensive simulations have been performed in OPNET. The rest of the chapter is organized as follows.

Section 7.1 discusses the changes required in the OPNET simulator and the scheduler that are necessary to implement our RAC scheme. Section 7.2 comprehensively describes the proposed LTE-FIAC scheme. Section 7.3 discusses the performance evaluation of LTE-FIAC scheme. It describes the simulation setup and discusses the simulation results.

The notations for system level and bearer level parameters that we use in this chapter are as follows.

Notations

$Total\_PRB_x$	Total amount of bandwidth available in terms of Physical Resource Blocks (PRBs) for each direction 'x' that is Uplink (UL) or Downlink (DL).
PRB_{used_x}	Total number of PRBs used in 'x' direction.
PRB_{avail_x}	Available PRBs in the resource pool. $PRB_{avail_x} = Total\_PRB_x - PRB_{used_x}$
PRB_{REQ_x}	PRB Requested by an incoming connection in 'x' direction.
PRB_{REM}	It is the gap between PRBs requested by an incoming connection and PRBs available in the available resource pool. $PRB_{REM} = PRB_{REQ_x} - PRB_{avail_x}$ (If $PRB_{REQ_x} > PRB_{avail_x}$ Else $PRB_{REM} = 0$)
PRB_{CAP}	PRB Capacity in bits.
PRB_{0_x}	Target total number of PRBs in use after LTE-FIAC completes degradation procedure, ($PRB_{used_x} - PRB_{REM}$).
r_{MBR}, r_{GBR}	Maximum Bit Rate (MBR) and Guaranteed Bit Rate (GBR) requested by an incoming connection.

7.1 eNodeB Scheduler

The LTE module of OPNET release 17.1.A (OPNET, 2012) admits GBR connections through an admission control. However, non-GBR connections are admitted by default.

The manner in which LTE module of OPNET allocates resources to existing connections is discussed in Section 5.2. Section 5.4 discussed the enhancements to the simulator to support LTE-Advanced feature that allows to set the MBR attribute higher than the GBR of an EPS bearer. To effectively manage the added feature of LTE-Advanced to operate with LTE-FIAC, we propose the allocation of resources to bearers in the following order. 1) GBR connections should be given the resources using Proportional Fair (PF) scheduling scheme to meet their GBR and delay requirements. 2) Once GBR and delay requirements of the GBR connections are satisfied, the bearers are given resources to gain their respective MBR. Thus, to avoid the situation, where non-GBR connections are deprived of resources, we assign a minimum GBR to non-GBR connections as well as admit them through the admission control procedure. 3) In situations, when resources are left after step 2, the non-GBR or GBR connections are allocated

resources whichever of either two can take them using the Modified-RR (M-RR) algorithm (Section 5.2).

It is possible to use round robin or Modified-RR scheduler in step 2 to assign resources to GBR connections above their GBR requirements to gain their MBR. The issues with using RR and M-RR schedulers in step 2 are: a) they can keep on scheduling connections in a round robin manner until queues are empty. So, in step 2 GBR connections may be given resources above their MBR requirements, which can adversely affect throughput of non-GBR connections. b) These scheduling schemes do not consider the MBR parameter of bearers. They may select a connection and schedule it when it is not given resources above its GBR requirements by the admission control (such as in the proposed RAC scheme, non-GBR connections are given resources only for their GBR requirements).

To address these issues, we enhanced the PF scheduler given in the simulator to assure that the connections, which are given resources to meet their respective MBR by LTE-FIAC, get resources by the scheduler to gain up to their MBR in step 2.

To guarantee MBR, the enhanced PF (E-PF) operates in two phases. It separately grants resources to connections to meet their GBR and GBR_Ad[1] requirements. In the first phase, it serves GBR connections to guarantee their GBR and delay requirements, similar to original the PF scheme in the simulator. In the proposed second phase, it selects and serves GBR connections, which are assigned resources by LTE-FIAC to attain their MBR. It allocates PRBs to connections corresponding their GBR_Ad and delay requirements to guarantee up to their respective MBR.

The significant feature of E-PF is that it enters in the second phase when GBR and delay requirements of all GBR connections are gained after the first phase. This feature of E-PF is very useful in case of varying channel conditions. In situations, when the channel condition degrades,

---

[1] The bandwidth allocated above GBR requirements of GBR connections to meet their MBR is referred as GBR_Ad in this thesis. GBR_Ad = MBR-GBR. Say MBR=128 kbps and GBR=32 kbps, GBR_Ad =96 kbps

User Equipment's (UE) demand increases for PRBs to guarantee GBR. The connections, which are given PRBs by the RAC procedure to guarantee up to their GBR, their QoS degrades significantly in this situation. As discussed earlier, in every subframe the E-PF executes the first phase. It continues scheduling GBR connections until their GBR and delay requirements are satisfied. In the first phase of E-PF the PRBs, which are assigned by LTE-FIAC to connections in addition to their GBR requirements, can be allocated by the scheduler to connections whose channel conditions degrade to guarantee their GBR. Hence, the E-PF enables the network to manage changes in demand of PRBs caused by channel variations and guarantees the GBR of connections.

7.2 Description of LTE-FIAC

In this section, we describe proposed RAC, the Fair Intelligent Admission Control scheme for LTE-Advanced networks (LTE-FIAC). LTE-FIAC includes four main components: Congestion Control Module (CCM); Extra Resource Reservation Module (ERRM); Connection Arrival Procedure (CAP); and Connection Departure Procedure (CDP). The coordination among the modules and information, which is shared among the components, is shown in Figure 7.1.

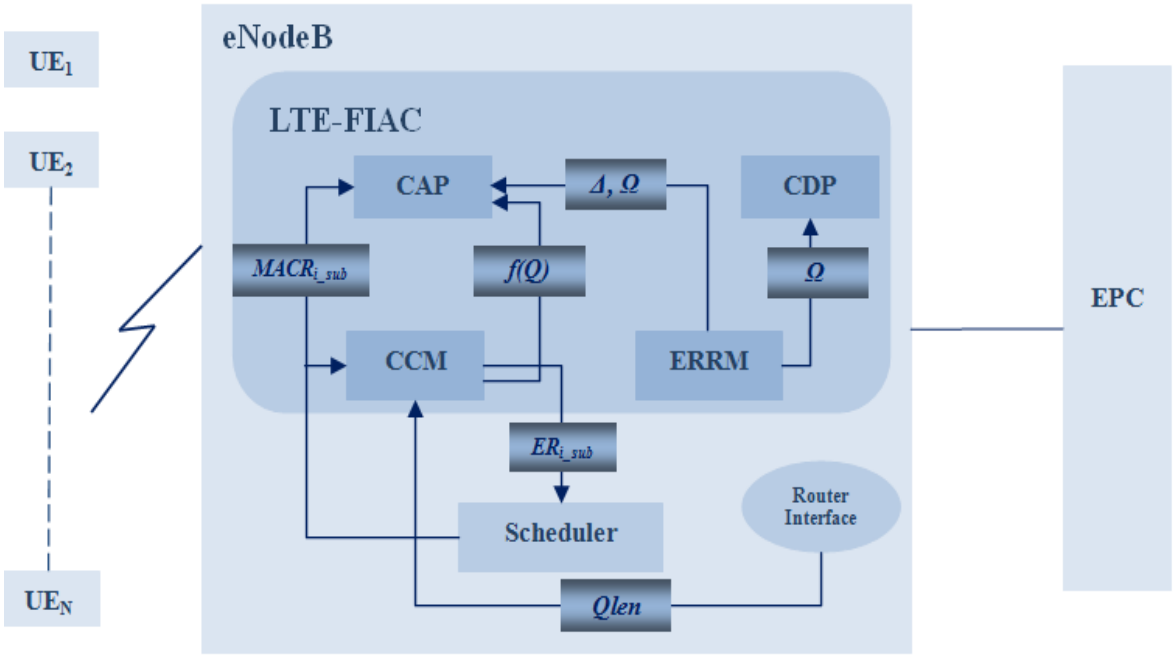


Figure 7.1 LTE-FIAC at an eNodeB

LTE-FIAC based on congestion conditions on the core side of the network decides new connection admission to avoid QoS degradation of existing traffic flows. The CCM of LTE-FIAC identifies congestion in the core network by evaluating the queue length at an output buffer of an eNodeB (Figure 7.1).

LTE-FIAC, at the time of admitting an incoming request, reserves extra resources to enable the network to provide GBR to connections with degrading channel conditions. The ERRM determines the amount of extra resources to be reserved based on the history of the resource usage of connections to gain their GBR.

The CAP is invoked, when a request for new EPS bearer arrives. To determine whether to admit or reject request for new connection it coordinates with the modules of extra resource reservation and congestion control (Figure 7.1). From ERRM, it obtains the amount of extra resources required to be reserved with an incoming connection. From CCM, it takes feedback about the queue control function to estimate the rate that the network can offer to a new bearer.

The CDP is invoked at the termination of an existing EPS bearer. Prior to allowing CAP to admit any waiting connection, CDP coordinates with ERRM (Figure 7.1). It obtains the amount of extra resource reserve required for each connection and reserves the required resources in the network, thus enables the network to deal with channel fluctuations.

The detail of all these components is discussed in following sub-sections.

7.2.1 Congestion Control Module (CCM)

We use LTE-FICC, described in Chapter 5, as the CCM of LTE-FIAC. LTE-FICC defines a target operating point (Q_0) at an output buffer of an eNodeB and dynamically maintains network traffic around the target operating point. It employs a rate allocation scheme, which takes into account the congestion at an output buffer of an eNodeB. It ensures that the queue length at an eNodeB's output buffer neither reaches the maximum buffer capacity nor becomes empty ensuring link is never idle unnecessarily.

7.2.2 Extra Resource Reservation Module (ERRM)

In this chapter, with respect to resource availability the network is termed as lightly, moderately or highly loaded. The network is lightly or moderately loaded when many or at least few connections have additional PRBs above their GBR requirements, to gain their respective MBR. When the network is lightly or moderately loaded, the E-PF scheduler enables the network to deal with channel fluctuations and guarantee the GBR of connections (Section 7.1). The LTE-FIAC continues to admit connections, through degrading the rate allocated to low priority connections. When the connections are admitted to a point where almost all connections in the network are at or close to their respective GBR, the network is in high load. In this situation, E-PF shall not be able to manage changes in user's resource demand caused by channel fluctuations. Consequently, the QoS of most of the connections will degrade.

To overcome this issue, LTE-FIAC reserves additional resources for an incoming bearer. It reserves extra resources with an incoming bearer and existing connections in form of additional resources allocated above their GBR requirements. However, in situations when there is no such connection, which can take additional resources, extra resource reserve is maintained in the available resource pool. Additional reserve is maintained only to guarantee GBR of connections.

ERRM estimates the required extra reserve of resources based on the history of mean variation in PRBs usage to attain the GBR**.** LTE-FIAC estimates the required reserve per connection during time 't' (Δ(t)) as follows.

$$\Delta(t) = \frac{\sum_{i=0}^{N}\sum_{j=0}^{n_i} \begin{cases} Sch\_PRB_j^{GBR}(t) - \dfrac{PRB_j^{GBR}}{t}, & Sch\_PRB_j^{GBR}(t) > \dfrac{PRB_j^{GBR}(t)}{t} \\ 0, & otherwise \end{cases}}{N_{total}} \quad 7.1$$

In Eq. 7.1, 'i' is an index of each QCI and 'j' refers to each data flow within each QCI. The PRB_j^{GBR} refers to the number of PRBs allocated by LTE-FIAC to a connection 'j' to meet its GBR per second. It depends on Signal-to-Noise Ratio (SNR) of user 'j' at time of its admission. The $Sch\_PRB_j^{GBR}(t)$ is the number of PRBs allocated by a scheduler to a connection 'j' to meet

its GBR in time 't' (say 300 ms (.3 s) or a second (1 s)). It depends on variation in SNR of user 'j' in time 't'. The N_{total} refers to the total number of connections in the network.

In Eq. 7.1, PRBs utilized to guarantee GBR are considered only, as this rate is guaranteed by the network. Also, LTE-FICC does not apply to GBR of bearers (Section 5.3). Whereas, to control congestion at an eNodeB output buffer, LTE-FICC can reduce MBR of a connection 'j' up to its GBR. Therefore, changes in the number of PRBs to achieve the GBR of connections in time 't' are employed to measure the affect of channel fluctuations on the PRBs demand. In Eq. 7.1, Δ(t) is the estimation of change in PRBs demand in time 't'. To obtain the history of changes in PRBs demand, we further take exponential average so that Δ changes smoothly.

$$\Delta(t) = \Delta(t-1) + \varepsilon(\Delta(t) - \Delta(t-1)) \quad 7.2$$

In Eq. 7.2, ε is an exponential average factor. It determines how fast old value of Δ converges to the new estimated value determined in Eq.7.1.

The network guarantees only GBR to connections, so LTE-FIAC takes into account the additional PRBs above the GBR requirements of connections and estimates an average available reserve per connection (Δ_e(t)). The available reserve indicates the number of PRBs already reserved with existing connections above their GBR requirements. It assists the network to avoid an over reservation of resources. It is estimated as follows.

$$\Delta\_e(t) = \frac{\sum_{i=0}^{N} \sum_{j=0}^{n_i} PRB_{jx}^{MBR} - PRB_{jx}^{GBR}}{N_{total}} \quad 7.3$$

In Eq. 7.3, PRB_{jx}^{MBR} and PRB_{jx}^{GBR} represents PRBs allocated per second to a bearer 'j' in direction 'x' (UL or DL) to gain its MBR and GBR requirements, respectively. Equation 7.3 is used when the ERRM is either invoked by the CDP, or it is executed periodically to update the extra resource reservation parameters. However, the value of available reserve per connection changes, depending upon whether an incoming connection is admitted at its GBR or MBR. Consequently, when the ERRM is invoked by the CAP, it utilizes Eq. 7.4 or Eq. 7.5.

When the admission procedure attempts to admit a connection at its MBR, LTE-FIAC employs Eq. 7.4 to estimate the value of available reserve per connection (Δ_e(t)).

$$\Delta\_e(t) = \frac{\sum_{i=0}^{N}\sum_{j=0}^{n_i} PRB_{jx}^{MBR} - PRB_{jx}^{GBR} + (PRB_{REQ_x}^{MBR} - PRB_{REQ_x}^{GBR})}{N_{total} + 1} \quad 7.4$$

Equation 7.4 considers the resources that are to be allocated to an incoming connection above its GBR requirements, and avoids the over reservation of resources. However, when resources in the available resource pool are enough to guarantee the GBR of an incoming connection, LTE-FIAC utilizes Eq. 7.5 to estimate Δ_e(t). As a result, it avoids the under reservation of the extra resources.

$$\Delta\_e(t) = \frac{\sum_{i=0}^{N}\sum_{j=0}^{n_i} PRB_{jx}^{MBR} - PRB_{jx}^{GBR}}{N_{total} + 1} \quad 7.5$$

LTE-FIAC estimates the gap Ω between the required reserve per connection and the available reserve per connection as shown in the following equation.

$$\Omega = \Delta(t) - \Delta\_e(t) \quad 7.6$$

The gap Ω in Eq. 7.6 indicates an averaged number of PRBs per connection, which actually needs to be reserved per connection before admitting a new connection. Figure 7.2 shows the steps involve in the ERRM.

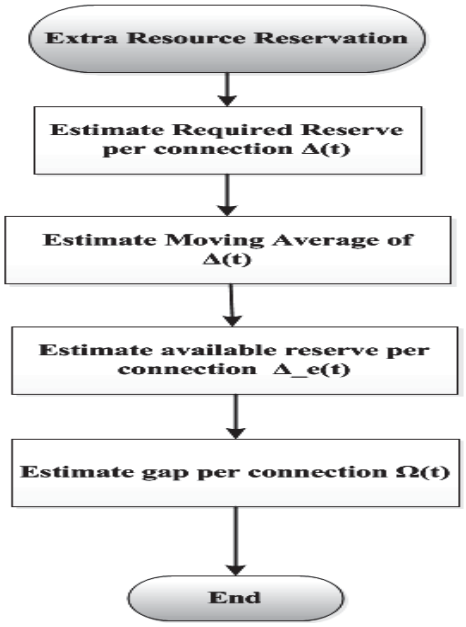


Figure 7.2 Procedure of Extra Resource Reservation Module of LTE-FIAC

CAP should set admission conditions in a manner that a new connection is admitted, when the network has enough resources to gain the GBR and the extra resource reserve per connection for new and existing connections.

7.2.3 Connection Arrival Procedure (CAP)

When a new connection request arrives with PRB_{REQ}, LTE-FIAC executes the CAP. The main steps of arrival procedure are: a) it sets connection status to *ready to admit*, when resources are sufficient in the available resource pool to provide either MBR or GBR of an incoming connection and the extra resource reserve. b) It sets a connection status to *ready to admit* only when it estimates that resources can be obtained by reducing the rate allocated to existing low priority connections. Otherwise, it sets connection status to reject. This step is invoked only when resources are not enough and an incoming connection has a high priority. c) It sets a connection status to *ready to admit* when it estimates that LTE-FICC can handle the load of an incoming connection. Otherwise, it sets connection status to reject. This step is executed when connection status is ready to admit after step a or b. d) It applies degradation only when the CCM sets the connection status to ready to admit, if degradation is required, this step is invoked and degradation is applied on connections of low priority.

The detail of each step is given below.

a. No Degradation

When a connection request arrives, LTE-FIAC estimates resource availability in the available resource pool using following equation.

$$PRB_{REQx} = \begin{cases} PRB_{REQ_x}^{MBR} \quad , & \left[\sum_{i=0}^{N}\sum_{j=0}^{n_i}(PRB_{jx}^{MBR}+\Omega)\right] + (PRB_{REQ_x}^{MBR}+\Omega) \le Total\_PRB_x \\ PRB_{REQ_x}^{GBR} \quad , & \left[\sum_{i=0}^{N}\sum_{j=0}^{n_i}(PRB_{jx}^{MBR}+\Omega)\right] + (PRB_{REQ_x}^{GBR}+\Omega) \le Total\_PRB_x \end{cases} \quad 7.7$$

In Eq. 7.7, LTE-FIAC verifies resource availability for both UL and DL directions as indicated by ‘x’. The Ω indicates PRBs per connection, which needs to be set-aside in the available

resource pool to effectively deal with channel fluctuations. Hence, employing Ω is equivalent to set a limit on the number of PRBs in the available resource pool that can be allocated to new connections, with the aim to maintain the QoS of existing bearers. When channel condition improves, value of Ω and thus the limit reduces. As a result, LTE-FIAC can admit more connections in the network.

When resources are sufficient in the available resource pool, to meet MBR or GBR as well as the extra resource reserve requirements for an incoming request and ongoing connections in the network, LTE-FIAC sets the connection status to *ready to admit*. When a connection is admitted at its GBR, it also sets r_{MBR} *equal to* r_{GBR}.

b. Degradation

In situations, when enough resources are not available in the network to admit an incoming connection, LTE-FIAC admits high priority connections using a step-wise degradation scheme. The degradation allows RAC to ensure differentiation among the connections at different priority levels. It is applied for the maximum resource gap either in UL, or DL direction. Following equations are applied to validate that enough resources can be obtained by degrading the rate allocated to connections with lower priorities.

$$PRB_{REQX} = \begin{cases} PRB_{REQ_x}^{GBR}\ , & \left[\sum_{i=0}^{N}\sum_{j=0}^{n_i} PRB_{j_x}^{MBR} - \{f(pr_j, pr_{new})\,(PRB_{j_x}^{MBR} - PRB_{j_x}^{GBR})\}\right] \\ & \quad + PRB_{REQ_x}^{GBR} \le Total\_PRB_x \\ 0\ , & otherwise \end{cases} \qquad 7.8$$

In Eq. 7.8, 'pr_{new}' and 'pr_j' represent degradation priorities of an incoming new bearer and an existing bearer 'j', respectively. In LTE-FIAC, degradation priority of a bearer is determined based on both QCI and ARP priority. It is estimated using Eq. A.1 in Appendix A. Function '$f(pr_{new}, pr_j)$' in Eq. 7.8 is similar to a function given by (Kwan. et al., 2010). It returns 1 only if bearer 'j' has higher degradation priority than new bearer. Otherwise, it returns 0. Equation 7.8 shows that degradation is applied to attain resources only to gain GBR of an incoming connection. It does not consider Δ(t). It is due to the fact that extra-required resources may already be reserved with existing connections at any priority level in the form of additional

resources above their GBR requirements. When Eq. 7.8 returns a value greater than 0, indicating resources for an incoming connection can be obtained by degrading connections with lower priorities, Eq. 7.9 is executed.

$$PRB_{REQ_X} = \begin{cases} PRB_{REQ_x}^{GBR} , & \left[\sum_{i=0}^{N}\sum_{j=0}^{n_i} PRB_{j_x}^{MBR} - \left(PRB_{j_x}^{MBR} - PRB_{j_x}^{GBR}\right) + \Delta(t)\right] \\ & + PRB_{REQ_x}^{GBR} + \Delta(t) \leq Total\_PRB_x \\ 0 , & otherwise \end{cases} \quad 7.9$$

In Eq. 7.9, Δ(t) is an estimate of an average number of PRBs required per connection as estimated in Eq. 7.2, which should be allocated to cope with an increase in PRBs demand due to the channel fluctuations. Equation 7.9 shows that degradation is applied only if Total_PRBs$_x$ in the system in direction 'x' is sufficient to provide: a) PRBs to guarantee the GBR and the required reserve (Δ(t)) for all existing connections; and b) requested PRBs to attain the GBR and the required reserve Δ(t) for an incoming connection. Similar to Eq. 7.7, employing Δ(t) in Eq. 7.9 is equivalent to setting a limit on PRBs that can be allocated to an incoming connection with a view to ensure the QoS of existing bearers. When channel condition improves, Δ(t) and hence the limit reduces. Consequently, the RAC can admit more connections in the network.

Equation 7.8 indicates that LTE-FIAC degrades the rate above the GBR of only the connections with priority lower than the incoming connection's priority. However, during channel fluctuations, the resources from connections of any priority level can be given to a connection in order to meet its GBR requirements. Consequently Eq. 7.9 considers resources allocated above the GBR of connections on all priority levels. Using Eq. 7.8 and Eq. 7.9, if LTE-FIAC estimates that adequate resources can be obtained for both directions by degrading connections of lower priorities, it sets the new connection status to *ready to admit* and also sets r_{MBR} *equal to* $r_{GBR.}$. Otherwise, it rejects an incoming request.

c. Load Estimation (LE)

In times when the core network is congested, output buffer at an eNodeB becomes overloaded. When a RAC unknowingly keeps admitting connections, it aggravates congestion state at an output buffer of an eNodeB. It has been observed that during congestion, the network cannot

guarantee desired QoS to connections of different priority classes (Li and Hoang, 2005). LTE-FIAC with load estimation addresses this issue.

Once an incoming connection status is *ready to admit*, CAP of LTE-FIAC performs the load estimation to avoid buffer overloading and to ensure stability in the network. It admits an incoming bearer only if it determines that LTE-FICC will be able to manage the load of an incoming connection. Consequently, traffic of new connection is not going to overload an output buffer of an eNodeB. To determine this, LTE-FIAC estimates the current network capability for each service type. The network capability indicates the rate that the network can offer to an incoming connection of a specific service type based on the current congestion state at an output buffer of an eNodeB. LTE-FIAC compares the QoS requirements of an incoming connection with the network capability, and admits or rejects it to enable LTE-FICC to manage the network load.

To determine the network capability for each traffic class 'i' at connection arrival time 't', LTE-FIAC estimates Expected Admission Rate ($EAR_i(t)$). EAR_i is rate that the network can allow to a connection of class 'i' for its requested PRBs. It is based on the average resource usage of class 'i' and the buffer status at time 't' (Li and Hoang, 2005). Expected admission rate is defined as follows.

$$EAR_i(t) = f\big(Q(t)\big) * MACR_i(t) \qquad 7.10$$

In Eq. 7.10, $MACR_i$ is Mean Allowed Class Rate of class 'i'. It maintains the mean value of the rate allocated to all active connections per CoB (GBR or non-GBR) (Section 5.3.2). LTE-FIAC receives MACR for each class 'i' from the scheduler (Figure 7.1).

LTE-FIAC obtains the value of function of queue from LTE-FICC (Figure 7.1). The queue control function indicates the status of congestion at an output buffer of an eNodeB and hence controls the network capacity for an incoming flow. When the queue length at a buffer operates above the target operating point, EAR reduces due to f(Q(t)) <1 (refer to Figure 3.3). Consequently, it limits admission of new connections in the network to maintain the QoS of existing connections. However, when the queue length is below the target operating point, EAR increases and allows RAC to admit more connections in the network.

LTE-FIAC grants a new connection request of class 'i', if its rate [r_{MBR}] (established in Eq. 7.7 or Eq. 7.8) is less than or equal to the network capacity offered to its class that is $EAR_{i,}$. Otherwise, it rejects an incoming request to avoid buffer congestion at an eNodeB.

LTE-FIAC Load Estimation (LE) Algorithm

Calculate f(Q) as follows.

IF (Qlen > Q_0)

$$f(Q) = \frac{(Buffer\_Size - Qlen)}{(Buffer\_Size - Q_0)}$$

Else

$$f(Q) = \frac{(\alpha - 1) * (Q_0 - Qlen)}{Q_0} + 1$$

End IF

Estimate EAR_i using Eq. 7.10

IF $r_{MBR} \leq EAR_i$

Admit = TRUE

Else

Admit = FALSE

END IF

Figure 7.3 provides the steps involve in the procedure of load estimation of LTE-FIAC.

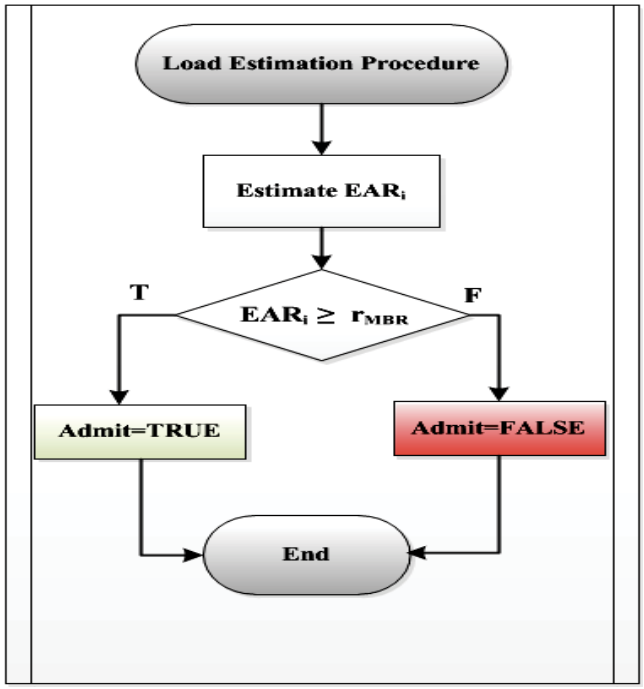


Figure 7.3 Procedure of Load Estimation of LTE-FIAC

d. Step Wise Degradation

When the network resources are limited, a RAC can admit high priority connections using a degradation scheme. LTE-FIAC uses a step-wise degradation scheme and applies a variable size degradation step. The proposed step size takes two parameters as an input, resources allocated to a bearer above its GBR requirements (GBR_Ad), and an intelligent function of PRBs. The variable size degradation step is calculated as follows.

$$\delta_{(\mathrm{pr,j})_x} = f(PRBs)_{(\mathrm{pr,j})_x} * GBR\_Ad_{(\mathrm{pr,j})_x} \quad 7.11$$

In Eq. 7.11, $\delta_{(\mathrm{pr,j})_x}$ is a degradation step size and represents an estimate of resources, which can be taken from a connection 'j' with priority level 'pr', stated by its QCI and ARP priorities. LTE-FIAC admits a connection either at its MBR, or GBR. The GBR_Ad is introduced in the degradation step size due to the upgradation of the rate allocated to the existing connections by the CDP of LTE-FIAC. The CDP will be discussed in detail in sub section 7.2.4. Additionally, it is possible that due to the degradation some connections are operating at a rate between their GBR and MBR. Whereas some connections may be admitted later at the same priority level with their MBR. Hence, at the same priority level there can be some connections with MBR, while other can have a rate above GBR but less than MBR. The parameter GBR_Ad in degradation step size allows more degradation to be applied on connections that have higher rate compared to other bearers at the same priority level.

LTE-FIAC degradation procedure does not take into account channel conditions of users directly, but these are considered automatically during the degradation process. It is because, the degradation procedure converts the step size in Eq. 7.11 to the number of Physical Resource Blocks (PRBs), which can be taken from a connection 'j', based on its channel conditions, using Eq. 7.12.

$$\delta_{(\mathrm{pr,j})_x} = f\left(\delta_{(\mathrm{pr,j})_x}, MCS_j\right) \quad 7.12$$

In Eq. 7.12, $f\left(\delta_{(\mathrm{pr,j})_x}, \mathrm{MCS_j}\right)$ returns the step size in terms of PRBs, which can be taken from a connection 'j' based on its step size in Eq. 7.11, and its channel conditions. The capacity of a PRB is small for users, which are in bad channel conditions and have lower MCS. Consequently,

the degradation step size for bearers of these users is relatively higher compared to the users, having good channel conditions and higher MCS. The degradation step size, therefore, applies more degradation on connections, which are over provisioned, and are in less favorable channel conditions. Both these type of users occupy more resources compared to other users with the same priority in the network. In this way, LTE-FIAC scheme ensures fair resource allocation among the connections at the same priority level.

As discussed, the degradation is applied only on connections with lower priority compared to the priority of an incoming connection. LTE-FIAC degradation process also ensures fairness among the service flows of different priority levels, which are candidates to be degraded. The degradation step size employs a function of PRBs to ensure that different size of degradation is applied to connections at different priority levels. The function of PRBs to determine the contribution of a bearer in the degradation procedure is as follows.

$$f(PRBs)_{(pr,j)_x} = 1 - \frac{Total_{PRB_x} - \max\left(\left(PRB_{used_x} - w\_deg_{(pr_j)}\right),\ PRB_{0_x}\right)}{Total\_PRB_x - PRB_{0_x}} \quad 7.13$$

In Eq. 7.13, $f(PRBs)_{(pr,j)_x}$ adjusts the contribution of a connection 'j' in the degradation process in accordance to its priority level 'pr'. The PRB_{used_x} is the total number of PRBs occupied by the existing connections in direction 'x'. The PRB_{0_x} is the target number of used PRBs after the degradation process completes. The $w\_deg_{(pr,j)}$ represents the total number of PRBs, which can be obtained from lower priority connections compared to the priority of a connection 'j'. It is estimated as follows.

$$w\_deg_{(pr_j)} = \sum_{i=0}^{N}\sum_{k=0}^{n_i} f(pr_k, pr_j)\left(PRB_{k_x}^{MBR} - PRB_{k_x}^{GBR}\right) \quad 7.14$$

Equation 7.14 returns zero for the lowest priority connections. Hence, $f(PRBs)_{(pr,j)_x}$ is very high for connections with the lowest priority. The value of $f(PRBs)_{(pr,j)_x}$ in Eq. 7.13 keeps reducing for the next lower priority levels. So, $f(PRBs)_{(pr,j)_x}$ allows variable size degradation step and ensures different amount of degradation is applied to connections with different priorities. Importantly, when resources can be obtained from lower priority connections, no

degradation is applied at the next lower priority, though its priority is lower than the priority of an incoming connection.

LTE-FIAC step-wise degrades existing connections until the required resources are obtained. To convert the degradation process to a step-wise degradation, input of each bearer 'j' in the degradation procedure is estimated as follows.

$$\S_{(\mathrm{pr,j})_x} = \frac{\delta_{(\mathrm{pr,j})_x}}{\sum_{i=0}^{N}\sum_{j=0}^{n_i}\delta_{(\mathrm{pr,j})_x}} * PRB_{REM} \qquad 7.15$$

As a result of step-wise degradation, LTE-FIAC brings slightly higher resources in the available resource pool than needed by an incoming connection as opposed to the RAC scheme given by (Priya and Franklin, 2012, Qian et al., 2009). Thus, the f(PRBs) ensures that resources are retained with existing connections to admit later arriving connections of high priority by applying the degradation procedure. Therefore, LTE-FIAC during the degradation prioritizes high priority connections to ensure better QoS provisioning to them. However, it does not monopolize network resources for high priority connection, as it also introduces upgradation procedure as part of the CDP, which upgrades the rate allocated to the connections. This enables the network to admit connections of other priorities as well, by applying degradation on connections of lower priorities. So, LTE-FIAC scheme ensures the fair share of bandwidth among the different service types.

The preemption scheme proposed by (Khabazian et al., 2013) also considers three parameters, including priority, QoS over provisioning, and channel condition of a user. The scheme applies a partial preemption on all connections at all priority levels. Depending upon the value of parameters, it may also apply degradation to connections of high priority, even when all resources can be obtained from connections of lower priority services. Additionally, after partial preemption, when as little as one additional PRB is required, it applies a full preemption, and drops connections starting from the lowest priority. Consequently, resources more than the required are added back in the available resource pool. The scheme does not consider redistributing the excess resources to existing connections using an upgradation scheme. The RAC can use these extra resources to admit connections of any QoS class. In this way, in times

of resource scarcity, the resources reduce that can be given to connections of high priority through the degradation. Consequently, blocking probability of high priority applications, especially with high bandwidth demand, increases and affects fairness among the connections at different priority levels. While, when sufficient resources are obtained from low priority connections, LTE-FIAC degrades only low priority connections. Thus, in times of network resources scarcity, it sets aside some extra resources with ongoing sessions for incoming high priority connections. Hence, LTE-FIAC prefers reserving resources with existing connections, instead of dropping ongoing connections to admit high priority connections.

The results by (Khabazian et al., 2012) shows that in times when the connection arrival rate increases, better fairness among the different service types can be achieved if priority of connections is given a weight close to 1. Also, when resource over-provisioning is assigned a weight close to 1, better fairness can be achieved among connections of the same priority level. The proposed degradation procedure of LTE-FIAC considers both the priority and the resource over-provisioning of bearers. It does not involve fine-tuning of parameters. As a result, it enables the network to ensure the QoS among the connections at the same as well as different priority levels with less dependency on parameters fine-tuning by the service provider.

LTE-FIAC degradation scheme follows the following rules.

1. LTE-FIAC applies degradation only if it estimates that LTE-FICC can manage the load introduced by an incoming connection.

2. The degradation is applied on flows with lower priority compared to the priority of an incoming connection.

3. When executing degradation, a connection bandwidth can be reduced only to its GBR.

The detailed steps of the degradation procedure to obtain PRB_{REM} are as follows.

Step 1: LTE-FIAC estimates the degradation step size for each connection at each priority level. *Step 2:* All data flows are degraded equal to the step size. Consequently, step 2 enables LTE-FIAC to ensure fairness among the data flows with the same priority. *Step 3:* Based on degradation applied, LTE-FIAC updates GBR-Ad of degraded connections, remaining PRBs

(PRB_{REM}), used PRBs (PRB_{used_x}) and available PRBs (PRB_{avail_x}). In Eq. 7.8, it has already been validated that enough resources can be obtained so LTE-FIAC continues the degradation process until enough resources are obtained. It degrades bandwidth equal to exactly what is required. Hence, it retains resources with the low priority connections to admit incoming high priority connections. Consequently, it enables the network to provide the fair share of bandwidth among different priority classes by holding resources with existing connections to admit incoming high priority connections. Figure 7.4 shows the steps of the CAP.

LTE-FIAC Degradation Procedure Algorithm

```
While PRB_avail_x < PRB_REQ
      For each connection 'j'
            IF f(pr_j , pr_inc ) // Connection 'j' has higher degradation
                                    priority than an incoming connection
                  Estimate §_(pr,j)_x using Eq. 7.15
                  δ_bits = §_(pr,j)_x * PRB_CAP_jx
                  IF MBR_(pr,j)_x − δ_bits >= GBR_(pr,j)_x
                        MBR_(pr,j)_x = MBR_(pr,j)_x − δ_bits
                  ELSE IF MBR_(pr,j)_x > GBR_(pr,j)_x
                        δ_bits= MBR_(pr,j)_x− GBR_(pr,j)_x
                        MBR_(pr,j)_x = MBR_(pr,j)_x − δ_bits
                  END IF
                  GBR_Ad_(pr,j)_x= MBR_(pr,j)_x − GBR_(pr,j)_x
                  Update PRB^MBR_(pr,j)_x, PRB_avail_x, PRB_used_x and PRB_REM_xbased
                  on the new MBR_(pr,j)_x
            END IF
      END For
End While
```

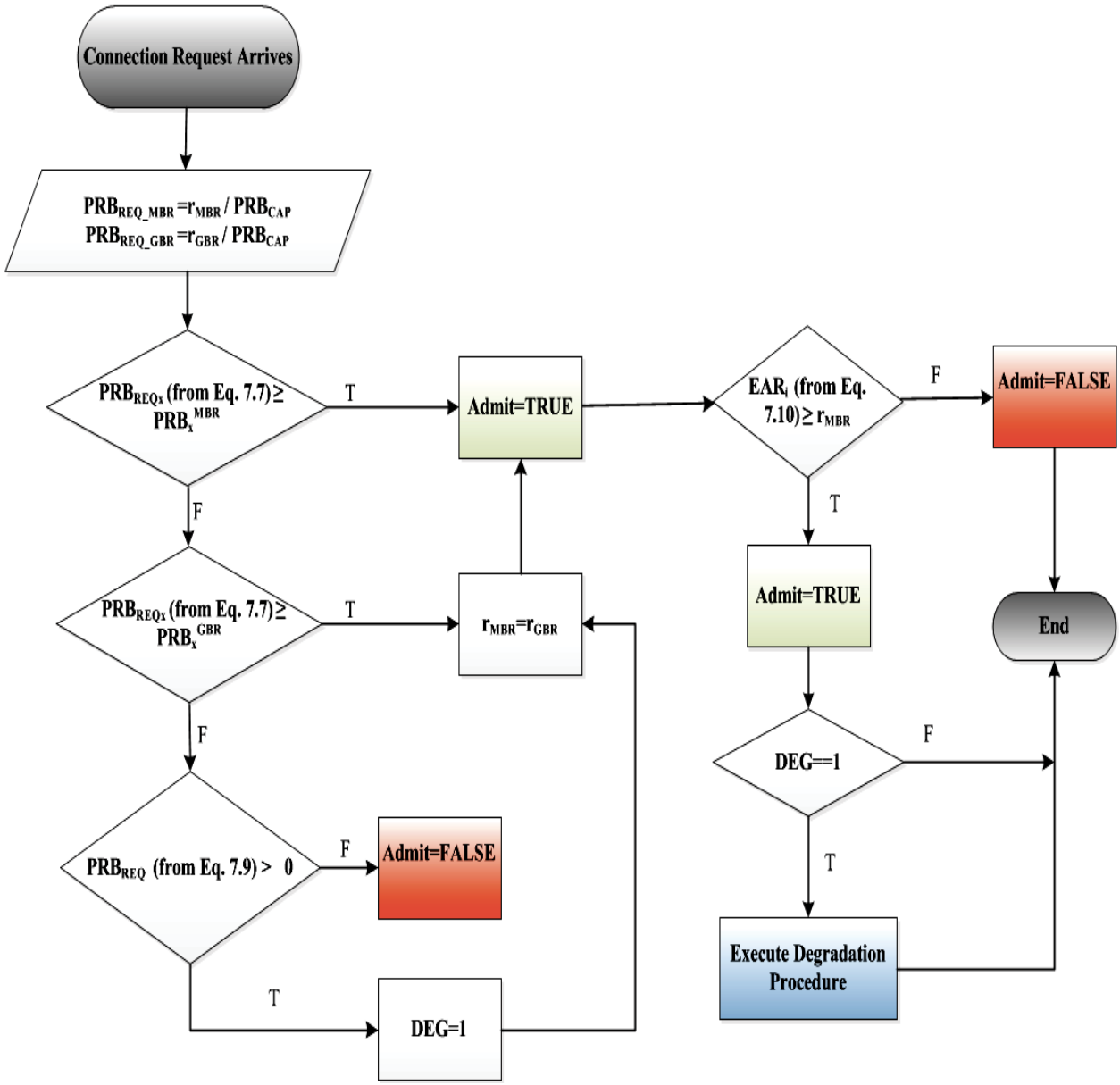


Figure 7.4 Connection Arrival Procedure of LTE-FIAC

7.2.4 Connection Departure Procedure (CDP)

When a dedicated EPS bearer is torn down due to inactivity, its resources are added back in the available resource pool of both UL and DL, respectively. With the aim to maximize throughput of a cell and to maintain the QoS of ongoing sessions, the CDP operates in three phases.

In the first phase, LTE-FIAC estimates the gap (Ω) using Eq. 7.6. To ensure that the network has sufficient resource reserve to deal with channel fluctuations, it reserves the gap (Ω) for all network connections. It uses a step-wise upgradation to allocate the resource reserve to the existing connections to gain their MBR. It reserves either the PRBs in the available resource pool, or the total reserve required for all connections ($\Omega * N_{total}$), whichever is the minimum. It enables the network to deal with channel variations and ensures minimum guaranteed QoS.

LTE-FIAC aims to maximize the network resource utilization. So, in times when resources are left after phase one, the CDP executes the second phase. In the second phase, it checks for any waiting connection in the queue and tries to admit it using the CAP.

In scenarios when resources are left after the second phase, it moves to the third phase. In the third phase, it uses a step-wise upgradation to allocate free resources to the existing connections to gain their MBR. It also enables the network to deal with channel variations and to put aside resources with the existing connections for incoming connections of high priority.

Step Wise Upgradation

The upgradation procedure allocates available resources to connections based on their priority, stated by both QCI priority and ARP priority. To ensure the fair bandwidth share to connections of different priority classes, LTE-FIAC uses a stepwise upgradation scheme. The step size is a function of the f(PRBs); the gap between MBR of QoS class with priority level 'pr' in direction 'x' (MBR_{pr_x}) and MBR which a connection 'j' avails in direction 'x' ($MBR_{(pr,j)_x}$); and the capacity of a PRB of a connection 'j' in direction 'x' ($PRB\_CAP_{j_x}$) . The step size is calculated as follows.

$$\delta_{(pr,j)} = f(PRBs)_{(pr,j)} * \left(MBR_{pr_x} - MBR_{(pr,j)_x}\right) * PRB\_CAP_{j_x} \quad 7.16$$

In equation Eq. 7.16, $\delta_{(pr,j)}$ is the upgradation step size and represents an estimate of resources to be allocated to connection 'j' with QCI 'q' and ARP 'r'. The gap between MBR of a priority level 'pr' and MBR of a connection 'j' in the upgradation step size allows more upgradation to be applied on connections that avail less resources above their GBR requirements compared to other bearers at the same priority level. In Eq. 7.16, $PRB\_CAP_{j_x}$, represents the capacity of a PRB. It is

obtained in terms of bits from TB size table given in (3GPP 36.213) corresponding the MCS of a user 'j'. The $PRB\_CAP_{j_x}$ is small for users having bad channel conditions and lower MCS. As a result, the upgradation step size for such bearers is relatively lower compared to users with higher MCS. Hence, the upgradation step size ensures fairness among users at the same priority as more upgradation is applied to connections, which are less over provisioned and have favorable channel conditions compared to other users at the same priority level. In this way, it also helps to ensure that overall throughput of a cell increases.

LTE-FIAC upgradation process ensures fairness among the connections of different priority levels. The upgradation step size employs the function of PRBs to ensure that different size of upgradation is applied to connections at different priority levels. The function of PRBs for the upgradation step size is calculated as follows.

$$f(PRBs)_{(pr,j)_x} = \frac{\max\left(Total\_PRB_x - \left(PRB_{used_x} + w\_up_{(pr_j)}\right),\ 0\right)}{Total\_PRB_x} \quad 7.17$$

In Eq. 7.17, $w\_up_{(q_j,u_j)}$ represents the total number of PRBs, which are to be allocated to connections with higher priority compared to the priority of connection 'j'. It is estimated as follows.

$$w\_up_{pr_j} = \sum_{i=0}^{N}\sum_{k=0}^{n_i} f(pr_j, pr_k)\left(PRB_{pr_x}^{MBR} - PRB_{k_x}^{MBR}\right) \quad 7.18$$

Equation 7.18 returns zero for the highest priority connections. Therefore, the function of PRBs (Eq. 7.17) is very high for connections with high priority. Its value keeps reducing for the next higher priority levels. In this way, the function of PRBs allows variable size upgradation step and ensures different amount of upgradation is applied to connections with various priorities. To perform step wise upgradation and to ensure connections do not get more than the available resource, share of each bearer 'j' at a specific priority level is estimated using Eq. 7.19.

$$\S_{(pr,j)_x} = \frac{\delta_{(pr,j)_x}}{\sum_{i=0}^{N}\sum_{j=0}^{n_i}\delta_{(pr,j)_x}} * PRB_{Avail_x} \quad 7.19$$

So, LTE-FIAC scheme ensures fair share of bandwidth among different service types.

LTE-FIAC Upgradation Procedure Algorithm

```
While PRB_availx > 0
    Estimate S_(pr,j)x using 7.19
    For each connection 'j'
        δ_bits = S_(pr,j)x * PRB_CAP_jx
        IF MBR_(pr,j)x + δ_bits <= MBR_prx AND PRB_availx >0
                MBR_(pr,j)x = MBR_(pr,j)x + δ_bits
        IF MBR_(pr,j)x + δ_bits > MBR_prx AND PRB_availx >0
                δ_bits = MBR_prx - MBR_(pr,j)x
                MBR_(pr,j)x = MBR_(pr,j)x + δ_bits
        End IF
        Update PRB^MBR_(pr,j)x, PRB_availx, PRB_usedx based on new MBR_(pr,j)x
     End For
End While
```

The detailed steps of upgradation procedure are as follows.

Step 1: LTE-FIAC estimates the upgradation step size for each connection at each priority level. *Step 2:* All data flows at a certain priority level are upgraded, equal to the step size. Consequently, step 2 enables LTE-FIAC to ensure fairness among the data flows with the same ARP and QCI. LTE-FIAC applies the upgradation on MBR of data flows. Based on the upgradation applied, LTE-FIAC updates available PRBs (PRB_{avail_x}) and PRB_{used_x}. LTE-FIAC continues the upgradation process until either the free resources in the network are exhausted, or the MBR of all data flows increases to MBR of their respective priority level 'pr'. It upgrades bandwidth according to the priority of connections. Consequently, it ensures fair share of bandwidth among the different priority classes.

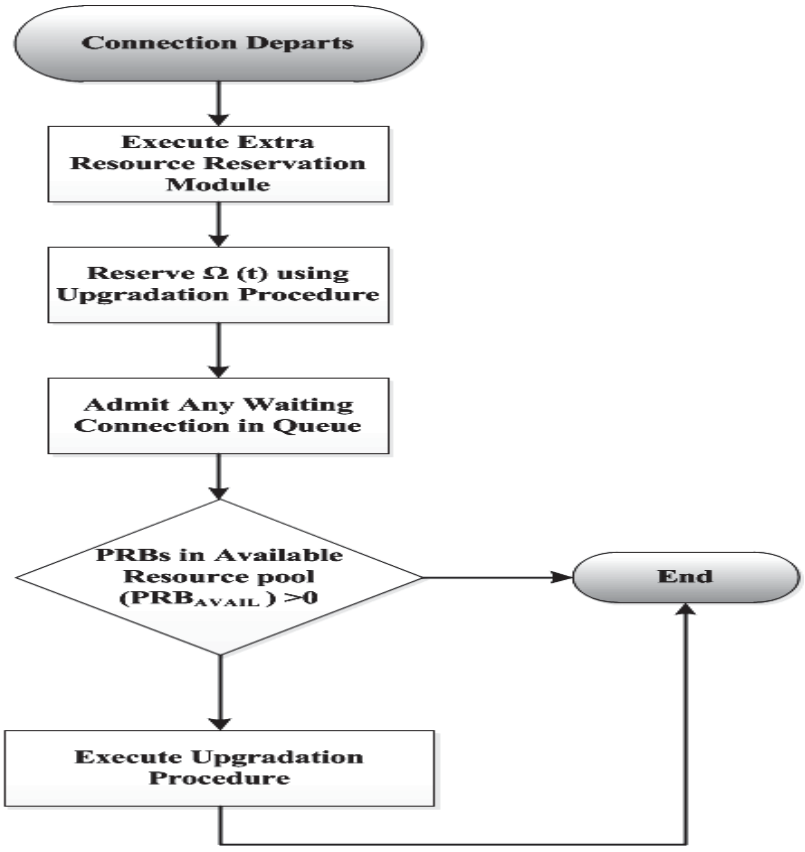


Figure 7.5 Connection Departure Procedure of LTE-FIAC

As discussed, to allocate resources to GBR connections, we employ the E-PF scheduler. In times when channel condition varies say due to change in weather, the E-PF scheduler treats PRBs allocated to GBR connections in excess to their GBR requirements, just as they are in the available resource pool and allocate them to any GBR connection to meet its GBR. This is the reason that the upgradation procedure is allowed to allocate resources until available resource pool is exhausted.

7.3 Performance Evaluation of LTE-FIAC

The overall goal of the simulation is to investigate the effectiveness of the proposed scheme in terms of call blocking probability, fairness among the service flows and QoS provisioning to existing connections. The simulation results are presented in the following sequence.

First, we discuss the performance of LTE-FIAC scheme in terms of connection Blocking Probability (BP). The BP of LTE-FIAC is compared with the BP of a tuned reference admission control algorithm per Transmission Time Interval (TTI) (Anas et al., 2008, Anas, 2008). The reference scheme operates as follows.

$$\sum_{i=1}^{K} N_i + N_{new} \leq N_{total} \qquad 7.20$$

In Eq. 7.20, N_{total} is the total number of PRBs in the network per second and K is the number of existing connections in a cell. The N_i represents the number of PRBs required per second by an active connection 'i' to satisfy its MBR requirements. The reference Admission Control (AC), hereafter referred as *Ref scheme*, grants a request if the sum of PRBs requested by a new (N_{new}) and the active connections is less than or equal to N_{total}.

The BP of LTE-FIAC is also compared with the BP of an admission control scheme (referred here as *Ref-Deg*) proposed by (Qian et al., 2009). In times of resource scarcity, the Ref_Deg scheme admits high priority connections by applying a degradation procedure to connections of low priority. LTE-FIAC degradation procedure is compared with the degradation process proposed by (Qian et al., 2009) in terms of fairness among the connections at the same as well as different priority levels. The Ref_Deg scheme one-by-one degrades connections of low priority. It degrades the rate of a connection directly to its GBR with a fixed size degradation step of $r_{MBR} - r_{GBR}$.

In times when the core network is congested, the queue at an eNodeB output buffer starts building up and results in high delay and reduced throughput. So, we investigate the effect of the proposed load estimation module on the QoS of existing connections in terms of throughput and queuing delay.

In state of channel variations, the demand of PRBs increases. In case, when no extra resources are reserved, the QoS of connections degrades. So, we also evaluate and verify the effect of the proposed extra resource reservation module on the QoS of ongoing connections and the BP of new connections.

7.3.1 Simulation Setup

The simulations are performed using LTE module of the system level simulator, Optimized Network Engineering Tool release 17.1.A. In the current simulation setup, UEs are connected to an eNodeB that in turn is connected to an EPC. The EPC is connected to a server through the Internet to reflect the actual deployment of an end-to-end network. The eNodeB operates in FDD mode and employs 3 MHz bandwidth. The target operating point is set at 1/32 of the total buffer capacity of 3 Mbps. The link capacity between the eNodeB and the EPC is set at 44.7Mbps.

For simplicity of analysis, the simulation model consists of a single cell based on the 3GPP LTE system model. The requests arrivals in the system are modeled by a Poisson process. The simulation includes three service classes (voice, video and web). The arrival rate of connections of each service type is set to be the same that is $\lambda_{voice} = \lambda_{video} = \lambda_{web}$. Calls from the same service class have the same QCI and ARP values, thus the same priority level. The detailed traffic model settings are given in Table 7.1.

Table 7.1. QoS Requirements of Applications

Services	QCI	ARP-Priority	Delay Budget	MBR (kbps)	GBR (kbps)
Voice	1	1	100 ms	68	68
Video	2	4	150 ms	256	96
Web	8	5	300 ms	128	32

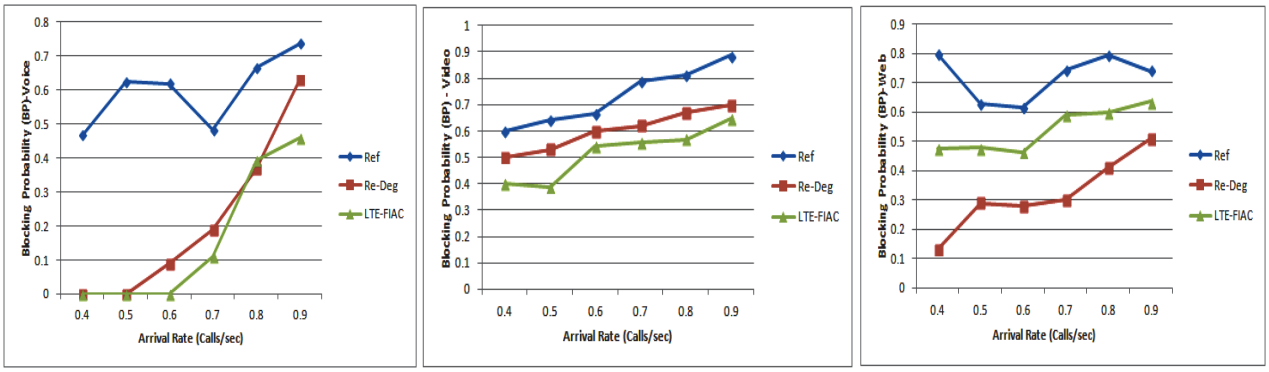
The voice UEs transmit traffic using VoIP G.711. The video service users transmit 256 kbps H.263 video streams. The web traffic sources generate 128 kbps FTP traffic. The total simulation time is 300 seconds.

7.3.2 Simulation Results

This section presents and discusses the results of our simulations to show the performance of the proposed LTE-FIAC. First, we discuss the performance of LTE-FIAC in terms of connection blocking probability. In order to compare the BP of LTE-FIAC to the BP of existing schemes, simulations are performed with the non-congested core network and stable channel conditions.

7.3.2.1 Blocking Probability

Figure 7.6 (a), (b) and (c) show the BP of voice, video and web traffics using Ref, Ref-Deg and LTE-FIAC schemes. They present the connection BP as a function of call arrival rate. From these figures we can conclude that as the connection arrival rate increases, the BP also increases.



(a) (b) (c)

Figure 7.6 Blocking Probability for different service types (a). Voice (b). Video (c). Web

Figure 7.6 (a), (b) and (c) indicate the BP of all traffic types is the highest using the Ref scheme, as it admits connections only when the resources are sufficient. LTE-FIAC and Ref-Deg schemes apply the degradation on lower priorities connections to admit high priority connections. Therefore, the main comparison is between Ref-Deg scheme and LTE-FIAC. Figure 7.6 (a) and (b) show the BP of higher priority services such as voice and video traffic is low using LTE-FIAC scheme compared to Ref-Deg scheme. However, the BP of the lowest priority service, web, is low using the Ref-Deg scheme compared to LTE-FIAC. It is because LTE-FIAC step wise degrades the rate of lower priority connections to exactly what is required and keeps the resources with them to admit incoming high priority connections. As a result, the BP of high priority traffic such as voice and video is the lowest for LTE-FIAC scheme.

The Ref-Deg scheme degrades the rate of lower priority connections directly to their GBR. Consequently, after the degradation completes, it acquires extra resources from the existing connections than requested by a new connection. The extra resources obtained are enough to meet the GBR of narrow bandwidth applications. As a result, the BP of the lowest priority web traffic is the lowest with Ref-Deg scheme.

The affect of using each admission scheme on the fairness among the flows of different and the same priority levels is discussed in the following sections.

7.3.2.2 Fairness among connections of Different priority classes

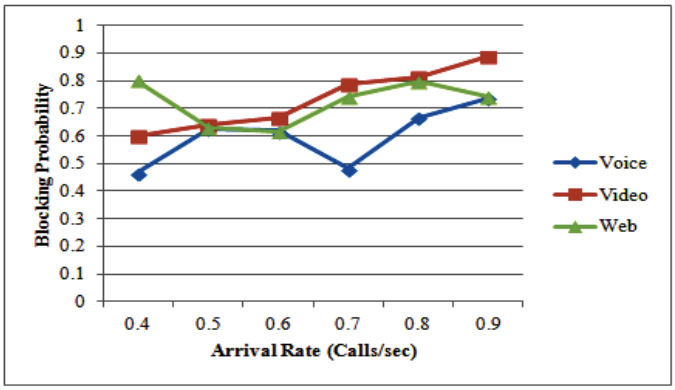


Figure 7.7 Blocking Probability of connections at the eNodeB with Ref Scheme

Figure 7.7 shows the connection blocking probability for the Ref scheme. It shows the BP of voice traffic is less compared to web and video traffic, but the BP of high priority video traffic is higher than low priority web traffic. It is because the bandwidth demand of the video traffic is much higher than the web traffic, so the system accommodates more web connections compared to video traffic. Thus, Ref scheme does not guarantee differentiation among different priority classes.

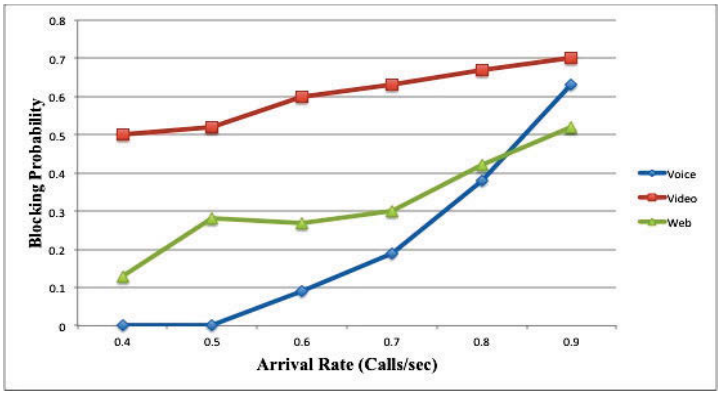


Figure 7.8 Blocking Probability of connections at the eNodeB with Ref –Deg Scheme

Figure 7.8 shows the BP for Ref-Deg scheme. It shows the BP of voice traffic is less compared to the BP of web and video traffic, but as the arrival rate increases the BP of voice traffic becomes higher than the BP of web traffic. Additionally, the BP of high priority video traffic is higher than the BP of low priority web traffic. It is due to the fact that in times of resource limitations, in order to admit a connection of high priority, Ref-Deg scheme one-by-one degrades connections of lower priority to their GBR, until sufficient resources are obtained. So, when the degradation procedure completes, the network acquires extra resources in its available resource pool than requested by an incoming connection. The available resources pool is based on the Complete Sharing (CS) and therefore these additional resources can be assigned to an incoming connection of any service type. As the data rate (GBR) requested by web traffic is less than voice and video traffic, the system with extra available resources accommodates more incoming connections of web compared to voice and video traffic. This situation becomes significant as the arrival rate increases to 0.9 connections per second and the BP of voice becomes higher than web as shown in Figure 7.8. Thus, Ref-Deg scheme does not ensure fair bandwidth allocation among different priority classes.

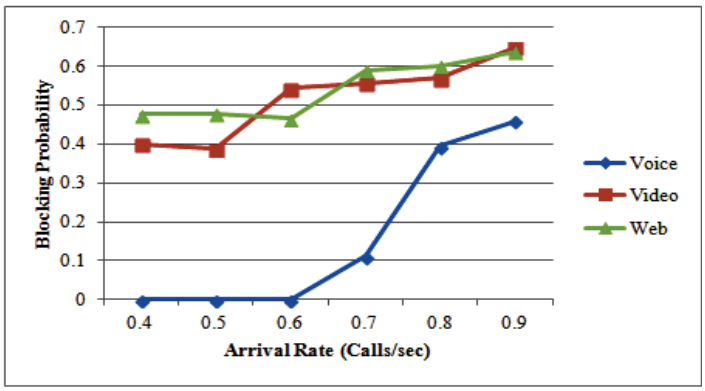


Figure 7.9 Blocking Probability of connections at an eNodeB with LTE-FIAC

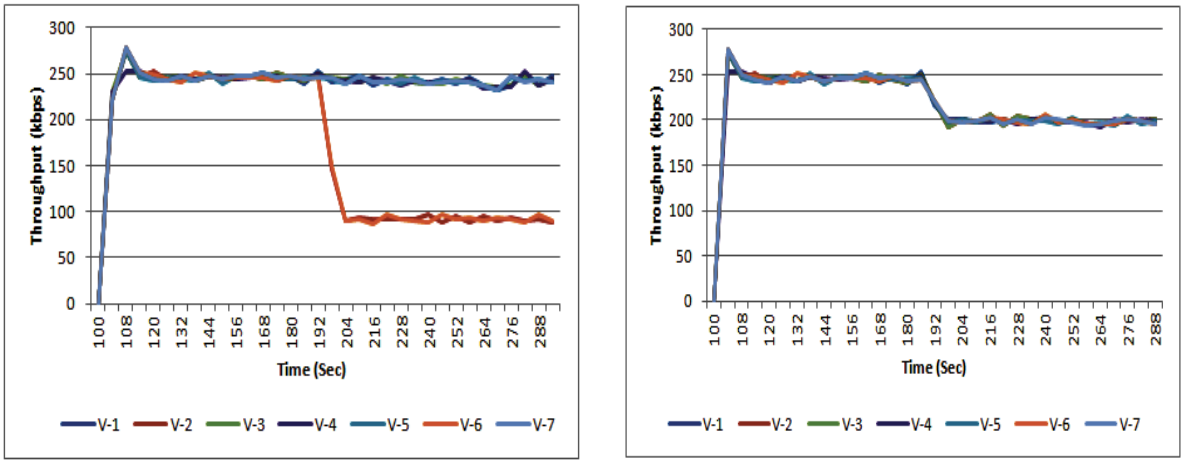
Figure 7.9 shows the BP with LTE-FIAC scheme. With LTE-FIAC scheme the BP of voice traffic is the lowest compared to the BP of video and web traffics. More importantly, the BP of high priority video is overall lower than the BP of low priority web traffic. So, LTE-FIAC ensures service differentiation. It can be contributed to the fact that in order to admit a connection of high priority, LTE-FIAC step wise degrades connections of lower priority until enough resources are obtained. So, when LTE-FIAC performs the degradation, the CS based available

resource pool becomes a little higher than the resources requested by an incoming connection. When high priority connections arrive, LTE-FIAC admits them by degrading the lower priority connections.

In this way, LTE-FIAC scheme by degrading the lower priority connections to only what is exactly required holds the resources for incoming high priority connections and guarantees the service differentiation.

7.3.2.3 Fairness among flows within the same priority class

In times when the resources are inadequate in the network, as discussed earlier, the Ref-Deg scheme degrades a connection with the lowest priority with a step size of $r_{MBR} - r_{GBR}$. When enough resources are obtained, it stops the degradation procedure; otherwise it selects the next connection and degrades it to its GBR. This type of degradation leads to unfairness among the connections at the same priority level.



(a) (b)

Figure 7.10 Throughput (kbps) of video bearers (a) with Ref-Deg scheme (b) With LTE-FIAC

Figure 7.10 (a) shows that at around 204 sec of the simulation, connections with high priority arrive. To obtain sufficient resources, it degrades only video connections v-2 and v-6. So, the throughput of v-2 and v-6 reduces to their GBR. However, the throughput of other video connections, with the same priority and with similar values of MBR and GBR, is still at their MBR. This leads to unfair resource allocation among the connections at the same priority level.

So, Ref-Deg scheme cannot ensure fair bandwidth share among the connections at the same priority level.

LTE-FIAC degrades all connections at the lowest priority with the degradation step size determined using Eq. 7.15. When enough resources are obtained, it stops degrading connections. This type of degradation results in fairness among the connections at the same priority level.

Figure 7.10 (b) indicates that at around 204 sec of the simulation, when voice connections arrive with high priority, LTE-FIAC degrades all video flows equally. So all video connections v-1…v-7, with same priority and with similar values of MBR and GBR, have the throughput at closely the same rate. Thus, LTE-FIAC provides the fair share of bandwidth to connections at the same priority level.

In short, LTE-FIAC achieves lower blocking probability and maintains priority among the connections of different traffic types. Furthermore, LTE-FIAC ensures fairness among the service data flows at the same priority level during the degradation procedure.

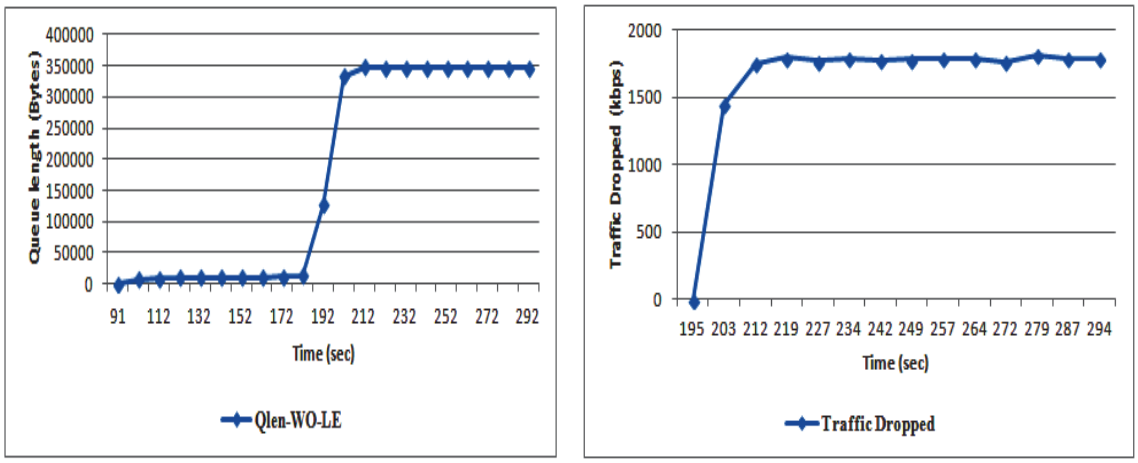
7.3.2.4 Effect of Load Estimation on QoS

To show the effect of the load estimation on the QoS of existing connections, simulation is performed with a congested core network. To depict congestion in the core network, link capacity between the eNodeB and the EPC is set at 1.9 Mbps. LTE-FIAC and LTE-FICC schemes are employed at the eNodeB as the RAC and the CC schemes, respectively. The arrival rate of high priority voice is increased to 0.9 connections per second after 150 seconds of the simulation. To show the effect of load estimation, the connection departure is set to be the end of the simulation. For comparison, simulations are performed without the load estimation.

Queue length Variation

Figure 7.11 (a) shows the scenario when connections are admitted by LTE-FIAC without the load estimation (WO-LE). In this scenario, LTE-FIAC only considers the resource availability and the QoS requirements of an incoming connection. It does not take into account the EAR_i estimated in Eq. 7.10.

In Figure 7.11 (a), initially the queue length is around the target operating point. This is because to manage the load at the core, LTE-FICC degrades the extra bandwidth allocated to non-GBR, and GBR connections above their GBR requirements. After 150 second of the simulation, LTE-FIAC admits high priority voice connections without the load estimation and by applying degradation on low priority video and web connections.



(a) (b)

Figure 7.11(a) Queue length (Bytes) (b) Traffic Dropped (kbps), without Load Estimation

In the simulation, VoIP connections have the same MBR and GBR (Table 7.1). So, during congestion, LTE-FICC cannot degrade the bandwidth allocated to VoIP connections. As a result, the queue length at the eNodeB output buffer starts increasing and reaches the maximum buffer capacity. Consequently, packets drop starts as shown in Figure 7.11 (b).

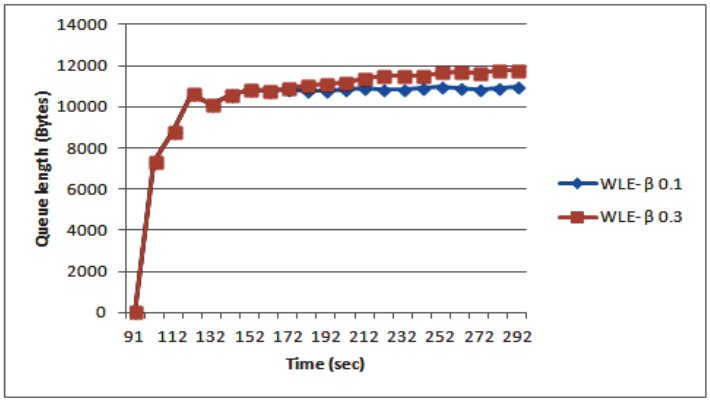


Figure 7.12 Queue length (Bytes) with Load Estimation

Figure 7.12 shows the scenario when LTE-FIAC is applied with the load estimation (WLE). LTE-FIAC with the load estimation admits even a high priority connection only if it estimates that the rate requested by a connection is less than or equal to the rate expected by the network (EAR_i). It facilitates LTE-FICC to maintain the queue length around the target operating point as clearly illustrated in Figure 7.12. So, LTE-FIAC by intelligently admitting new connections resolves the output buffer overloading problem.

The value of parameter β largely impacts the period $MACR_i$ takes to reach the current class allowed rate. In turn, it widely affects the estimation of EAR_i and so the admission rate. When the parameter β is set to a small value such as 0.1, it takes longer for the MACR to converge to the current allowed rate. As a result, the load estimation becomes more conservative (less responsive to slight reduction in the load at the buffer) and admits less number of connections in the network compared to β equal to 0.3. Consequently, the queue length for β equal to 0.1 changes smoothly and closely around the target operating point compared to β equal to 0.3 as indicated in Figure 7.12.

Queuing Delay Variation

Figure 7.13 and Figure 7.14 show the effect of the load estimation on the average queuing delay at an eNodeB output buffer.

Without the load estimation, the queue length reaches the maximum buffer capacity (Figure 7.11 (a)). As a result the queuing delay becomes very high as demonstrated in Figure 7.13.

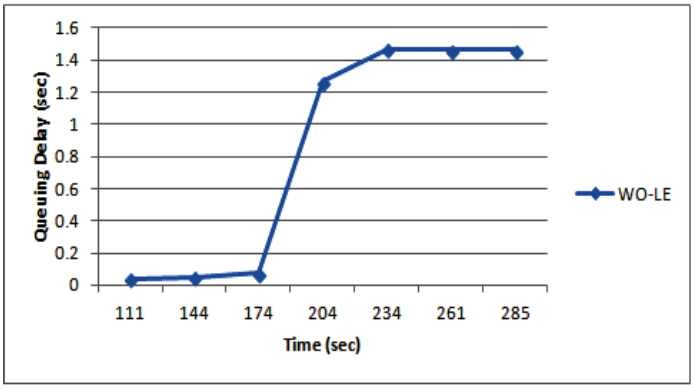


Figure 7.13 Average Queuing Delay (sec) without Load Estimation

LTE-FIAC with the load estimation admits a connection only if its requested rate matches to the rate offered by the network. Hence, it enables LTE-FICC to maintain the queue length around the target point (Figure 7.12). Consequently, the average queuing delay is very low as illustrated in Figure 7.14.

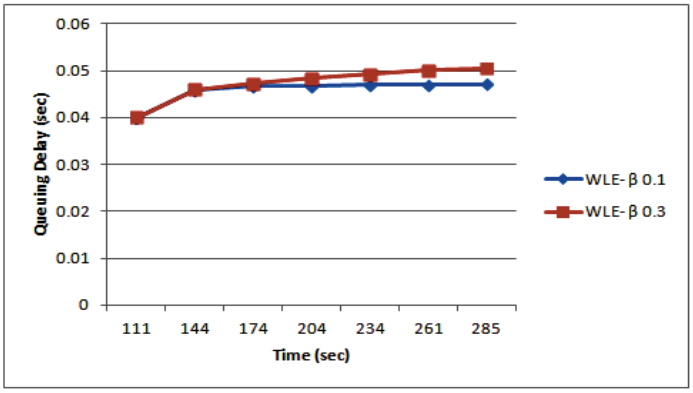


Figure 7.14 Average Queuing Delay (sec) with Load Estimation

With β equal to 0.1, the queue length fluctuates closely around the target point (Figure 7.12). As a result, the average queuing delay for β equal to 0.1 is relatively less than the β equal to 0.3 (Figure 7.14).

Throughput

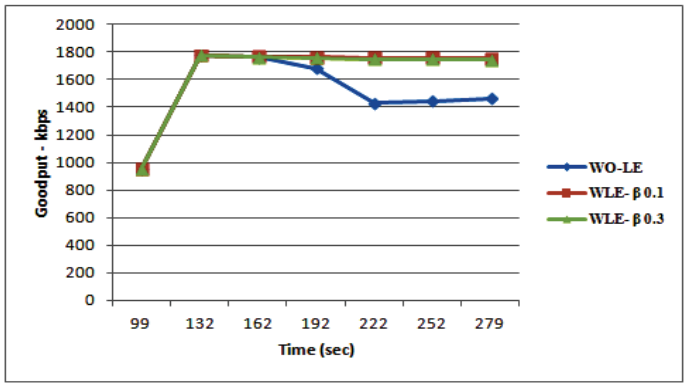


Figure 7.15 Average Throughput (kbps) of the network with and without Load Estimation

Figure 7.15 shows the throughput of the network with and without the load estimation. It shows before 150 sec of the simulation, with and without the load estimation, the average

throughput of all connections in the network is the same. After 150 sec, the throughput of LTE-FIAC without the load estimation becomes lower than the throughput with the load estimation. This is because, without the load estimation LTE-FIAC admits connections without knowing the current congestion state in the core network. Consequently, the queue length reaches the maximum buffer capacity and traffic drops (Figure 7.11 (b)) and results in reduced throughput of the network.

LTE-FIAC with the load estimation admits a connection only if it estimates that incoming request will not overload the buffer. As a result, the network achieves stable throughput without any loss at the buffer as shown in Figure 7.15. In Figure 7.15, the overall throughput for β equal to 0.3 and 0.1 is the same, which is explained with Figure 7.16.

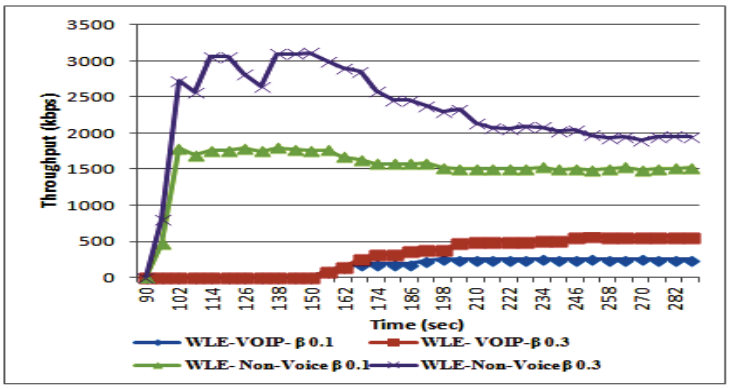


Figure 7.16 Average Throughput (kbps) of Voice and non-Voice traffics with Load Estimation

In order to accommodate incoming connections of high priority voice traffic, LTE-FIAC degrades resources allocated to connections of low priority services. Figure 7.16 clearly indicates that after 150 sec of the simulation, the throughputs of low priority services such as video and web with β equal to 0.3 reduce greatly compared to the throughputs with β equal to 0.1. It is because the admission rate of voice connections is high for β equal to 0.3 compare12d to β equal to 0.1 (Figure 7.12). However, the cumulative throughput of all voice and non-voice connections is the same for β equal to 0.1 and 0.3 (Figure 7.15).

In short, when LTE-FIAC admits incoming connections based on the load estimation in the network, it facilitates LTE-FICC to maintain the queue length at the eNodeB output buffer

around the target operating point. LTE-FIAC with the load estimation achieves smaller average queuing delay and stable throughput in the network.

7.3.2.5 Effect of Upgradation and Extra Resource Reservation on Blocking Probability and QoS degradation Probability

In times when channel condition degrades, the demand of PRBs to guarantee GBR of connections increases. When RAC does not take the increased demand of PRBs of existing connections into account and continues admitting connections based on the resource availability, the QoS of ongoing sessions reduces. The RAC scheme by (Mehdi. et al., 2012) reserves additional resources with an incoming connection to offset the changes in the user demand. The scheme estimates the extra-required resources as an average of the extra used resources at an end of each mobility epoch. In LTE-FIAC, extra resource reservation scheme determines the amount of extra resources required (Δ) from the past call history. Furthermore, before reserving additional resources for an incoming connection, it considers the resources that are already allocated to existing sessions above their GBR requirements.

To show the effect of the proposed extra resource reservation on the BP and the QoS of existing connections, we run the simulation in a scenario when the channel conditions significantly degrade. For comparison, the simulation is run without the proposed extra resource reservation and without the connection departure procedure.

Blocking Probability

Figure 7.17 shows the blocking probability of new calls without the extra resource reservation (WoRR) and with the extra resource reservation (WRR) schemes. The CDP executes upgradation procedure, which allocates free resources to existing connections above their GBR requirements. As discussed earlier, the scheduler can allocate the resources, above the GBR requirements of existing connections, to any connection to gain its GBR. So, when the simulation is run without the extra resource reservation (WoRR), the CDP is not executed to show the accurate affect of the channel degradation on the QoS of ongoing sessions.

Figure 7.17 clearly illustrates that WoRR and without CDP (WoCDP), the BP of different priorities is not in the order of precedence stated by their QCIs and ARP. The BP of high priority video traffic is quite high compared to the BP of low priority web traffic. This is because when the connection arrival rate increases, more degradation is applied to existing connections of low priority traffic to admit high priority connections. Whereas, the upgradation is not applied to increase the rate of existing connections and also the channel conditions do not support high MCS for an incoming connection. As a result, when a connection departs the resources added back in the CS based available resource pool are mostly enough to admit connections with low data rate requirements. Consequently, low priority web connections are admitted more compared to high priority video traffic.

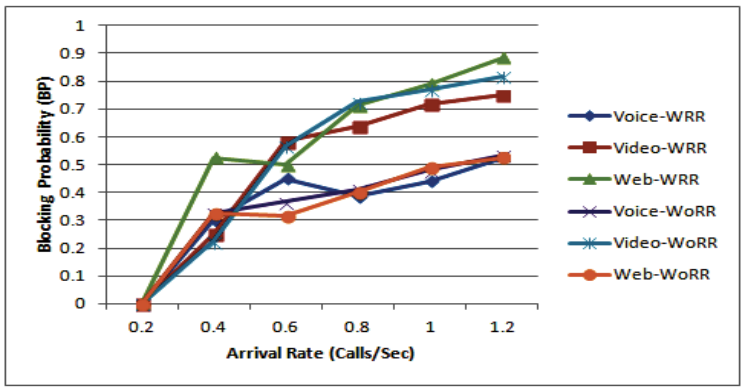


Figure 7.17 Blocking Probability of new calls with and without ERRM and CDP

Figure 7.17 indicates WRR and with CDP (WCDP), the BP of different priorities is in the order of precedence. The BP of the highest priority voice is the lowest. Additionally, the BP of video is lower compared to the BP of web traffic although video is a high bandwidth demanding application. It is because the upgradation procedure allocates the resources left by departing connections to existing connections to enable them to gain their MBR. As the degradation is applied only for high priority connections, so it admits more connections of voice and video traffic. Consequently, it ensures differentiation among the connections at different priority levels is maintained.

Figure 7.17 shows with the resource reservation, the BP of the lowest priority web is very high compared to the BP of web without the resource reservation. However, overall with the resource

reservation and WCDP increase in the BP of high priority traffics is negligible. Rather, for the arrival rates until 0.6 calls per second, the BPs of voice and video with the resource reservation and WCDP are overall equal to the BPs of voice and video traffic without the resource reservation and WoCDP. Moreover, for the call arrival rate above 0.8 calls per second, the BP of high priority voice and video with the resource reservation and WCDP even becomes less than the BP of voice and video traffic without the resource reservation and WoCDP.

QoS Degradation Probability (QDP)

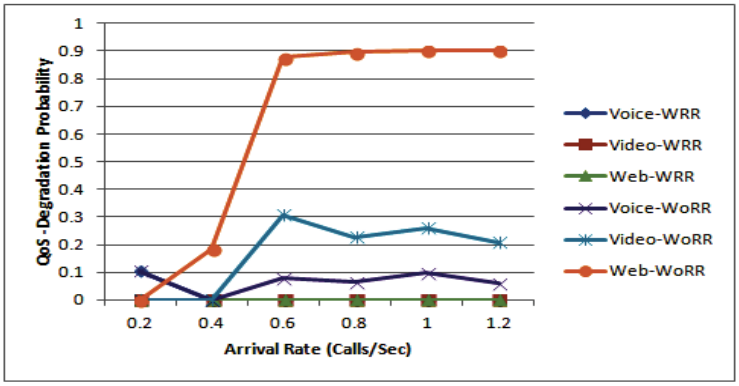


Figure 7.18 QoS Degradation Probability of ongoing calls with and without ERRM and CDP

Figure 7.18 shows the QoS degradation probability (QDP) of ongoing sessions. The QDP indicates probability that the network cannot provide the minimum guaranteed rate to connections of a specific service type during the call holding time (Mehdi. et al., 2012). It shows WoRR, the QDP of web connections is very high and is close to 1. However, the QDP of video and voice traffic is relatively low compared to web traffic. This is because the PF scheduler prioritizes traffic according to their QCIs using the scheduling weights. It allocates resources first to connections of voice traffic, video and then web traffic to gain their GBR. As a result, the effect of channel fluctuations on the demand of PRBs is automatically moved to the lowest priority web connections and thus their QDP is very high.

Figure 7.18 clearly demonstrates that WRR, the QDP of all traffic types is even lower than the QDP of the highest priority voice traffic WORR. It is due to the fact that LTE-FIAC with the extra resources reservation, reserves additional resources at the time of connection arrival in the form of extra resources allocated to existing connections using the upgradation, or sets aside

them in the available resource pool to enable the network to cope with channel fluctuations. With the extra resource reservation, the QDP of all traffic types is around zero, which indicates that the minimum QoS to all connections is guaranteed. It can be attributed to the fact that the amount of extra resource required (Δ) is determined as a moving average of the change in the demand of PRBs. Also, as the connection departure procedure is applied with the resource reservation, so before admitting any new connection in the network it updates the gap (Ω). It allocates resources to existing connections or sets aside them in the available resource pool to cover the gap. This enables the network to update the reserve of the extra resources to match the actual channel conditions.

With the extra resource reservation, there is a clear tradeoff between the BP and the QDP. With the CDP and WRR, the network is able to: ensure fairness among the connections at different priority levels; and provides the fair share of bandwidth to existing connections to guarantee their QoS.

7.4 Summary

In this chapter we presented LTE-FIAC, an intelligent admission control scheme. By using different network scenarios, we evaluated the performance of FIAC scheme. The results showed that LTE-FIAC achieves lower BP and guarantees the fair share of bandwidth among the service flows at the same as well as different priority levels. When the core network is congested, LTE-FIAC with the load estimation module matches the network capability with the QoS requirements of an incoming connection to avoid buffer overloading at an eNodeB. Results show that in times the core network is congested or when the channel condition degrades, there is a trade-off between the BP of new connections and the QoS of existing connections. LTE-FIAC with the load estimation and the extra resource reservation preserves the QoS of existing connections. So, LTE-FIAC is well suited to the networks that aim to ensure the QoS of existing connections in terms of stable throughput and reduced delay.

Chapter 8 Impact of QoS Schemes on Capacity and Coverage Analysis

The network dimensioning activities include coverage analysis and capacity estimation. These tasks allow the network operators to determine the overall site configurations in terms of total number of base stations and the capacity of the core network required to guarantee the data rate demands of the area of interest. It facilitates to determine the point to which the network can cope with the dynamic fluctuations in application demand; population distributions; and load at the core network, without the need to re-dimension while also maintaining the QoS of connections. In doing so, this chapter suggests a novel and efficient way to the network operators for estimating the total number of users, which may be supported under dynamically varying network load and demand parameters by employing our proposed QoS schemes at a base station.

This chapter focuses on several aspects, which did not receive much attention for WiMAX and LTE dimensioning. Section 8.1 discusses the factors, which impact the coverage and capacity of a wireless network. Section 8.2 provides coverage analysis for both 4G technologies, WiMAX and LTE. Section 8.3 describes the parameters involve in the capacity analysis of both 4G technologies. Section 8.4 comprehensively discusses the effect of various factors, including bandwidth and applications distribution, on the resource utilization and the capacity of the 4G networks. It also discusses the impact of compressed overhead employed in WiMAX networks on the capacity and the coverage of the WiMAX network. Furthermore, it provides the capacity of a cell in terms of the users supported with the proposed QoS schemes including the Congestion Control (CC) and the Radio Admission Control (RAC). Finally, Section 8.5 summarizes the chapter.

8.1 Factors effecting Coverage and Capacity

This section discusses the factors, which impact coverage and capacity offered by a wireless network.

Coverage (CVG) of a cell is a function of transmit power (P_{Tx}), transmitter antenna gain (G_{Tx}), receiver antenna gain (G_{Rx}), Signal-to-Noise-Ratio (SNR), frequency (f), base station antenna height (h_b), subscriber station antenna height (h_s), bandwidth (BW), and overhead repetition factor (R) as shown in Eq. 8.1.

$$\text{CVG} = f\ (\ P_{Tx}, G_{Tx}, G_{Rx}\ , \text{SNR}, \text{f}, h_b, h_s, \text{BW}, \text{R}) \tag{8.1}$$

The effect of the height of a base station and a mobile station and also the effect of the SNR are discussed by (Afric et al., 2006). This chapter focuses on the effect of the frequency and the bandwidth on the coverage of the network.

In urban, suburban and rural environments of India, measurements were taken by (Sharma and Singh, 2010) at 900 MHz and 1800 MHz frequencies. For suburban areas, results of SUI and Cost-231 models were close to the measurements. The Hata model and Cost-231 gave better results for rural areas. So, in this analysis to calculate the cell radius, Terrain type 'C' of SUI-model and 'flat environment 'of COST-231 Hata model (231, 1999) are used for the rural areas with the frequency of 2300 MHz. The Hata model (Hata, 1980) is used to calculate the cell range for the rural areas with the frequency of 900 MHz.

Path loss (PL) is determined using the following equation.

$$PL = (P_{Tx} + G_{Tx} + G_{Rx} - L_{others} - R_{ss}) \tag{8.2}$$

In Eq. 8.2, L_{others} refers to other losses in the network and includes Inference Margin (IM), body and indoor losses. The R_{ss} is the Receiver Sensitivity and refers to the minimum received power required at a receiver to guarantee the bit error rate (BER) of 10^{-6}. For OFDMA PHY, it is determined using the following equation.

$$R_{ss} = -174 + 10log_{10}(\Delta f) + NF + SNR_{Rx} \tag{8.3}$$

In Eq. 8.3, SNR determines the Modulation and Coding Scheme (MCS) that can be used by a subscriber station. The Δf refers to the system bandwidth in Hz and NF refers to the Noise Figure at the receiver.

Effective Capacity (CPC) of a cell indicates the total number of users that can be served with the available network resources. It is a function of frequency (f), bandwidth (BW), frame duration (T_f), SNR, modulation and code rate, cyclic prefix (CP), overhead (OH) and Application demand Distribution (AD) as shown in Eq. 8.4.

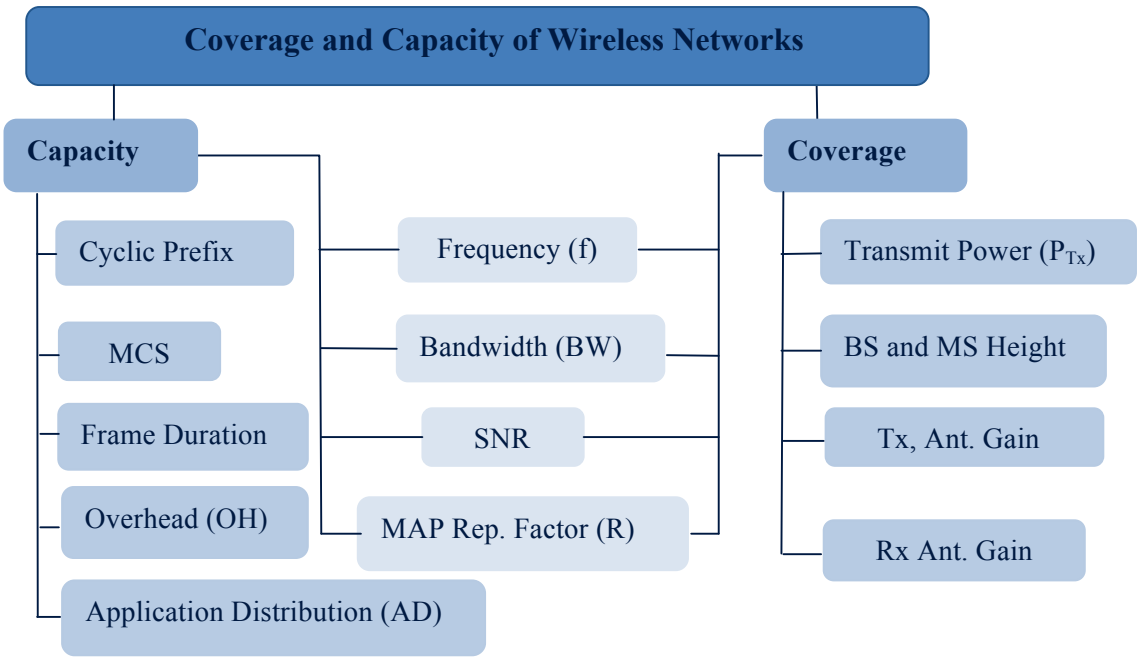


Figure 8.1 Factors Contributing to Coverage and Capacity of Wireless Networks

$$\text{CPC} = f\,(\text{f}, \text{BW}, \text{T}_\text{f}, \text{SNR}, \text{MCS}, \text{CP}, \text{OH}, \text{AD}) \qquad 8.4$$

Almost every research on the capacity of the system discussed the effect of cyclic prefix and code rate. The lower the values of cyclic prefix and code rate, higher is the capacity, as less redundant data is sent. Similarly, the use of high capacity modulations give high capacity as shown by (WiMAX Forum, 2006). This chapter discusses the effect of frequency, bandwidth, application distribution and overhead on the capacity of the network.

8.2 Coverage Analysis

This section provides a comprehensive coverage analysis of the both 4G technologies, WiMAX and LTE.

8.2.1 Coverage Analysis of Mobile WiMAX

This section discusses the coverage analysis of the WiMAX networks. By applying the typical values of a base station antenna height of 35m and a mobile station antenna height of 2m, the closed form to determine the path loss for the rural area using Hata, Cost-231 Hata and (Erceg et al., 1999, V.Erceg, January 2001) SUI-C models are given in equations 8.5, 8.6 and 8.7, respectively.

$$PL_{Hata} = 98.9 + 34.78log_{10}(d) \quad 8.5$$

$$PL_{Cost231-Hata} = 137.4 + 34.78\, log_{10}(d) \quad 8.6$$

$$PL_{SUI-C} = 89.6 + 39.96\, log_{10}\left({}^{d}/_{d_0}\right) \quad 8.7$$

Where 'd' is the distance between a base station and a Mobile Station (MS). The PL refers to the path loss and is determined using Eq. 8.2. In the current analysis, the WiMAX BS uses a directed antenna with a typical transmit power of 35 dBm and an antenna gain (G_{Tx}) of 18 dBi. The receiver antenna gain of the MS (G_{Rx}) is assigned 0dBi (Andrews et al., 2007). The transmit power (P_{Tx}) of the MS in UL is assigned a typical value of 23 dBm. In the current analysis, NF and L_{others} are assumed to be 8 dB and 5 dB, respectively. In the coverage analysis of Mobile WiMAX, we use the SNR values given by (Ahmadzadeh, 2008).

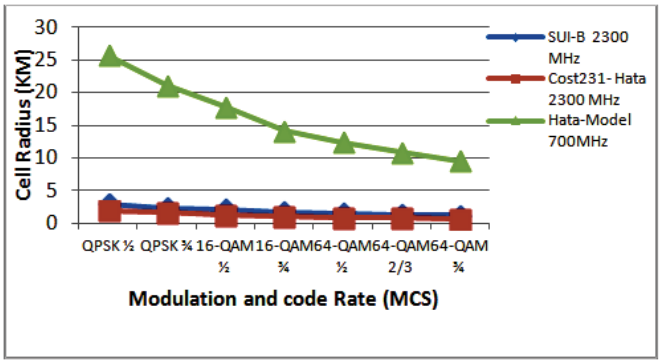


Figure 8.2 Cell Radius of Rural Area with Various frequencies for WiMAX networks

Figure 8.2 illustrates the cell radius of the rural area for different modulation and code rates. It provides the cell radius for 700 MHz and 2300 MHz frequencies with 5MHz bandwidth. It clearly indicates that the cell radius with 700 MHz frequency is very large compared to the cell

radius with 2300 MHz. Hence, lower frequencies give better cell coverage compared to higher frequencies. Figure 8.2 also illustrates that the cell radius reduces for higher capacity MCS. Consequently, there is a clear tradeoff between the capacity of a cell and its coverage.

8.2.2 Coverage Analysis of LTE Networks

This section discusses the coverage analysis of the LTE networks. The NBN Australia has selected LTE to deploy its fixed broadband wireless network. So, to perform the coverage analysis of LTE, the parameters are selected to match the scenarios and values considered by NBN (Amrish Kacker et al., 2012). The NBN Australia selected 2.3 GHz and 3.4 GHz frequencies to deploy in the Broadband Wireless Access networks (BWA). So, this analysis considers these two frequencies for LTE networks. By applying an eNodeB antenna height of 40m and a Customer Premises Equipment (CPE) antenna height of 6m (Amrish Kacker et al., 2012), the closed form to determine the path loss for the urban, the suburban and the rural area using Cost-231 Hata model are given in equations 8.8, 8.9 and 8.10, respectively.

$$PL_{Cost231-Hata_{urban}} = 127.7 + 34.78\, log_{10}(d) \quad 8.8$$

$$PL_{Cost231-Hata_{sub}} = 114.6 + 34.78\, log_{10}(d) \quad 8.9$$

$$PL_{Cost231-Hata_{rural}} = 103.5 + 34.78\, log_{10}(d) \quad 8.10$$

In the current analysis, the P_{Tx} of the eNodeB for DL is set at 46 dBm. For UL, transmit power of the CPE is set at 23 dBm (Amrish Kacker et al., 2012). The eNodeB deploys 3 sector antennas, so its antenna gain is set at 15 dBi including feeder loss. The antenna gain of the CPE is defined as 10 dBi as it is a fixed wireless terminal and uses an outdoor directed antenna (Harri Holma and Antti Toskala, 2009). The uplink is orthogonal, so there is no intra cell interference. The NBN frequency plan is to re-use three frequencies across the network. Hence, each of the three sectors of the eNodeB deploys a unique channel and there is no co-channel interference in the cell (Amrish Kacker et al., 2012). Consequently, an IM is set to 0 dB. The antenna is fixed outdoor so body and indoor losses are also set to 0 dB. As a result, for the LTE coverage analysis, the value of L_{others} is 0 dB in Eq. 8.2.

The Multiple-Input and Multiple-Output (MIMO) antenna diversity reduces the effect of multipath fading on the SNR of the received signal (Mohammad T. Kawser et al., 2012). For the coverage analysis, the downlink transmit diversity is set at 2x2 and UL transmit diversity is set at 1x2. In the coverage analysis of LTE we utilize the SNR values determined by (Mohammad T. Kawser et al., 2012) for the transmit diversity of 2x2 corresponding to 10% BLER. The NF is set at 9 dB and 5 dB for DL and UL, respectively.

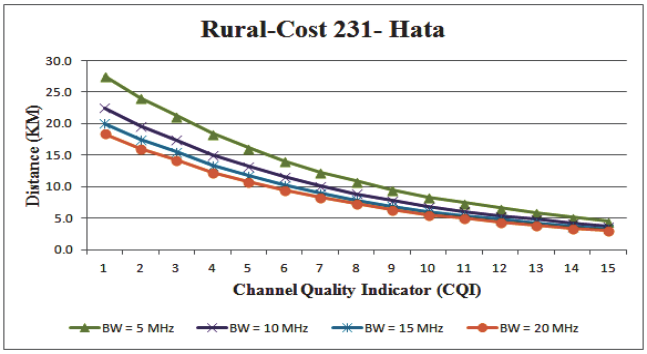


Figure 8.3 Cell Radius of Rural Area with various bandwidths for LTE networks

Figure 8.3 shows the cell radius of the rural area for different values of CQIs. It provides the cell radius for 2.3 GHz frequency with 5, 10, 15 and 20 MHz bandwidths, respectively. Figure 8.3 indicates that for any specific CQI when the bandwidth is doubled, the cell radius reduces by around 0.2 percent. Hence, lower bandwidth gives better cell coverage compared to higher bandwidth due to the receiver sensitivity. The component, $10.\log10(\Delta f)$, of the receiver sensitivity in Eq. 8.3 encompasses all subcarriers in the bandwidth. As a result, when the bandwidth increases, the minimum power required at the receiver (R_{ss}) also increases and results in reduced cell radius.

The above estimations demonstrate that the cell radius of the suburban and the urban areas is 52% and 80% smaller compared to the cell radius of the rural area given in Figure 8.2 and Figure 8.3. They also indicate that for 3.4 GHz frequency the cell radius is 32 % smaller compared to the cell radius of 2.3 GHz. So as discussed earlier, the lower the frequency the higher is the cell radius. Figure 8.2 and Figure 8.3 indicate that the cell radius reduces for higher values of MCS and CQIs, respectively. It is because higher SNRs are required for higher capacity modulations. When the users are close to their base stations and have higher SNR, they can avail higher

capacity modulation schemes, and as they move away from the base station their respective SNR reduces leading to usage of low capacity modulations.

Normally, the site configurations derived from the coverage analysis remain the same until or unless the coverage area converges from rural to suburban or suburban to urban. So, our analysis afterwards is restricted to the factors that affect the supported capacity of the network.

8.3 Parameters of Capacity Analysis

This section discusses the capacity analysis of the 4G technologies, WiMAX and LTE. It provides the effect of various factors such as bandwidth, frequency, application distribution, overheads and QoS schemes on the capacity of the network.

The current capacity analysis considers following applications.

VoIP: It is used to make voice calls over the packet switched networks. The VoIP source generates fixed size packets periodically. The packet size and the inter-arrival time depend on the encoder scheme in use.

Streaming: It is used for video conferencing and videos upload or download. The streaming applications generate variable size packets. Its data rate depends on the quality and size of the display.

Online Gaming: It allows players to play games online over the Internet.

File Transfer: The FTP session is used to download or upload a file. Its main parameters include file size and inter-request time.

Web: It specifies the Hyper Text Transfer Protocol (HTTP) session between a client and a server. Its main parameters include page properties and inter-arrival time. The page properties include size of the main page, number of objects per page, and the size of each object. In the current analysis, web browser uses HTTP 1.1 protocol.

To appropriately define the applications demand distribution, users in a cell are grouped as follows.

Households: The household users normally use VoIP, streaming to upload or download videos, online gaming and web services.

Public and Health Services: The public and health services offices usually employ web and VoIP services.

Business: The businesses at any level mostly utilize applications of web, VoIP, streaming for video conferencing and file transfer.

8.3.1 Parameters for Capacity Analysis of Mobile WiMAX

To incorporate the impact of varying channel conditions on the ability to deploy a particular MCS, an average Slot Size (SS) is defined using the following equation (So-In. et al., 2010).

$$SS = \sum P_{QPSK-1/2} \times S_{QPSK-1/2} + P_{QPSK-3/4} \times S_{QPSK-3/4} + P_{16QAM-1/2} \times S_{16QAM-1/2} + P_{16QAM-3/4} \times S_{16QAM-3/4} + P_{64QAM-1/2} \times S_{64QAM-1/2} + P_{64QAM-2/3} \times S_{64QAM-2/3} + P_{64QAM-3/4} \times S_{64QAM-3/4} \quad 8.11$$

In Eq. 8.11, 'P' is the probability to use a particular MCS and 'S' is the slot size with a specific modulation. In this analysis, the probability distribution of different MCS for 700MHz and 2500 MHz frequencies given by (MWG/AWG, 2008) is used. Table 8.1 gives the detail probability of each MCS.

Table 8.1. Probability of MCS at 2300 MHz and 700 MHz- WiMAX networks

MCS	**Probability Distribution of MCS**	
	2300 MHz	**700 MHz**
QPSK ½	0.51	0.24
QPSK ¾	0.16	0.05
16-QAM ½	0.08	0.04
16-QAM ¾	0.05	0.06
64-QAM ½	0.03	0.08
64-QAM 2/3	0.05	0.15
64-QAM ¾	0.12	0.38

To apply the compressed MAPS as discussed in 0, the users in the region of 64QAM-3/4 to 16QAM-1/2 MCSs are grouped in group #1. All the users that are in the region of QPSK-3/4 MCS are grouped in group #2. The group #3 includes all the users that are in the region of QPSK-1/2 MCS as shown in Figure 8.4.

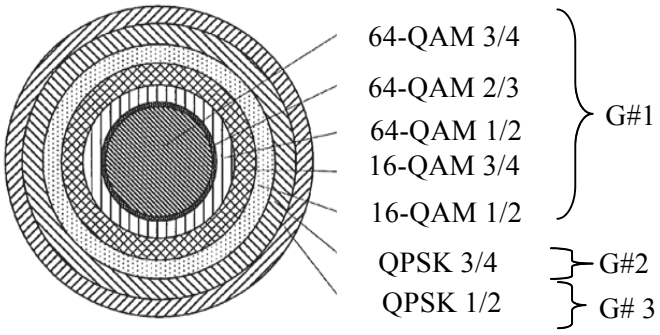


Figure 8.4 Groups of different Modulation and Coding Scheme (MCS)

As suggested by (Sanker et al, Oct. 15 2009), the users in group #1 are considered to have good channel conditions so they can receive the MAP message modulated using QPSK-1/2 and with a repetition of 1. For the users in group #2, the MAP message can be transmitted using QPSK-1/2 and a repetition of 2. However, the users in group#1 are considered to be in bad channel conditions so they can receive the MAP message modulated with QPSK-1/2 and with a high repetition of 4. So, in this analysis, to transmit the MAP message we use the same MCS (QPSK-1/2) for all groups but use a varied repetition factor for various groups to reduce the MAP overhead. Table 8.2 shows groups that are defined in Figure 8.4 with their corresponding probabilities derived from Table 8.1.

Table 8.2. Probability of Each Group of MCS for MAP Transmission

G#	**MCS**	**Probability of MCS**		**MCS for MAP**
		700 MHz	**2300 MHz**	
1	64QAM-3/4… 16QAM-1/2	0.71	0.33	QPSK ½ , (R: 1)
2	QPSK 3/4	0.05	0.16	QPSK ½ , (R: 2)
3	QPSK 1/2	0.24	0.51	QPSK ½ , (R:4)

To analyze the capacity of WiMAX networks, the two application distributions AD-1 and AD-2 are defined in Table 8.3. For VoIP, wireless codec standard GSM-EFR is chosen. It operates at the data rate of 12.2 Kbps with a sampling frequency of 8 KHz. For mobile TV, video codec format H.264 for a resolution of 640×480 and with an average data rate of 1215 kbps (Seung-Eun and Woo-Yong, 2010), is selected. The data rate for web traffic and online gaming are set to 15 Kbps and 50 Kbps, respectively (Ahmadzadeh, 2008).

Table 8.3. Application Distributions- WiMAX Networks

Application	**Data Rate (Kbps)**	**(AD-1)**	**(AD-2)**
Mobile TV	1215	10%	20%
VoIP	12.2	30%	20%
Web	15	35%	35%
Online Gaming	50	25%	25%

For the WiMAX capacity estimation, the parameters including frame duration, cyclic prefix, DL to UL ratio and bandwidth are set at 5ms, 1/8, 2:1 and 10 MHz, respectively.

8.3.2 Parameters for Capacity Analysis of LTE Networks

Similar to the WiMAX networks, to incorporate the fact that varying channel conditions impact the deployment of a particular MCS, the average cell throughput of LTE networks is defined using the following equation (Abdul Basit Syed, 2009).

$$Avg\_cell\_TH = \sum_{i=0}^{15} P_{SNR_i} * TH_{SNR_i} \qquad 8.12$$

The P_{SNR_i} is the probability of occurrence of an SNR value at a cell edge corresponding the CQI 'i'. The TH_{SNR_i} is the throughput of a cell with a specific MCS corresponding the CQI 'i'. The throughput of a cell also depends on the bandwidth of the system. To perform the capacity analysis of LTE networks, we assumed a fixed network and the uniform probability distribution for different MCS at 2.3GHz frequency.

For LTE capacity analysis, the VoIP traffic is generated using G.711 encoder scheme with a voice payload size of 160 bytes and an interval of 20ms. To generate the video traffic we use a frame size of 128 X 120 pixels. The frame inter-arrival time is set to 160 ms. In the current analysis, the data rate of online gaming is set to 50 kbps. For the FTP session, the file size is set at 1 Mbps and the inter-request time is 4.0 second. The web page properties and the main attributes of the web traffic are given in Table 8.4 (So-In. et al., 2010).

Table 8.4. Parameters of Web Traffic

Parameters	**Values**
Main Page Size (Bytes)	10710
Embedded Object Size (Bytes)	Uniform (4000-8000)
Number of Embedded objects	10
Reading Time (sec)	10
Request size (Bytes)	350

Keeping in view the demand of each user-group discussed in section 8.1, two application distributions given in Table 8.5 are defined for the LTE capacity analysis.

Table 8.5. Application Distributions- LTE Networks

Application	**QCI**	**MBR (Kbps)**	**GBR (Kbps)**	**(AD-1)**	**(AD-2)**
Streaming	2	850	850	20%	40%
VoIP	1	64	64	30%	10%
Online Gaming	3	50	50	20%	20%
Web	4	32	15	20%	20%
FTP	4	256	50	10%	10%

For the LTE capacity estimation, normal CP is employed. For simplicity of the analysis, capacity estimation is performed for the eNodeB with a 1x1 antenna and 20 MHz channel. It does not consider the effect of H-ARQ.

The scheduler we considered in the LTE capacity analysis serves connections based on their delay requirements (OPNET, 2012). Instead of scheduling each service flow in every subframe, the scheduler holds its data for certain number of subframes hereafter referred to as "Interval". The Interval is determined based on the following equation.

$$Interval = \frac{delay\_budget}{subframe\_dur} \quad 8.13$$

In Eq. 8.13, Subframe_dur is the duration of subframe. The delay_budget is an acceptable delay of the QoS class to which the service flow belongs. The services are mapped to QoS classes according to the mapping given in Table 8.5.

Table 8.6. Protocol Overhead with Proportional Fair Scheduler and 20 MHz bandwidth- LTE Networks

	Streaming	**VoIP**	**Online Gaming**	**FTP**	**Web**
User-Data at IP layer (Bytes)	1377	640	250	1377	960
Data+ Headers +CRC (Bytes)	1394	659	269	1394	979
Served_sub	77	13	25	23	4
Total OH (kbps)	10.0	1.9	3.8	3.5	0.41
Total rate (kbps)	860.5	65.9	53.8	259.1	32.6

Based on the Interval estimated in Eq. 8.13, Table 8.6 provides the required data rate of each application to meet its data and protocol overhead requirements with 20 MHz bandwidth. For example, the streaming application has data requirements of 850 kbps. It's QCI is 2 and hence its delay budget is 150 ms. The scheduler estimates to serve the service flow in every 120th subframe, based on the 80% of the interval estimated in Eq. 8.13 (OPNET, 2012). We took 80% of the interval to allow a margin for the delay at the core network. In the current scenario, the streaming generates a frame every 30 ms. The scheduler, therefore, estimates that in one LTE subframe it can send four video frames based on its delay requirements. This allows the scheduler to reduce the protocol overhead. However, the maximum bits that can be served in a subframe depend on

the MCS of a CPE and the bandwidth of the system. For example, the 20MHz bandwidth and the 64QAM-3/5 MCS allows the maximum transmission of 3542 bytes per subframe (3GPP 36.213). Moreover, in LTE networks, GPRS Tunneling Protocol (GTP) tunnels carry bearers in an all IP based core network. The Maximum Transmission Unit (MTU) of an IP core is 1500 bytes. The GTP/UDP/IP and IPsec tunneling also add overheads to the packet. To enable a packet to fit to an Ethernet interface's Service Data Unit (SDU) of the core network, one way is to set the MTU at an IP layer of a CPE to be less than or equal to 1394 bytes for IPV4 (1358 bytes in case of IPv6 transport) (Harri Holma and Antti Toskala, 2011).

As discussed in Section 2.4.3, packets scheduled also accommodate protocol headers of maximum 20 bytes using ROHC and allows the transmission of approximately 1374 bytes of user's data per subframe. As a result, to meet the throughput and delay requirements of the streaming application, the scheduler is required to send its data every 14^{th} subframe. So, the video data flow needs to be served in total of 77 subframes hereafter referred to as served_sub. As a result, the protocol overhead is estimated to be 10.5 kbps and hence the total rate requirement of the video flow increases to 860.5 kbps. The total required data rate for other applications is estimated in the similar way. The VoIP, online gaming and web applications generates relatively small packet sizes. The scheduler based on the delay requirements of each application concatenates its packets to reduce the protocol overhead.

Table 8.6 clearly indicates that the scheduler requires resources in addition to the data rates given in Table 8.5 to cover the protocol overhead of connections. In this analysis, the UL and DL loading factors are set to 1, which means all resources are available to the RAC. Consequently, it keeps on allocating the left over resources to incoming connections without the consideration of the protocol OH. As a result, the QoS of connections degrades. Therefore, to ensure the QoS of all users, GBR and MBR given in Table 8.5 should be updated to cover the amount of total protocol OH estimated in Table 8.6.

8.4 Analysis of Capacity Estimation

The capacity estimation determines the total number of users that can be supported for specific traffic types, application distributions and system settings. Initially, the capacity estimation is

performed for the application distribution AD-1 of both WiMAX and LTE network. Afterwards, to analyse the impact of change in the application distribution on the capacity of the system, AD-2 is applied. Table 8.7 gives capacity of the system for the downlink of WiMAX networks in terms of the number of users served, using standard (Std) and compressed (Comp) overhead, for AD-1 (Table 8.3). The capacity analysis is given for 700 MHz and 2300 MHz frequencies. The table also provides the slot usage distribution for the 700 MHz frequency.

Table 8.7. Number of Supported Users with 700 MHz and 2300 MHz frequencies and Slot Utilization with 700 MHz, with AD-1 – WiMAX Networks

AD-1						
BW	**5 MHz**		**10 MHz**		**20 MHz**	
	Std. OH	**Comp. OH**	**Std. OH**	**Comp. OH**	**Std. OH**	**Comp. OH**
Users - 2300 MHz	15	18	33	40	70	87
Users - 700 MHz	20	30	42	64	88	137
Slot Utilization at 700 MHz for AD-1						
Data-Slots	32	48	33	51	34	53
OH-Slots	65	52	65	48	64	46
Unused-Slots	3	0	2	1	1	1

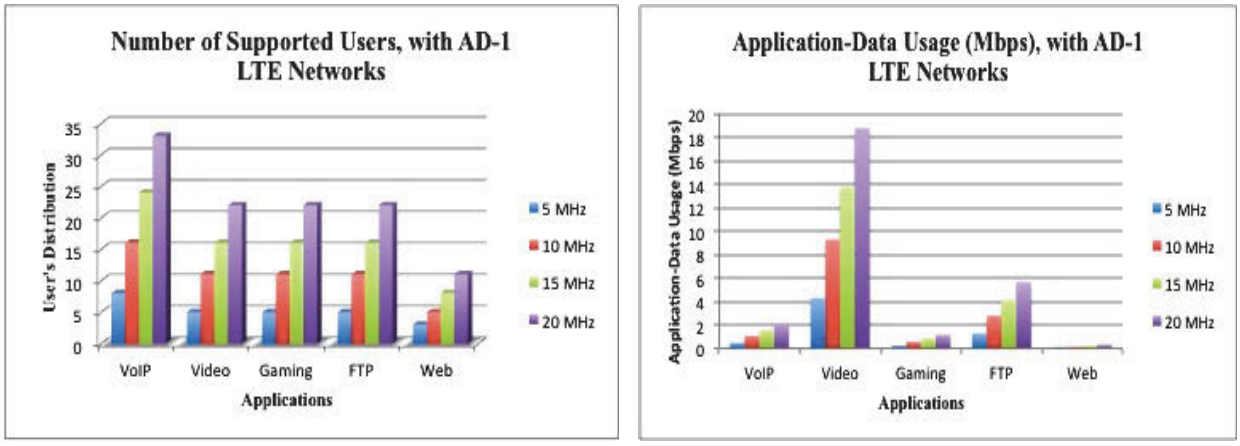


Figure 8.5 Number of Supported Users and Applications Data-Usage (Mbps), with AD-1 LTE Networks

Figure 8.5 provides the capacity of the system for the downlink of LTE networks in terms of the number of users served for AD-1 (Table 8.5). The figure also indicates the total data usage of each application in kbps. This information is provided to analyze the impact of the proposed QoS schemes on the capacity of the network.

The following subsections discuss the impact of various factors such as frequency, bandwidth, application distribution, repetition factor and QoS schemes on the capacity of WiMAX and LTE networks.

8.4.1 Impact of Frequency

The effect of the frequency on the coverage of the system is shown in Figure 8.2. Table 8.7 shows the effect of frequency on the capacity of the system. It indicates that the total number of users that can be served with 2300 MHz frequency is less compared to the number of users that can be served with 700 MHz frequency. This can be contributed to the fact that for 2300 MHz frequency, the probability of using high capacity modulations is low as shown in

Table 8.1. However, for 700 MHz frequency the probability of employing high capacity modulations is high. It results in high capacity of the system with 700 MHz frequency as demonstrated in Table 8.7. So, the use of lower frequency in the network gives better coverage as well as better capacity.

8.4.2 Impact of Bandwidth

The bandwidth of the system impacts both the capacity and the coverage. An increase in the bandwidth also increases the capacity of the system in terms of the number of users served as shown in Table 8.7 and Figure 8.5. It is because as the bandwidth increases the total number of slots or TBPS available to the RAC increases and hence it admits more users in the network. In this scenario, when the bandwidth is doubled, the capacity of system also doubles. Contrary to the throughput, the coverage reduces as the bandwidth increases (Figure 8.3).

The bandwidth also has a direct impact on the number of packets used to transmit user's data. For example, in case of low bandwidth, a subframe in LTE and a frame in WiMAX can accommodate less amount of user's data. Consequently, more packets need to be sent to meet the data rate requirements. As a result, the protocol OH increases. For an instance, in LTE for 3MHz bandwidth and with the same 64QAM-3/5 MCS, the protocol OH for the video applications increases approximately by 5kbps compared to the protocol OH estimated in Table 8.6. As a

result, it reduces the capability of the network to serve the users with the same bandwidth. Hence, lower bandwidths decrease the total number of users that can be supported in the network.

In WiMAX, an alternative to increasing the bandwidth is to increase the frame duration (T_f) in order to increase the capacity of the system with the same coverage.

Table 8.8. Effect of Change in frame Duration- WiMAX Networks

Mobile TV	**A-1**	**A-2**	**A-3**
Data Rate(Kbps)	1215		
Bytes per frame Duration	759.375	1518.75	759.375
MAC PDU-OH	12	12	12
MAP-OH	12	12	12
Total bytes per user	783.375	1542.75	783.375
Slots per user Data using QPSK1/2	130.5625	257.125	130.5625
Total slots per DL-Subframe	455	942	942
Total Users (QPSK1/2)	3	4	7

For three different scenarios, Table 8.8 shows the capacity of the system in terms of the number of users served for Mobile TV using the standard overhead. The A-1 is the scenario with the basic parameters defined for the capacity estimation of WiMAX in section 8.3. Table 8.8 shows that in the scenario A-2 when only the frame duration is increased from 5 ms to 10 ms, the amount of data that each user receives also increases. Consequently, the number of users that the base station can serve increases. As the amount of data that each user can receive is increased, more slots are used for the user traffic and less for the overhead. However, in the scenario A-3 when the bandwidth is increased from 10 MHz to 20 MHz, each user receives the same amount of data. Consequently, the base station can serve higher number of users compared to the aforementioned scenario (A-2).

8.4.3 Impact of Repetition Factor (R)

In the WiMAX networks, the compressed overhead with the Sub-MAPS also affects both the coverage and the capacity. The result in Table 8.7 shows that the compressed overhead with the Sub-MAPS reduces the number of slots used for the overhead along with an increase in the number of users. However, the use of compressed overhead reduces the coverage of the base station, and this is explained with the equation of receiver sensitivity provided in the WiMAX standard (802.16-2005, 2004).

$$R_{ss} = -114 + SNR_{Rx} - 10log_{10}(R) + 10log_{10}\left(\frac{F_s\ X\ N_{Used}}{N_{FFT}}\right) + ImpLoss + NF \qquad 8.14$$

In Eq. 8.14, the component $10\log_{10}(R)$ depends on the repetition (R). When the value of repetition reduces, the value of $10\log_{10}(R)$ also reduces. For the repetition of 1, the value of this component becomes 0. So, when the repetition is reduced, the minimum received power required at the receiver to guarantee BER of 10^{-6} increases, which in turn reduces the radius that can be covered by the base station.

The cell radius for 2300 MHz and 700 MHz frequencies illustrated in Figure 8.2 is given with the MAP repetition of 4. The cell radius for 2300 MHz reduces 0.18 percent and 0.3 percent for the repetition values of 2 and 1, respectively. The cell radius for 700 MHz frequency reduces 0.16 percent and 0.3 percent for the repetition values of 2 and 1, respectively. The cell radius with 700 MHz frequency is larger compared to the radius covered with 2300 MHz frequency (Figure 8.2). So, if the SUB-MAPS are to be used, it is better to use the 700 MHz frequency for better coverage along with high capacity.

8.4.4 Impact of Application Distribution (AD)

Along with the bandwidth employed in the cell, the application distribution also has an impact on the number of supported users as indicated in Table 8.9 and Figure 8.6.

Table 8.9. Number of Supported Users and the Slot Utilization with 700 MHz frequency, with AD-2 –WiMAX Networks

700 MHz – (AD-2)						
BW	**5 MHz**		**10 MHz**		**20 MHz**	
	Std. OH	**Comp. OH**	**Std. OH**	**Comp. OH**	**Std. OH**	**Comp. OH**
Users- 700 MHz	14	18	30	40	64	86
Slot Utilization at 700 MHz for AD-2						
Data-Slots	46	58	48	64	49	67
OH-Slots	54	38	51	34	51	33
Unused-Slots	0	2	0	0	0	0

The AD-1 in Table 8.3 and Table 8.5 covers the areas, which have relatively more businesses, health and public service offices. Whereas, AD-2 covers the areas that have more households, which relatively utilize more streaming services to upload or download videos.

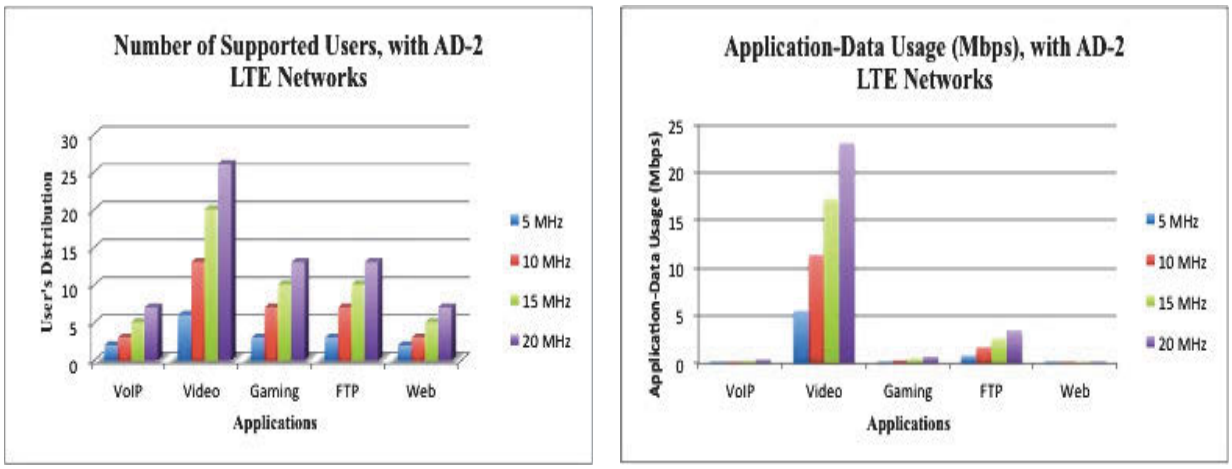


Figure 8.6 Number of Supported Users and Applications Data-Usage (Mbps), with AD-2 LTE Networks

Table 8.9 shows the capacity of WiMAX networks for AD-2 (Table 8.3). Figure 8.6 shows the capacity of LTE networks for AD-2 (Table 8.5). They clearly demonstrate that by increasing the distribution of the streaming application, the total number of users that a BS or an eNodeB can serve with a particular BW reduces. It is because the bandwidth demand of the video users is very high compared to the voice users. Table 8.9 indicates that with AD-2, the total number of users in WiMAX networks with the standard overhead reduces approximately 0.1 % compared to the total users with AD-1 (Table 8.7). Figure 8.6 shows that the total number of users for LTE networks with AD-2 decreases by approximately 0.4 % compared to the total users with AD-1 (Figure 8.5). The decrease in the LTE users is higher compared to the WiMAX users. It is due to the fact that the application distribution of LTE video users is increased from 20% to 40%, whereas the ratio of WiMAX video users is increased from 10% to 20% only. Thus for the application distribution with higher percentage of the application with high bandwidth demand, the total number of users served in a cell significantly reduces.

The higher ratio of applications with the large packet size reduces the number of users in a cell and this in turn decreases the amount of overhead. Table 8.9 illustrates that by increasing the

distribution of Mobile TV from 10% to 20% in WiMAX networks, the percentage of slots used for user's data increases compared to Table 8.7. As a result, there is decrease in the percentage of the slots used for the overhead. In addition to that, the application distribution also affects the amount of unused slots in a frame. For example, Table 8.7 shows with AD-1 and standard overhead, the number of unused slots is 3, 2, and 1 for 5MHz, 10MHz and 20 MHz, respectively. Whereas, Table 8.9 shows with AD-2 and standard overhead, the number of unused slots is 0, 0 and 0 for 5MHz, 10MHz and 20 MHz, respectively. Consequently, better utilization of resources is achieved at the cost of reduction in the number of users in the cell.

8.4.5 Impact of QoS Schemes

When performing network dimensioning, the bandwidth requirements of network connections are planned based on the condition that the network traffic is pre-determined with limited room for variations. Whereas, some services such as file transfer and email, which employs Transmission Control Protocol (TCP) at the transport layer are elastic. They adjust their source rates according to the available bandwidth in the network (Arthur W. Berger and Yaakov Kogan, July, 2001). Several results on dimensioning of LTE networks have been made available in other publications. The authors (Nafiz Imtiaz Bin Hamid et al., May 2012) presented a coverage and capacity analysis of LTE networks for the city of Dhaka. The authors (Abdul Basit Syed, 2009) explained different steps involved in dimensioning procedure. They suggested models and methods for the capacity and coverage planning of LTE networks. The researchers (Xi Li et al., 2013) explained dimensioning models for LTE access transport network including X1 and S1 interfaces. The authors considered both elastic and real time traffics in analysis. The authors (Fredrik Persson, Sept. 2007) discussed the effect of scheduling, antenna diversity and Multiple Input and Multiple Output (MIMO) on the capacity of VoIP in LTE. All of these schemes confine their discussions to only the dimensioning process. They do not deal with the changes in network characteristics once the network has been dimensioned.

Whereas, once the network is dimensioned, the demand for any specific service or mix of services can vary due to changes in the population density, or introduction of an event or a new application. Furthermore, the load in the core network can impose a requirement on the network operators to re-dimension the network. Whereas, by applying QoS mechanisms including RAC

and CC, the network operators may delay the re-dimensioning while still ensuring the QoS to connections.

With RAC, the network can effectively control the admission of each type of service for the advantage of both the users and the network operators. The CC determines the extent to which the load in the core network can be handled by an access network with an acceptable delay. In this section, we discuss the impact of the QoS schemes on LTE networks. The proposed Fair Intelligent Congestion Control (LTE-FICC) and Fair Intelligent Admission Control (LTE-FIAC) schemes utilize the elastic feature of the TCP traffic and the flexibility in bandwidth allocation between the MBR and GBR of a bearer in the latest releases of 3GPP. These schemes employ the bandwidth adaptation (BA) algorithms to accommodate variations in the traffic demand and load at the core network. Consequently, for a given network configuration, they impact the ability of the network to deal with the dynamic variations in the network.

The above sections discuss the capacity of the network for a given traffic distribution without taking any QoS scheme into consideration. Let us first discuss the impact of the proposed RAC scheme, the LTE-FIAC. In times of resource scarcity, LTE-FIAC admits a connection of high priority by applying bandwidth borrowing. It stepwise degrades the bandwidth allocated to connections of low priority services. The bandwidth of a connection is allowed to reduce only to its minimum rate stated by the GBR. Hence, it enables the network to handle changes in the demand of high priority services like voice, video and online games.

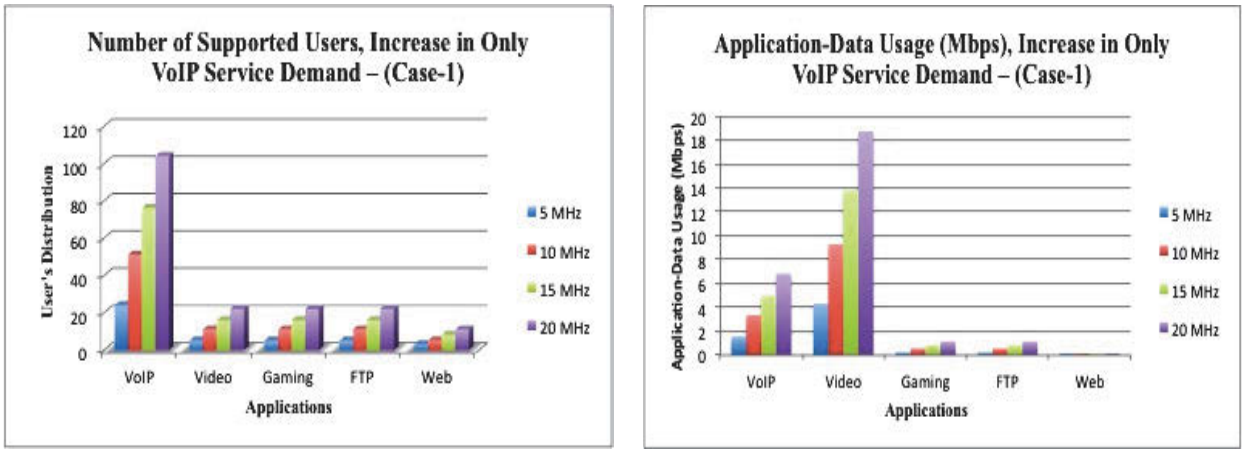


Figure 8.7 Number of Supported Users and Applications Data-Usage (Mbps), Increase in Only VoIP Service Demand – (Case-1)

Figure 8.7 shows the scenario hereafter referred to as case-1, where only the demand of the voice traffic increases. It demonstrates that with the use of the proposed RAC, the network operator with the same network configurations can gain 68% increase in the number of voice users supported compared to the voice users in the base scenario given in Figure 8.5. Figure 8.7 shows after the RAC admits new voice users to the network, the total number of connections of all other services remains the same (Figure 8.5) but the overall data usage of FTP and web applications reduces. This is because LTE-FIAC admits the high priority voice connections by degrading the rates allocated to low priority data flows of FTP and web to their GBR.

Table 8.10 shows three cases. For all the three cases, the table shows the results only for the applications, which undergo a change in terms of the users supported.

In case-2, only the demand of the video application increases. By applying LTE-FIAC, the network can gain 20% increase in the video users compared to the video users supported in the base scenario (Figure 8.5).

In case-3, only the demand of gaming application increases. The RAC by degrading the low priority connections can support 80% increase in the number of gaming users with respect to the gaming users in the base scenario (Figure 8.5). The increase in gaming users is much bigger than the increase in voice and video users as discussed in case 1 and 2. This is because the bandwidth requirement of the online gaming application is the minimum of the three applications according to the specifications given in Table 8.5.

Table 8.10. Number of Supported Users and Applications Data-Usage (kbps) – LTE-Networks

	Case-2	**Case-3**	**Case-4**		
	Video	**Gaming**	**Voice**	**Video**	**Gaming**
BW	**Users- distribution**				
5	6	25	16	5	10
10	13	54	34	11	22
15	20	80	50	17	32
20	27	109	68	23	44
BW	**Application- Data Usage- Kbps**				
5	5378	1341	1030	4574	536
10	11617	2895	2224	9879	1158
15	17210	4290	3295	14636	1716
20	23448	5845	4490	19941	2338

In case-4, there in is an increase in the demand of all three services, voice, video and online gaming. The probability of admitting voice, video and gaming services are assumed to be 0.5, 0.25 and 0.25, respectively. By applying the RAC, the network gains 52%, 5% and 50% increase in the number of users supported for voice, video and online gaming applications, respectively. Table 8.10 also indicates the corresponding changes in the total data usage of the applications for each case. The data usage of FTP and web applications in case-2, 3 and 4 remains the same as given in case-1. This is because in all four cases, the RAC stepwise degrades rates to their minimum to admit incoming high priority connections.

The results show that in times of resources scarcity, the RAC does not admit low priority FTP and web connections by degrading the connections at the same priority, with a view to keep resources for incoming high priority connections. So, during the periods of resource scarcity, using the proposed RAC the BP of low priority services will be high compared to the BP of high priority services.

The proposed congestion control scheme, the LTE-FICC, defines a target operating point at an output buffer of an eNodeB and maintains the queue length around the target point. It employs a queue control function to detect congestion at the core network. In response to congestion, it applies a bandwidth degradation algorithm, which reduces the rate offered to the connections starting with the lowest priority service until the queue length reaches the define target point. It also upgrades connection's expected rate when the queue operates below the target point. In the current scenario, when an eNodeB detects congestion, with LTE-FICC it can reduce the rate allocated to low priority FTP and web connections up to their GBR. Hence, for AD-1, LTE-FICC at maximum reduces the data rate requirements of the cell to around 5.4, 11.6, 17.2 and 23.4 Mbps for 5, 10, 15 and 20 MHz bandwidths, respectively. The eNodeB can ensure the minimum guaranteed QoS to the data flows until the capacity of the core network can support both that is the control channel rate and the data channel's minimum rate offered by LTE-FICC. In situations when capacity of the core is lower than the minimum rate offered by LTE-FICC, the queue at an output buffer of an eNodeB starts building up. As a result, the QoS of connections degrades in terms of throughput and delay.

The network operators, by employing the proposed QoS schemes at an eNodeB, to a certain extent are able to manage the changes in the demand and the load in the core network without the need for re-dimensioning. Once the RAC admits high priority connections by degrading all low priority connections to their respective minimum rate (GBR), the BP of high priority services also starts increasing. The network operators can define a range of acceptable BPs for different services. When the BP of a service such as voice or mix of service such as voice and video goes beyond the range, they can mark it as an indication to re-dimension the network. Similarly, when the capacity of a bottleneck link at the core is lower than the minimum throughput offered by LTE-FICC, the QoS of connections degrades. The network operator can define a threshold for the acceptable minimum QoS. When the QoS of connections reduces below the minimum limit, the condition can trigger an alert for provisioning additional resources or re-dimensioning the network.

In the current analysis, the bandwidth allocations of only the TCP based FTP and web applications are specified by a range (Table 8.5). Therefore, the bandwidth borrowing algorithm of LTE-FIAC is applied to utilize the gap between their respective MBR and GBR. The network operators depending on the preferences may also vary priority and values of MBR and GBR of other applications such as video. This can enable the operators to cope with more dynamic variations in the demand of high priority services (such as voice), and also the load at the core network.

8.5 Summary

In this chapter, we presented the coverage analysis of mobile WiMAX and LTE networks. We also discussed the capacity of the networks under various scenarios. Crucial parameters that affect both coverage and capacity were thoroughly investigated including carrier frequency, channel bandwidth and repetition factor. The impact of overhead and application distribution was also analyzed. It is concluded that they have significant effect on the utilization of resources. By assigning proper values to these parameters, an operator can achieve the desired level of capacity and coverage.

More importantly, this chapter investigated the impacts of the QoS schemes on the capacity of the networks. It presented a general and an efficient approach for the network operators to determine the extent to which the current network configurations with the employed QoS schemes can effectively manage the dynamic variations in the access and the core side of the network. With this approach, an indication of the need for re-dimensioning can be presented to network operators to ensure the QoS of connections in terms of throughput and delay.

Chapter 9 Conclusion and Future Work

In this chapter, a summary of the whole research has been presented. It also outlines the main contributions the thesis has made towards the existing body of knowledge and state of practice. It also maps the directions for the future research by proposing ideas based on the current work.

9.1 Summary and Contribution of This Thesis

In this thesis, we identified critical problems of the existing QoS approaches in the 4G networks. The current 4G standards do not specify any QoS scheme and left them to be vendor specific. Existing approaches in the literature are not efficient enough to cope with the dynamic changes in the traffic demand and load at the core network. Consequently, they are unable to provide fair resource allocation and to guarantee QoS to the users in dynamically changing environment. The primary focus of this research is to enable the network operators in the 4G broadband to effectively manage an increase in the traffic demand and load at the core network and provide the requested QoS to the end users. In this thesis, we proposed innovative and intelligent QoS schemes. These proposed QoS schemes include congestion control and admission control modules. The control algorithms are always active in the network. The parameters involved in the schemes are function of the current network usage and the current network state, which makes the schemes scalable and robust in changing network scenarios.

The research is this thesis opened up a distinct and new way of QoS provisioning in the 4G networks, where QoS mechanisms namely RAC, load control and scheduler work together to keep the network traffic around the target operating point. The schemes are always operational whether the network is overloaded or underutilised and ensure the network resources are optimally utilised.

In this thesis, the proposed QoS schemes are implemented for both 4G technologies, WiMAX and LTE. WiMAX networks do not support the legacy networks such as 2G and 3G. LTE is an evolution of UMTS/HSPA network technologies. Hence, LTE has the ability to coexist with the legacy networks including GSM, GPRS, UMTS, WCDMA, CDMA, CDMA 2000 and EVDO.

In developing countries where the wireless networks are required to be deployed in order to replace the Digital Subscriber Line (DSL), WiMAX networks are a choice for installing the low cost networks. Whereas, LTE is a natural upgrade path for the operators of GSM/UMTS and CDMA 2000 networks. For this reason, NBN Australia selected LTE to deploy fixed wireless in Australia.

Each of the 4G technologies has potential markets. Hence, we implemented and evaluated the performance of our proposed QoS schemes for both 4G technologies, WiMAX and LTE.

The research contribution of this thesis can be summarised as follows:

- ✓ The proposed control algorithms, CC and admission control, are proactive and allow the network to employ the same control policies in all network situations. They enable the network operators to ensure stability in the network.

- ✓ We proposed an innovative CC algorithm for each of the 4G technologies, WiMAX and LTE networks, the WFICC and the LTE-FICC, respectively. Our new CC algorithms change the conventional load control in a way that instead of using thresholds to detect and control network congestion, they employ a target operating point. They estimate an accurate level of fair share for all type of services, with the aim to maintain the desired target operating point. They take into account QoS constraints of each type of service in the network.

- ✓ Simulation results demonstrate that WFICC and LTE-FICC controls traffic at an output buffer effectively, prevents overflows, and ensures the QoS of flows in terms of fair bandwidth allocation, improved throughput and reduced queuing delay.

- ✓ In this thesis, we thoroughly and comprehensively investigated the critical parameters of the CC scheme and discussed the impact of various settings of these parameters on the network performance. We also discussed the value selection of these parameters for the efficient load control. The simulations results illustrated that the proposed CC scheme is robust and relatively insensitive to minor mistuning of the parameters.

- ✓ In this thesis, we presented an intelligent RAC schemes, the WFIAC and the LTE-FIAC, for WiMAX and LTE networks, respectively. The new admission control mechanism changes the conventional admission control to maximise the resource utilisation. The RAC has following main responsibilities.

  Firstly, it does not employ thresholds and avoids resource reservation for any specific type of service. In state of limited resource availability, it employs a step-wise degradation scheme to obtain resources for an incoming high priority connection. The degradation procedure utilises a variable size degradation step.

  Secondly, it includes a load estimation module, which detects congestion at the core side of the networks by measuring the queue length at an output buffer of a base station. The RAC with the load estimation module ensures that an incoming connection will not overload the output buffer of a base station and preserves QoS of existing traffic flows in a fair manner.

  Simulations results demonstrate that when network is lightly loaded, the proposed RAC performs extremely well in terms of lower blocking probability. It guarantees the fair share of bandwidth among the service flows at the same as well as different priority levels. When the network is congested, RAC with the load estimation matches the network capability with the QoS requirements of an incoming connection. Consequently, in times of network congestion, there is a trade-off between the BP of new connections and the QoS of existing connections.

  The results also show that when the network is lightly loaded, higher bandwidth is allocated to the existing connections to gain their maximum rate. The extra resource allocation increases the efficiency of the network; still the network traffic is maintained around the target operating point due to the operation of the load control module.

  Thirdly, RAC reserves additional resources to allow the connections to deal with the channel fluctuations during their holding time and obtain at least their guaranteed rate. Simulations clearly indicate that when performing extra resource reservation, the proposed RAC avoids over reservation of the resources by considering the resources,

which are already allocated to the existing connections above their minimum requirements. As a result, BP of new connections reduces. Hence, the proposed RAC takes into account the specifications of LTE-Advanced that the maximum rate can be assigned a value higher than the guaranteed rate.

The proposed RAC with load estimation and extra resource reservation preserves the QoS of the existing connections.

- ✓ We comprehensively and systematically investigated the parameters that affect coverage and capacity of mobile WiMAX and LTE networks. Importantly, various overhead components are investigated and how they can be optimised to improve the network performance is discussed. The impact of the application distribution is analysed, and the overall conclusion is that it has significant effect on the utilisation of resources and hence the capacity of the network.

  Finally, we investigated impact of QoS schemes on the capacity and dimensioning of the networks. The results of this study are of value to the network designers to effectively determine the capability of the network to deal with the variations in the demography of the covered area and the user's traffic profile with the proposed QoS schemes. It determines when to update the network configurations to provide sufficient QoS to the users in covered area in a cost effective manner.

9.2 Future Work

The results of the research presented in this thesis have opened up further directions for future research. It is of interest to deploy the proposed QoS framework also in femtocells or Wi-Fi hotspots to perform the load balancing. Following new dimensions can be pursued for the research based on the current work:

The work presented in this thesis is based on the simulators. Simulations were performed for various traffic loads and application distributions. The results are still to be tested in a large-scale environment. In future, we intend to test this framework and control algorithms in a real environment, which can represent various applications and population distribution features.

The research in this thesis focused on MAC and network layer. We did not discuss how these proposed schemes at lower layers could interact with transport and application layer protocols to manage the load of the network. Hence, it would be interesting to investigate the cross-layer QoS issues.

Currently, we have proposed mechanisms to reduce user's traffic in the load conditions. However, control channel's traffic in times of congestion also aggravates load. This thesis did not focus particularly for LTE networks that how control channel overheads can be reduced. However, reducing the control channel traffic especially the bandwidth request overhead in both WiMAX and LTE networks can be an interesting future research direction.

In the current implementation of the proposed framework, a target operating point is defined manually. In future work, setting the target point dynamically at a level that is suitable for the throughput and delay can be included.

In the proposed work, admission control module degrades the connections rate to their minimum rate. In future, the scheme can be extended to initially degrade the rate of connections to only their current rate of usage per second that is the MACR. When BP of connections such as high priority voice reaches a certain level or threshold, the rate of connections can be reduced to their minimum.

The dimensioning work in this thesis can be extended to evaluate the performance of other existing schedulers. It can also serve as a basis to design system components efficiently to provide QoS depending on the applications in use; and the preferences of the network operators in terms of users served in a cell, the coverage of the network or the effective utilisation of the resources.

In this thesis, thorough analysis has been conducted to determine the sensitivity of the proposed RAC with the load estimation to the variations in the exponential average factor 'β'. In future we aim to extend current research and further investigate the sensitivity of the proposed RAC with the load estimation to an overselling factor 'α' and the BUR variations, for different network scenarios.

The current work is implemented for the fixed networks without taking into account the effects of handover. In future, the same work can provide a new direction. The refinement of our proposed QoS framework for targeting 5G networks is a significant future work. The 4G networks offer speed closer to Gigabit Ethernet whereas the expected speed of 5G networks is multiple of Gigabits Ethernet (Fagbohun, 2014). The promising technologies, which enable the 5G networks to provide high speed include spatial modulation, advanced error control, carrier aggregation, and massive MIMO techniques at the physical layer (Haider et al., 2014). The basic QoS architecture of the 5G networks is the same as in 4G networks. The proposed QoS framework will be employed in 5G networks in the similar manner as in 4G networks. The FICC in 5G networks will be employed at the MAC layer and executed before the scheduler allocates resources. The FICC for 5G networks will estimate the expected rate for each service type base on the queue length of the eNodeB. The FIAC in 5G networks will also execute in similar manner as in 4G networks and will obtain feedback from the FICC module before admitting any new connection. Hence, as the fundamentals of the proposed QoS schemes are the same for both 4G systems (WiMAX and LTE), the objectives and the basic steps involved in the scheme will remain same for the future 5G technologies to achieve QoS and fairness in networks.

References

3GPP 23.203. Technical Specification Group Services and System Aspects ; Policy and charging control architecture, (Release 12).

3GPP 23.401. Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 12).

3GPP 36.213. Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures, Release 11.

3GPP 36.211. Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 11).

3GPP TS 36.300. Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 11).

3GPP TS 36.321. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (Release 9).

3GPP TS 36.322. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification (Release 9).

231, C. A. Digital mobile radio towards future generation systems, final report. *European Communities***,** 1999.

802.16-2005, I. Part 16 : Air interface for Fixed and Mobile Broadband Wireless Access Sytems. *IEEE Standard for local and Metropolitan Area Networks.*2004.

802.16-2005, I. Part 16 : Air interface for Fixed and Mobile Broadband Wireless Access Sytems. *IEEE Standard for local and Metropolitan Area Networks***,** October 2004.

ABDUL BASIT SYED. Dimensioing of LTE Network, Description of Models and Tool, Coverage and Capacity Estimation of 3GPP Long Term Evolution radio Interface. 2009.

AFRIC, W., MATOSIC, N. & VAKANTE, D. WiMAX on 3.5 GHz Cell Size Calculation. *In:* Multimedia Signal Processing and Communications, 48th International Symposium ELMAR-2006 focused on, June 2006 2006. 195-198.

AHMADZADEH, A. M. Capacity and cell range estimation for mutitraffic user in mobile WiMAX. *Master thesis, University College of Boras, Sweden***,** 2008.

AL-MANTHARI, B., NASSER, N., ALI, N. A. & HASSANEIN, H. Congestion prevention in broadband wireless access systems: An economic approach. *Journal of Network and Computer Applications,* 34**,** 2011. 1836-1847.

ALDMOUR, I. LTE and WiMAX: Comparison and Future Perspective. *Communications and Network,* 2013**,** 2013.

AMRISH KACKER, FRANCK CHEVALIER, KHOOSHIRAM OODHORAH, TRICIA RAGOOBAR, PHILIP BATES & ROBSON, D. Review of efficiency and prudency of NBN Co's fibre, wireless and satellite network design. *Report for Webb Henderson- Public version.*2012.

ANAS, M., ROSA, C., CALABRESE, F. D., MICHAELSEN, P. H., PEDERSEN, K. I. & MOGENSEN, P. E. QoS-Aware Single Cell Admission Control for UTRAN LTE Uplink. *In:* IEEE Vehicular Technology Conference (VTC), 11-14 May 2008. 2487-2491.

ANAS, M., ROSA. C , CALABRESE. F. D, PEDERSEN.K.I, MOGENSEN.P.E. Combined Admission Control and Scheduling for QoS Differentiation in LTE Uplink. *In:* IEEE 68th Vehicular Technology Conference, 21-24 Sept 2008. 1-5.

ANDREWS, J. G., GHOSH, A. & MUHAMED, R. *Fundamentals of WiMAX: understanding broadband wireless networking*, Pearson Education. 2007.

ANTONOPOULOS, A. & VERIKOUKIS, C. Traffic-Aware Connection Admission Control Scheme for Broadband Mobile Systems. *IEEE Communications Letters,* 14**,** 2010. 719-721.

ARTHUR W. BERGER & YAAKOV KOGAN. July, 2001. *Dimensioning Bandwidth and Connection Admission Control for Elastic Traffic in High-Speed Communication Networks*. 09/092,422.

BAE, S., LEE, J., CHOI, B.-G., KWON, S. & CHUNG, M. 2009. A Resource-Estimated Call Admission Control Algorithm in 3GPP LTE System. *In:* GERVASI, O., TANIAR, D., MURGANTE, B., LAGANÀ, A., MUN, Y. & GAVRILOVA, M. (eds.) *Computational Science and Its Applications – ICCSA* Springer Berlin Heidelberg.

BAE., S. J., CHOI., B.-G., CHUNG., M. Y., LEE., J. J. & KWON., S. Delay-aware call admission control algorithm in 3GPP LTE system. *In:* TENCON, IEEE Region 10 Conference, 23-26 Jan 2009. 1-6.

BHANDARE, T. LTE and WiMAX comparison. *MSc, Santa Clara University***,** 2008.

BORODAKIY, V. Y., GUDKOVA, I. A., MARKOVA, E. V. & SAMOUYLOV, K. E. Modelling and performance analysis of pre-emption based radio admission control scheme for video conferencing over LTE. *In:* ITU Kaleidoscope Academic Conference: Living in a converged world - Impossible without standards?, 3-5 June 2014. 53-59.

CARVALHO, G. H. S., WOUNGANG, I., ANPALAGAN, A., COUTINHO, R. W. L. & COSTA, J. C. W. A. A semi-Markov decision process-based joint call admission control for inter-RAT cell re-selection in next generation wireless networks. *Computer Networks,* 57**,** 2013. 3545-3562.

CASEY, T., VESELINOVIC, N. & JANTTI, R. Base Station Controlled Load Balancing with Handovers in Mobile WiMAX. *In:* IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC, 15-18 Sept. 2008. 1-5.

CHANDRA, S. & SAHOO, A. An Efficient Call Admission Control for IEEE 802.16 Networks. *In:* 15th IEEE Workshop on Local & Metropolitan Area Networks (LANMAN), 10-13 June 2007. 188-193.

CHANG, B.-J., LIANG, Y.-H. & LEE, Y.-H. Dynamic-cost-reward connection admission control for maximizing system reward in 4G wireless multihop relaying networks. *Computer Networks,* 57**,** 2013. 2643-2655.

CHAUDHRY, S. B. & GUHA, R. K. Adaptive Connection Admission Control and Packet Scheduling for QoS Provisioning in Mobile WiMAX. *In:* IEEE International Conference on Signal Processing and Communications (ICSPC), 24-27 Nov 2007. 1355-1358.

CHEN, C.-L. Combining quality of services path first routing and admission control to support VoIP traffic. *Future Generation Computer Systems,* 29**,** 2013. 1742-1750.

CHEN, R. R. & KHORASANI, K. A robust adaptive congestion control strategy for large scale networks with differentiated services traffic. *Automatica,* 47**,** 2011. 26-38.

CHEN, R. R. & KHORASANI, K. Markovian jump guaranteed cost congestion control strategies for large scale mobile networks with differentiated services traffic. *Automatica,* 50**,** 2014. 1875-1883.

CHRISTOPHER COX. An introduction to LTE : LTE, LTE-advanced, SAE, and 4G mobile communications. *John Wiley & Sons Ltd***,** 2012.

CISCO. 2014. *Cisco Visual Networking Index: Forecast and Methodology, 2013–2018* [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html [Accessed June, 2014].

CYPRIEN, M. 2012. *WiMAX Maintains Growth in Africa and Middle East* [Online]. Available: http://www.goingwimax.com/wimax-maintains-growth-in-africa-and-middle-east-13978/ [Accessed May 2015].

DELGADO, O. & JAUMARD, B. Joint admission control and resource allocation with GoS and QoS in LTE uplink. *In:* IEEE GLOBECOM Workshops (GC Wkshps), 6-10 Dec 2010. 829-833.

E. O. LUCENA, F.R.M. LIMA, W. C. FREITAS & F. R. P CAVALCANTI. Overload Prediction Based on Delay in Wireless OFDMA Systems. *In:* IEEE Global Telecommunications Conference (GLOBECOM), 6-10 Dec. 2010. 1-5.

EL-SHINNAWY, A. H., NASSAR, A. M. & BADAWI, A. H. A switched scheduling algorithm for congestion relief in WiMAX wireless networks. *In:* International Computer Engineering Conference (ICENCO), 27-28 Dec 2010. 34-39.

EMANUEL B RODRIGUES & FRANCISCO RODGRIGO P CAVALCANTI. QoS-driven adaptive congestion control for voice over IP in multiservice wireless cellular networks. *In:* Communications Magazine, IEEE, 2008. 100-107.

ERCEG, V., GREENSTEIN, L. J., TJANDRA, S. Y., PARKOFF, S. R., GUPTA, A., KULIC, B., JULIUS, A. A. & BIANCHI, R. An empirically based path loss model for wireless channels in suburban environments. *Selected Areas in Communications, IEEE Journal on,* 17**,** 1999. 1205-1211.

FAGBOHUN, O. O. Comparative studies on 3G, 4G and 5G wireless technology. *IOSR Journal of Electronics and communication Engineering, ISSN***,** 2014. 2278-2834.

FANG-CHANG, K., HWANG-CHENG, W., CHIH-CHENG, T., KUO-CHANG, T. & PO-EN, L. Call admission control based resource allocation scheme for LTE uplink. *In:* 15th International Symposium on Wireless Personal Multimedia Communications (WPMC), 24-27 Sept. 2012. 554-558.

FITZEK, F. H. P. & REISSLEIN, M. 2001. *MPEG-4 H.263 Video Traces,* [Online]. Available: http://www2.tkn.tuberlin.de/research/trace/ltvt.html [Accessed October, 2011].

FREDRIK PERSSON. Voice over IP Realized for the 3GPP Long Term Evolution. *IEEE VTC Fall' 07.*Sept. 2007.

GE, Y. & KUO, G.-S. An Efficient Admission Control Scheme for Adaptive Multimedia Services in IEEE 802.16e Networks. *In:* IEEE 64th Vehicular Technology Conference (VTC-Fall), 25-28 Sept 2006. 1-5.

HAIDER, F., GAO, X., YOU, X.-H., YANG, Y., YUAN, D., AGGOUNE, H. M. & HAAS, H. Cellular architecture and key technologies for 5G wireless communication networks. *IEEE Communications Magazine***,** 2014. 123.

HARRI HOLMA & ANTTI TOSKALA. *LTE for UMTS: OFDMA and SC-FDMA based radio access*, John Wiley & Sons. 2009.

HARRI HOLMA & ANTTI TOSKALA. LTE for UMTS : evolution to LTE-advanced. *Chichester : Wiley,***,** 2011.

HATA, M. Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology,* 29**,** 1980. 317-325.

HOANG, D. B. & WANG, Z. Fair Intelligent congestion control Technical Report Number 524.December 1999.

HOU, F., HO, P.-H. & SHEN, X. WLC17-1: Performance Analysis of a Reservation Based Connection Admission Scheme in 802.16 Networks. *In:* IEEE Global Telecommunications Conference(GLOBECOM), Nov. 27-Dec. 1 2006. 1-5.

ITU-R. M.1645 : Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000. 2008.

IVESIC, K., SKORIN-KAPOV, L. & MATIJASEVIC, M. Cross-layer QoE-driven admission control and resource allocation for adaptive multimedia services in LTE. *Journal of Network and Computer Applications,* 46**,** 2014. 336-351.

JAIN, R., CHIU, D.-M. & HAWE, W. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. 1998.

JEFFERY G. ANDREWS, ARUNABHA GHOSH & MUHAMED, R. *Fundamentals of WiMAX Understanding Broadband Wireless Networking*, Prentice Hall Communications Engineering and Emerging Technologies Series. 2007.

JIANFENG, C., WENHUA, J. & HONGXI, W. A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode. *In:* IEEE International Conference on Communications (ICC), 16-20 May 2005. 3422-3426.

JIANG, C.-H. & TSAI, T.-C. Token bucket based CAC and packet scheduling for IEEE 802.16 broadband wireless access networks. *In:* 3rd IEEE Consumer Communications and Networking Conference (CCNC 06), 8-10 Jan 2006. 183-187.

KAMRAN ETEMAD. Overview of mobile WiMAX technology and evolution. *Communications Magazine, IEEE,* 46**,** 2008. 31-40.

KAUR, S. & SELVAMUTHU, D. Adaptive joint call admission control scheme in LTE-UMTS networks. *In:* IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), 4-5 Nov. 2014. 63-70.

KHABAZIAN, M., KUBBAR, O. & HASSANEIN, H. A fairness-based preemption algorithm for LTE-Advanced. *In:* IEEE Global Communications Conference (GLOBECOM), 3-7 Dec 2012. 5320-5325.

KHABAZIAN, M., KUBBAR, O. & HASSANEIN, H. An advanced bandwidth adaptation mechanism for LTE systems. *In:* IEEE International Conference on Communications (ICC), 9-13 June 2013. 6189-6193.

KWAN, R., ARNOTT. R, KUBOTA. M. On Radio Admission Control for LTE Systems. *In:* IEEE 72nd Vehicular Technology Conference Fall (VTC 2010-Fall), 6-9 Sept 2010. 1-5.

KWAN., R., ARNOTT., R., TRIVISONNO., R. & KUBOTA, M. On Pre-Emption and Congestion Control for LTE Systems. *In:* IEEE 72nd Vehicular Technology Conference Fall (VTC-Fall), 6-9 Sept 2010. 1-5.

LAKKAKORPI, J. & SAYENKO, A. Measurement-Based Connection Admission Control Methods for Real-Time Services in IEEE 802.16e. *In:* Second International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ '09) 20-25 July 2009. 37-41.

LEI., H., YU., M., ZHAO., A., CHANG., Y. & YANG., D. Adaptive Connection Admission Control Algorithm for LTE Systems. *In:* IEEE Vehicular Technology Conference (VTC), 11-14 May 2008. 2336-2340.

LI, M. & HOANG, D. B. FIAC: a resource discovery-based two-level admission control for differentiated service networks. *Computer Communications,* 28**,** 2005. 2094-2104.

LI, Y., PAPACHRISTODOULOU, A., CHIANG, M. & CALDERBANK, A. R. Congestion control and its stability in networks with delay sensitive traffic. *Computer Networks,* 55**,** 2011. 20-32.

LIM, H.-T., KIM, Y., JANG, I., PACK, S. & KANG, C.-H. A joint uplink/downlink connection admission control in WLAN/cellular integrated systems. *Mathematical and Computer Modelling,* 57**,** 2013. 2788-2800.

LIU., Z.-H. & CHEN., J.-C. Design and Analysis of the Gateway Relocation and Admission Control Algorithm in Mobile WiMAX Networks. *IEEE Transactions on Mobile Computing,* 11**,** 2012. 5-18.

LUO, S., LI, Z., HU, J., LIU, T. & CAI, B. A Policy-Based CAC Scheme for Fixed WiMAX System. *In:* International Conference on Communication Software and Networks (ICCSN '09), 27-28 Feb 2009. 376-379.

MEHDI., K., OSAMA., K. & HOSSAM., H. Call admission control with resource reservation for multi-service OFDM networks. *In:* International Conference on Computing, Networking and Communications (ICNC), 2012. IEEE, 781-785.

MOHAMMAD T. KAWSER, NAFIZ IMTIAZ BIN HAMID, MD. NAYEEMUL HASAN, M. SHAH ALAM & M. MUSFIQUR RAHMAN. Downlink SNR to CQI Mapping for Different Multiple Antenna Techniques in LTE. *International Journal of Information and Electronics Engineering,* 2 No.5**,** 2012. 756-760.

MURAWWAT, S., ASLAM, S. & SALEEMI, F. Urgency and Proficiency Based Packet Scheduling & CAC Method for IEEE 802.16. *In:* 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCom '09), 24-26 Sept 2009. 1-4.

MWG/AWG. A Comparative Analysis of Spectrum Alternatives for WiMAX Networks with Deployment Scenarios Based on the U.S. 700 MHz Band. 2008.

NAFIZ IMTIAZ BIN HAMID, MOHAMMAD T KAWSER & HOQUE, M. A. Coverage and Capacity Analysis of LTE Radio Network Planning considering Dhaka City. 46, no. 15**,** May 2012. 49-56.

NAGESHAR, N. & VAN OLST, R. A heuristic analysis approach to admission control for voice in packet switched wireless networks. *In:* AFRICON, 13-15 Sept 2011. 1-6.

NS2. 2010b. *The Network Simulator* [Online]. Available: http://www.isi.esu/nsnam/ns [Accessed Dec, 2010].

OPNET. 2012. *opnet modeler release 17.1.A,* [Online]. Available: http://www.opnet.com [Accessed December, 2012].

OVENGALT, C. B. T., DJOUANI, K. & KURIEN, A. "A Fuzzy Approach for Call Admission Control in LTE Networks". *Procedia Computer Science,* 32**,** 2014. 237-244.

PAZHYANNUR, R. S. Comparison of LTE and WiMAX. *IP NGN Architecture Thought Leadership Journal- Q1***,** 2010.

PETTER EDSTROM. Overhead Impacts on Long-Term Evolution Radio Networks. *Master of Science Thesis Stockholm, Sweden***,** 2007.

PHAN., H. T. & HOANG., D. B. FICC-DiffServ: A New QoS Architecture Supporting Resources Discovery, Admission and Congestion Controls. *In:* Third International Conference on Information Technology and Applications (ICITA), 4-7 July 2005. 710-715.

PRIYA, S. V. & FRANKLIN, J. V. Extensive DBA-CAC mechanism for maximizing efficiency in 3GPP: LTE networks. *In:* International Conference on Recent Advances in Computing and Software Systems (RACSS), 25-27 April 2012. 233-237.

QIAN, M., HUANG, Y., SHI, J., YUAN, Y., TIAN, L. & E., D. A Novel Radio Admission Control Scheme for Multiclass Services in LTE Systems. *In:* IEEE Global Telecommunications Conference (GLOBECOM '09), Nov. 30-Dec. 4 2009. 1-6.

QIU., Q., ZHAO., L., PING., L., WU., C. & YANG., Q. Avoiding the evolved node B buffer overflow by using advertisement window control. *In:* 11th International Symposium on Communications and Information Technologies (ISCIT), 12-14 Oct 2011. 268-273.

RAMKUMAR, V., STEFAN, A. L., NIELSEN, R. H., PRASAD, N. R. & PRASAD, R. Efficient admission control for next generation cellular networks. *In:* IEEE International Conference on Communications (ICC), 10-15 June 2012. 1362-1366.

REHMAN, S. 2012. *Overview Of WiMAX In Pakistan* [Online]. Available: http://propakistani.pk/2012/09/17/overview-of-wimax-in-pakistan/ [Accessed May 2015].

RODRIGUES, E. B. & CAVALCANTI, F. R. P. QoS-driven adaptive congestion control for voice over IP in multiservice wireless cellular networks. *IEEE Communications Magazine,* 46**,** 2008. 100-107.

SANKER ET AL. *Reduction of transmission overhead in a wireless communication System*. United States Patent Application Publication. Oct. 15 2009.

SCROXTON, A. 2015. *London businesses turn to Wimax after Holborn fire* [Online]. Available: http://www.computerweekly.com/news/4500244428/London-businesses-turn-to-Wimax-after-Holborn-fire [Accessed May 7 2015].

SEUNG-EUN, H. & WOO-YONG, L. Resource allocation and blocking analysis for mobile TV service over WiMAX network. *In:* International Conference on Information and Communication Technology Convergence (ICTC), 17-19 Nov 2010. 88-93.

SHARMA, P. K. & SINGH, R. Comparative analysis of propagation path loss models with field measured data. *International Journal of Engineering Science and Technology,* 2**,** 2010. 2008-2013.

SO-IN., C., JAIN., R. & TAMIMI., A.-K. Capacity Evaluation for IEEE 802.16e Mobile WiMAX. *Journal of Computer Systems, Networks, and Communications,* 2010**,** 2010.

SURESH., K., MISRA., I. S. & KLPANA., S. Bandwidth and Delay Guaranteed Call Admission Control Scheme for QOS Provisioning in IEEE 802.16e Mobile WiMAX. *In:* IEEE Global Telecommunications Conference (GLOBECOM), Nov. 30-Dec. 4 2008. 1-6.

TELESYSTEM INNOVATIONS INC. LTE in a Nutshell : System Overview.2010.

TSIROPOULOS, G. I., STRATOGIANNIS, D. G., KANELLOPOULOS, J. D. & COTTIS, P. G. Probabilistic framework and performance evaluation for prioritized call admission control in next generation networks. *Computer Communications,* 34**,** 2011. 1045-1054.

TUNG, H. Y., TSANG, K. F., LEE, L. T. & KO, K. T. QoS for Mobile WiMAX Networks: Call Admission Control and Bandwidth Allocation. *In:* Consumer Communications and Networking Conference, 2008. CCNC 2008. 5th IEEE, 10-12 Jan. 2008 2008. 576-580.

UKIL, A. & SEN, J. Proactive resource reservation in next-generation wireless networks. *In:* National Conference on Communications (NCC), 29-31 Jan 2010. 1-5.

V.ERCEG, K. V. S. H., ET AL. Channel models for fixed wireless applications. *IEEE 802.16 Broadband Wireless Access Working Group***,** January 2001.

VULKAN., C. & HEDER., B. Congestion Control in Evolved HSPA Systems. *In:* IEEE 73rd Vehicular Technology Conference (VTC Spring), 15-18 May 2011. 1-6.

WANG, H., LI, W. & AGRAWAL, D. P. Dynamic admission control and QoS for 802.16 wireless MAN. *In:* Wireless Telecommunications Symposium, April 28-30 2005. 60-66.

WANG, L., LIU, F., JI, Y. & RUANGCHAIJATUPON, N. Admission Control for Non-preprovisioned Service Flow in Wireless Metropolitan Area Networks. *In:* Fourth European Conference on Universal Multiservice Networks (ECUMN), Feb 2007. 243-249.

WIMAX FORUM. Mobile WiMAX-Part I: A Technical Overview and Performance Evaluation. *Copyright 2006***,** 2006.

WIMAX FORUM & NIST. 2011. *ns2-wimax-awg* [Online]. Available: http://code.google.com/p/ns2-wimax-awg/source/checkout [Accessed June, 2011].

WONGTHAVARAWAT, K. & GANZ, A. IEEE 802.16 based last mile broadband wireless military networks with quality of service support. *In:* Military Communications Conference, 2003. MILCOM '03. 2003 IEEE, 13-16 Oct. 2003 2003. 779-784 Vol.2.

XI LI, UMAR TOSEEF, DULAS, D., BIGOS, W., GÖRG, C., TIMM-GIEL, A. & KLUG, A. Dimensioning of the LTE access network. *Telecommunication Systems,* 52**,** 2013. 2637-2654.

YUEHONG, G., LI, C., XIN, Z. & YUMING, J. Performance Evaluation of Mobile WiMAX with Dynamic Overhead. *In:* IEEE 68th Vehicular Technology Conference (VTC-Fall), 21-24 Sept 2008. 1-5.

ZOLFAGHARI., A. & TAHERI., H. Queue-aware scheduling and congestion control for LTE. *In:* 18th IEEE International Conference on Networks (ICON), 12-14 Dec 2012. 131-136.

Appendix A. Equations

I. Degradation Priority of a Bearer

It is estimated as follows.

$$pr = \log(q + 1) * \exp(p) \qquad \text{A.1}$$

A scalar value for a priority level is defined based on the QCI priority and the preemption priority of the bearer. The above equation indicates connections with higher values of QCI priority 'q' and ARP priority 'p', have higher degradation priority.

QCI-Priority	log(QCI-Priority+1)	ARP-Priority														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		2.7182818	7.389056	20.08554	54.59815	148.4132	403.4288	1096.6332	2980.958	8103.0839	22026.4658	59874.1417	162754.7914	442413.392	1202604.284	3269017.372
1	0.30	0.82	2.22	6.05	16.44	44.68	121.44	330.12	897.36	2439.27	6630.63	18023.91	48994.07	133179.70	362019.96	984072.29
2	0.48	1.30	3.53	9.58	26.05	70.81	192.48	523.23	1422.28	3866.15	10509.29	28567.23	77653.77	211084.83	573788.06	1559717.67
3	0.60	1.64	4.45	12.09	32.87	89.35	242.89	660.24	1794.72	4878.54	13261.25	36047.83	97988.15	266359.40	724039.92	1968144.57
4	0.70	1.90	5.16	14.04	38.16	103.74	281.98	766.51	2083.60	5663.81	15395.84	41850.23	113760.72	309233.69	840584.32	2284945.09
5	0.78	2.12	5.75	15.63	42.49	115.49	313.93	853.35	2319.64	6305.42	17139.92	46591.14	126647.84	344264.53	935808.03	2543789.96
6	0.85	2.30	6.24	16.97	46.14	125.42	340.94	926.76	2519.20	6847.90	18614.52	50599.52	137543.76	373882.69	1016318.52	2762640.17
7	0.90	2.45	6.67	18.14	49.31	134.03	364.33	990.36	2692.07	7317.81	19891.88	54071.74	146982.22	399539.10	1086059.89	2952216.86
8	0.95	2.59	7.05	19.17	52.10	141.62	384.97	1046.45	2844.56	7732.31	21018.59	57134.45	155307.54	422169.67	1147576.13	3119435.34
9	1.00	2.72	7.39	20.09	54.60	148.41	403.43	1096.63	2980.96	8103.08	22026.47	59874.14	162754.79	442413.39	1202604.28	3269017.37

I. Degradation

$$w\_deg_{(pr,j)} = \sum_{i=0}^{N} \sum_{k=0}^{n_i} f(pr_k, pr_j) \left(PRB_{k_x}^{MBR} - PRB_{k_x}^{GBR}\right) \qquad \text{A.2}$$

Function '$f(pr_k, pr_j)$' in is a function given in (Kwan. et al., 2010). It returns 1 only if bearer 'k' has higher degradation priority than the bearer 'j'. Otherwise, it returns 0. So, $w\_deg_{(pr,j)}$ estimates the total number of PRBS, which can be obtained by degrading all the connections of lower priority than the priority of bearer 'j'.

II. Upgradation

$$w\_u_j = \sum_{i=0}^{N} \sum_{k=0}^{n_i} f(pr_j, pr_k) \left(PRB_{pr_x}^{MBR} - PRB_{k_x}^{MBR}\right) \qquad \text{A.3}$$

Function '$f(pr_k, pr_j)$' in is a function given in (Kwan. et al., 2010). It returns 1 only if bearer 'j' has higher degradation priority than the bearer 'k'. Otherwise, it returns 0. So, $w\_deg_{(pr,j)}$ estimates the total number of PRBS, which can be taken by all the connections of higher priority than the priority of bearer 'j'.

Appendix B. Adding MBR in OPNET

To support MBR in OPNET,

- Variables are added in an EPS bearer definition structure in **lte.h** as follows.

```
typedef struct LteT_EPS_Bearer_Def
    {
    char*                   name;
    unsigned int            qci;
    unsigned int            arp;
    unsigned int            ul_gbr;
    unsigned int            dl_gbr;
    // added by fatima
    unsigned int            ul_mbr;
    unsigned int            dl_mbr;
    double                  delay_budget;
    } LteT_EPS_Bearer_Def;
```

- Initialize the MBR variable in the **lte_attribute_definer.pr.m** as follows.

```
static void  lte_attribute_eps_bearer_definitions_parse (void)
    {
      // added by fatima
      op_ima_obj_attr_get (bearer_def_objid, "Uplink Maximum Bit Rate",
      &(bearer_def_ptr->ul_mbr));
      op_ima_obj_attr_get (bearer_def_objid, "Downlink Maximum Bit Rate",
      &(bearer_def_ptr->dl_mbr));
    }
```