Packet Scheduling Algorithms in LTE Systems

A Thesis

Submitted to

University of Technology Sydney

by

Roshanak Heidari

In accordance with
the requirements for the Degree of
Master of Engineering

Faculty of Engineering and Information Technology
University of Technology Sydney
New South Wales, Australia
October 2017

Certificate Of Authorship/Originality

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

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

Signature of Candidate

Acknowledgment

I would like to acknowledge and express gratitude to my supervisor, Dr Kumbesan Sandrasegaran for his guidance, support, logical thinking, tolerance, encouraging attitude and generous help. I would not be able to finish my work without his guidance.

I would like to thanks my close friends and all CRIN members for support and assistance.

Finally, my endless appreciation to my family for their kindness and help which supported me through the hard moments.

Because of their help, I was able to complete this research.

ABSTRACT

There has been a huge increase in demand towards improving the Quality of Service (QoS) of wireless services. Long Term Evolution (LTE) is a development of the Third-Generation Partnership Project (3GPP) with the aim to meet the needs of International Telecommunication Union (ITU). Some of its aspects are highlighted as follows: increase in data rate, scalable bandwidth, reduced latency and increase in coverage and capacity that result in better quality of service in communication.

LTE employs Orthogonal Frequency Division Multiple Access (OFDMA) to simultaneously deliver multimedia services at a high speed rate. Packet switching is used by LTE to support different media services. To meet the QoS requirements for LTE networks, packet scheduling has been employed. Packet scheduling decides when and how different packets are delivered to the receiver. It is responsible for smart user packet selection to allocate radio resources appropriately. Therefore, packet scheduling should be cleverly designed to achieve QoS that is similar to fixed line services. eNodeB is a node in LTE network which is responsible for radio resource management that involves packet scheduling.

There are two main categories of application in multimedia services: RT (Real Time) and NRT (None Real Time) services. RT services are either delay sensitive (e.g. voice over IP), loss sensitive (e.g. Buffered Video) or both (delay &loss sensitive) for example video conferencing. Best effort users are an example of NRT services that do not have exact requisites and have been allocated to spare resources.

Reaching higher throughput has sometimes resulted in unfair allocation to users who are located far from the base station or users who suffer from bad channel conditions. Therefore, a sufficient trade-off between throughput and fairness is essential. The scarce bandwidth, fading radio channels and the QoS requirement of the users, makes resource allocation a demanding issue. Different scheduling approaches have been suggested for different service demands described briefly throughout the thesis.

Initially, a comprehensive literature review of existing work on the packet scheduling topic has been accomplished in this thesis to realize the characteristics of packet scheduling and the resource allocation for the wireless network. Many packet scheduling algorithms developed to provide satisfactory QoS for multimedia services in downlink LTE systems. Several algorithms considered in this thesis include time and frequency domain algorithms and their way of approach has been investigated.

The next objective of this thesis is to improve the performance of packet scheduling in LTE downlink systems. A new packet scheduling algorithm has been introduced in this thesis. A study on VoLTE (Voice over LTE), video streaming and best effort traffic under three different scheduling algorithms has been conducted. Heterogeneous traffic based on precise modelling of packets has been used in the simulation. The main resource allocation and assignment technique used in this work namely Dynamic Subcarrier Allocation scheme is shown to provide a solution to solve the cross layer optimisation problem. It depends on Channel Quality Information (CQI) and has been broadly investigated for single carrier and multicarrier wireless networks. The problem is based on the maximisation of average utility functions. Different scheduling algorithms in this method consider to be utility functions. The throughput, fairness and Packet Loss Ratio have been considered as the requirements for examining the performance of algorithms. Simulation results show that the proposed algorithm significantly increases the performance of streaming and best effort users in terms of PLR and throughput. Fairness has also been improved with less computational complexity compared to previous algorithms that have been introduced in this thesis.

RELATED PUBLICATION

[1] Heidari, R., Afroz, F., Subramanian, R. Cong, S. Sandrasegaran, K. Kong, X 'Packet Scheduling Study for Heterogeneous Traffic in Downlink LTE 3GPP LTE System', *International Journal of Wireless & Mobile Networks (IJWMN)*, 2015, 7(5), pp. 91–106.

List of Contents

List of Figures

List of Tables

List of Acronyms

1G	First Generation
2G	Second Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
4G	Fourth Generation
ACK	Acknowledgement
AMC	Adaptive Modulation and Coding
AMPS	Analogue Mobile Phone System
ARQ	Automatic Repeat Request
BER	Bit Error Rate
BET	Blind Equal Throughput
BLER	Block Error Rate
BET	Blind Equal Throughput Scheduler
CC	Carrier Component
CD-EDD	Channel-Dependent Earliest Due Deadline
CDMA	Code Division Multiple Access
cdmaOne	code division multiple access One
CP	Cyclic Prefix
CQI	Channel Quality Indicator
CQA	Channel-QoS Aware
CRC	Cyclic Redundancy Check
CRA	Capacity Driven Resource Allocation

CRI	CQI Reporting Interval
CRR	CQI Reporting Rate
CSI	Channel State Information
DFS	Delay First Scheduling
DPS	Delay Prioritized Scheduling
DRA	Dynamic Resource Allocation
DSA	Dynamic Subcarrier Allocation
DSSS	Direct Sequence Spread Spectrum
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
eNB	eNodeB: enhanced Node B
EPC	Evolved Packet Core
EPS	Evolved Packet System
EXP	Exponential Rule
FIFO	First-In-First-Out
FD	Frequency Domain
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FDPS	Frequency Domain Packet Scheduling
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FTGS	Fair Throughput Guarantees Scheduler
GBR	Guaranteed Bit Rate
GPRS	General Packet Radio Services

GSM	Global System for Mobile Communication
HARQ	Hybrid Automatic Repeat Request
HO	Hard handoff
HOL	Head-of-Line
HSDPA	High-Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSPA+	High-Speed Packet Access+
HSUPA	High-Speed Uplink Packet Access
HTTP	Hypertext Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IR	Incremental Redundancy
ISI	Inter-Symbol Interference
ITU	International Telecommunication Union
JTACS	Japanese Total Access Communications System
JTFDS	Joint Time and Frequency Domain Scheduling
LOG-MLWDF	Log- Modified Largest Weighted Delay First
LTE	Long Term Evolution
MAC	Medium Access Control
Max-Rate	Maximum Rate
MTS	Maximum Throughput Scheduler
MCS	Modulation and Coding Scheme
MDU	Max Delay Unit

M-LWDF	Modified-Largest Weighted Delay First
MME	Mobility Management Entity
NACK	Negative Acknowledgement
NMT	Nordic Mobile Telephone
Non-GBR	Non-Guaranteed Bit Rate
NRT	None Real Time
OFDMA	Orthogonal Frequency Division Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
OFPF	OFDMA Frame-Based Proportional Fairness
ODS	Opportunistic and Delay Sensitive
OTFS	Oriented Time Frequency Scheduling
P-GW	Packet Gateway
PAPR	Peak-to-Average Power Ratio
PDC	Personal Digital Communications
PDN	Packet Data Network
PF	Proportional Fair
PHY	Physical
PLR	Packet Loss Ratio
PRB	Physical Resource Block
PSS	Priority Set Scheduler
PFS	Proportional Fair Scheduler
QAM	Quadrature Amplitude Modulation
QAS	Queue Aware Scheduling

QCI	QoS Class Identifier
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAP	Resource units Allocation Priority
RB	Resource Block
RE	Resource Element
RLC	Radio Link Control
RR	Round Robin
RAA	Resource Allocation and Assignment
RRM	Radio Resource Management
RRU	Radio Resource Unit
RT	Real Time
RU	Resource Unit
SAW	Stop-and-Wait
S-GW	Serving Gateway
SC-FDMA	Single Carrier-Frequency Division Multiple Access
SINR	Signal-to-Interference-Plus-Noise-Ratio
TACS	Total Access Communications System
TB	Transport Block
TBR	Target Bit Rate
TD	Time Domain
TDD	Time Division Duplex
TDMA	Time Division Multiple Access

TSN	Transmission Sequence Number
TTI	Transmission Time Interval
UMTS	Universal Mobile Telecommunications System
UE	User Equipment
WCDMA	Wideband Code Division Multiple Access
WIMAX	Worldwide Interoperability for Microwave Access

List of symbols

β	Constant parameter
φ	Constant parameter
ε	A small positive constant
σ	Shadow fading standard deviation
δ_i	Service-dependent PLR threshold of user *i*
$\mu_i(t)$	Priority of user *i* at scheduling interval *t*
λ	Proportion of the total RRUs allocated to real time users at TTI *t*.
$\lambda_{i,s}$	Average arrival data rate of user *i* with traffic type *s*
γ	Path loss exponent
γ	Rate of learning
ψ	Path loss
μ_{dB}	Mean of ψ_{dB}
μ_ψ	Mean of ψ
$\sigma_{\psi_{dB}}$	Standard deviation of ψ_{dB}
α_n	Attenuation of n^{th}path
τ_n	Delay of n^{th} path
θ_n	Phase of n^{th} path
ρ_i	Signal to Noise Ratio for user *i*
$\gamma_{i,j}(t)$	Instantaneous SINR of user *i* on RB_j at time *t*
$\xi_i(t)$	Shadow fading gain of user *i* at time *t*
v_i	The speed of user *i* at time *t*

A^n	Set of queues at time slot *n*
a_i	QoS requirement of user *i*
$B_i(t)$	Total size of all packets (in bits) in the buffer (at base station) of user *i* at scheduling interval *t*
B^n	Set of service types
$BU_i(t)$	Total size of receive packets (in bits) in the buffer (at user) of user *i* at scheduling interval *t*
$C_i[k,n]$	Feasible data rate for RB_k at time slot *n*
d_0	Reference distance for the antenna far field
$d_i(t)$	Time to live of the HOL packet of user *i* at TTI *t*
$dir_i(t)$	Direction of user *i* at time *t*
Dpc	Web browsing reading time
$DP_{l,i}(t)$	Delay of the lth packet of user *i* at time *t*
$E[r]$	Expected mean data rate *r*
$E[R_i(t)]$	Expected mean throughput of user *i* at scheduling interval *t*
$E_{i,j}(t)$	Element (in channel matrix) of user *i* on RRU j at scheduling interval *t*
Fa	Inter-arrival time between the beginnings of successive video frames
f_c	Career frequency
$h_b(t)$	Impulse response of baseband signal
ICI	Inter cell interference in watt
$I_i(t+1)$	Indicator function of the event that packets of user *i* are selected for transmission at scheduling interval *t+1*
K	Unit less constant

$loc_i(t)$ Location of user *i* at time *t*

$m(k,n)$ $RB_k$is assigned to user *m* at time slot *n*

$M_i(t)$ Maximum possible data rate (channel capacity) on all *RUs* for user *i* at TTI *t*.

$\bar{M}_i(t)$ Average subcarrier capacity of i_{th} user.

$mpath_{i,j}(t)$ Multi-path fading gain of user *i* on RB_j at time *t*

$n_i(t)$ Number of RRUs that is allocated to user *i* at scheduling interval *t*

N Total number of users

N_0 Thermal noise

N_d Number of embedded objects in a web browsing packet call

N_p Number of packets (slices) in a video frame

$n_{RT}(t)$ Number of resource units

$n_{NRT}(t)$ Number of resource units allocated to non-real time users at time interval t

$Nservice$ Total number of users of a service

P_a Inter-arrival time between packets (slices) in a video frame

P_{total} Total eNB transmit power in dB_m

PRB_{max} Maximum available number of RBs

$PB_i(t)$ Total number of frames in playback buffer of user *i* at scheduling interval *t*

$P_{li}(t)$ Path loss of user *i* at time *t* in dB_m

$PLR_i(t)$ PLR of user *i* at scheduling interval *t*

$r_{toti}(t)$ Total data rate on all RRUs of user *i* at scheduling interval *t*

$r(t)$	The received signal
$r_i(t)$	Instantaneous data rate (across the whole bandwidth) of user *i* at scheduling interval *t*
$r_i[n]$	Data transmission rate of user *i* at time slot *n*
$r_{i,j}(t)$	Instantaneous data rate of user *i* on RRB *j* at scheduling interval *t*
$R_i(t)$	Average throughput of user *i* at scheduling interval *t*
$R_{ofpfi}(t)$	Modified average throughput of user *i* at time interval *t*
$R_{-sch_i}(t)$	Average throughput of user *i* at simulation time *t*
$RB_{allocated}$	Total number of allocated RBs of all users
RB_{max}	Maximum available number of RBs
RE_{data}	Total number of REs specified for downlink data transmission
RRU_{max}	Maximum available number of RRUs
RRU_{rem}	Remaining RRUs
$RU_{tot}(t)$	Total available RUs at TTI *t*
Service PLR	PLR of a type of service (either GBR or Non-GBR service)
S'	Remaining unused RUs.
S_E	Web browsing embedded object size
$S_i(t)$	Satisfaction degree of user *i* at scheduling interval *t*
S_M	Web browsing main object size
S_p	Video packet (slice) size
t_c	Time constant
t_{-sch_i}	Time invariable equal to 30
T	Total simulation duration

T_i	Service-dependent buffer delay threshold of user *i*
T_p	Web browsing parsing time
$Twait_i(t)$	Waiting time of user *i* from the last scheduled interval until now.
$TOA_{l,i}$	Time of arrival of the l_{th} packet of user *i* in the eNB buffer
U_i	Utility function of user *i*
$V_i(t)$	Number of dead line packets due to violation of i^{th} user packets up to the time *t*.
$vec_i(t)$	Unit norm vector with positive real element of user *i* at scheduling interval *t*
$W_i(t)$	Delay of the HOL packet of user *i* at time interval *T*
$We_{RT}(t)$	Weight of allocation for real time users at time interval *t*
W_{ave}	Average delay of HoL packets for user *i* in simulation time *T*
$W_{wait\ i}(t)$	User *i*'s waiting time duration from the previous scheduling time until current time.
χ	Transmit band pass signal
x_b	Baseband signal

Chapter 1 Introduction

Mobile cellular systems have evolved rapidly over the last two decades. The report of the International Telecommunication Union (ITU) as shown in Figure 1.1 indicated that by the end of 2014 there were 7 billion mobile cellular subscriptions in the world, of which 3.6 billion were from the Asia-Pacific region.

The history of mobile cellular systems focusing on development of the mobile cellular technologies is briefly described here. The First Generation (1G) analogue systems introduced in 1980 were based on circuit-switched and voice communication. The Frequency Division Multiple Access (FDMA) [2] was used as a multiple access method in 1G. Analogue Mobile Phone System (AMPS) and Nordic Mobile Telephone (NMT) were important 1G systems that were introduced in North America, Scandinavia and some European countries. Total Access Communication System (TACS) in Japan under the name of the Japanese Total Access Communication (JTAC) system was a variant of AMPS that was used in the UK, some European countries and Japan in the early 1980's. The main disadvantage of the 1G technology was that it deployed analogue systems which means it was slower, and inefficient for data transmission in terms of spectrum usage and leaded to dropped calls.

In the early 1990s, the Second Generation (2G) mobile cellular system based on circuit-switched technology was introduced. Radio signals in 2G and 1G are digital and analogue respectively, where systems use digital signaling for connecting with radio towers. 2G technology uses two different types of multiple access methods: Time Division Multiple Access/ Frequency Division Multiple Access (TDMA/FDMA) and Code Division Multiple Access (CDMA). In addition to digital voice telephony, the introduction of 2G systems resulted in low rate services including mobile fax, voice mail and Short Message Service (SMS). Researchers in the mobile industry have been focused on optimizing interference mitigation proposals and providing a wider range of services. The Global System for Mobile Communications (GSM) was developed in 1987. GSM, Personal Digital Communication (PDC) and cdmaOne are examples of 2G systems.

In GSM Phase 1, the standard SMS was added to voice services. In Phase 2, fax and data services were added to the GSM specifications [3]. The need for getting higher data rate services resulted in the evolution of 2G networks. General Packet Radio Service (GPRS) was a significant step in the evolution of the GSM networks. GPRS (known as 2.5G) uses packet switched technology unlike GSM which deploys circuit-switched technology.

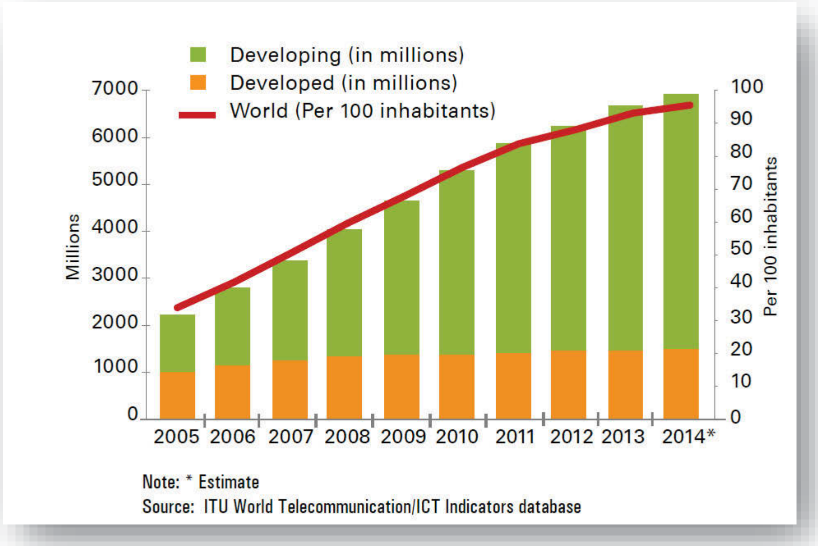

Figure 1.1: Worldwide mobile cellular systems subscription (2005-2014) [1]

The growing demand for accessing internet and broadband multimedia services for mobile users and the coherence of mobile and Internet Protocol (IP) based technologies are the major reasons behind the development of Third Generation (3G). The data rate of delivering services and applications in 3G mobile systems is up to and beyond 2Mbit/s [3].

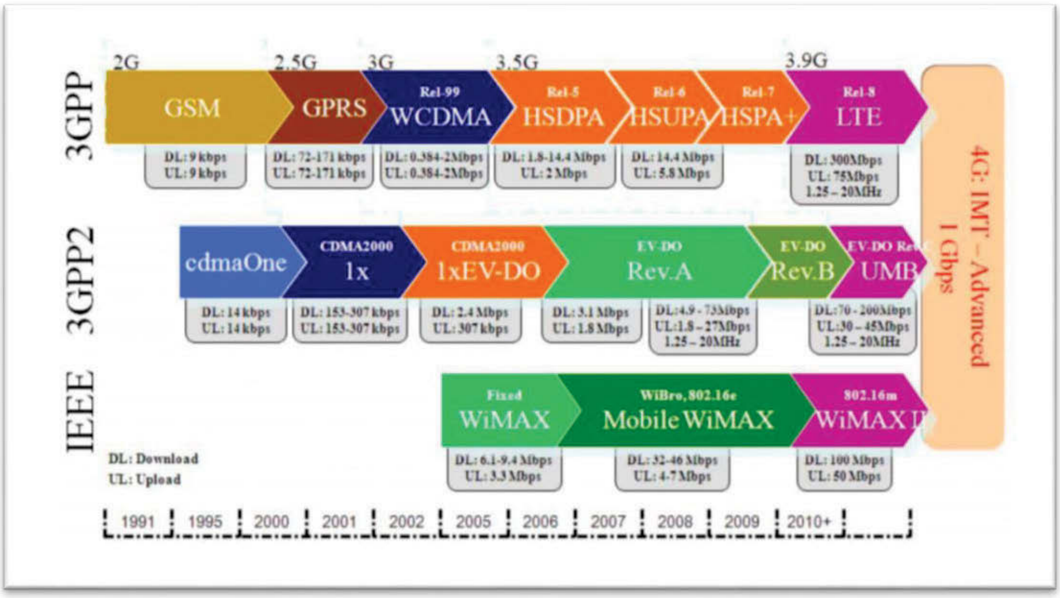

Figure 1.2: Progression in mobile cellular groups [4]

Universal Mobile Telecommunications System (UMTS) introduced in 2001 and is the widely accepted in the European 3G mobile cellular system. It deploys the Wideband CDMA (WCDMA) technology which is based on the Direct Sequence Spread Spectrum (DSSS) with a chip rate of 3.84 Mcps [5].

CDMA2000 was introduced by the Third Generation Partnership Project 2 (3GPP2) organization as a 3G system that is adaptable with the 2G cdmaOne system. Similarly, the UMTS system was used as a compatible version or GSM/GPRS systems. The next step in 3G cellular systems was the development of High-Speed Packet Access (HSPA) family, which was called High-Speed Downlink Packet Access (HSDPA). HSDPA deployed Hybrid Automatic Repeat Request (HARQ), Adaptive Modulation and Coding (AMC), and packet scheduling as important features that resulted in a higher data rate and more global success. HSUPA (High-Speed Uplink Packet Access) and HSPA+ (High-Speed Packet Access+) were further improvements in HSDPA and were known as Release 6 and 7, respectively. The Third Generation Partnership Project (3GPP), 3GPP2 and Institute of Electrical and Electronics Engineers (IEEE) (Figure 1.2) are three organizations whose researches have influenced the evolution towards Fourth Generation (4G).

It is important to note that 3GPP has the leading role in the world's mobile cellular systems.

The latest commercial mobile cellular system known as the Long Term Evolution (LTE) system, was first deployed on the 14th of December 2009 in two Scandinavian capitals. LTE systems are the result of their evolutionary research in this area. The objective behind the goal of the LTE project can be categorized in the following areas: flexibility in frequency and bandwidth, high peak data rate, high spectral efficiency and lower cost for multimedia delivery. The characteristics of 1G-4G technologies have been compared in Table 1.1.

The IEEE organization introduced Worldwide Interoperability for Microwave Access (WIMAX) that is also recognized as the 4G technology. LTE has become more successful and popular compared to WIMAX. The main reason is that LTE is more compatible with previous mobile technologies like GSM, GPRS and UMTS. But WIMAX does not support previous systems like 2G and 3G systems. WIMAX is compatible with previous versions of WIMAX but they are not commercial versions in mobile technology. Another reason is that-LTE provides much greater speed for mobile users (up to 450km/h) compared to WIMAX (120km/h). The other reason for LTE being more successful compared with WIMAX is that LTE uses SC-FDMA for uplink modulation technology, which provides better power consumption and longer battery lifetime of mobile terminals [6]. In September 2009, Long Term Evolution Advanced (LTE-A) has been introduced by 3GPP as an improvement on the LTE standard and is considered to be a true 4G system.

Table 1.1: Comparison of differences in 1-4 Generation in Telecommunication

Parameters	1G	2G	3G	4G
Period	1980-1990	1990-2000	2000-2010	2010-2020
Bandwidth	150/900MHz	900MHz	100MHz	100MHz
Frequency	Analog signal (30 KHz)	1.8GHz	1.6-2.0 GHz	2-8 GHz
Data rate	2.4-14.4kbps	9.6-144kbps	144kbps-2Mbps	100Mbps-1Gbps
Characteristic	First wireless communication	Digital	Digital Broad band, increased speed	High speed, all IP
Technology	Analog cellular	Digital cellular(GSM)	CDMA, UMTS,EDGE	LTE,WiFi
Features	Bad voice quality, poor battery	Allow txt msg, 2.5 G allows E-Mails, Web browsing, camera phones	Smart phones, video calls, fast communication, mobile TV	Mobile multimedia, global mobile support, integrated wireless solution, high security, good QoS

1.1 Introduction to LTE

To guarantee the maintenance of the GSM family of wireless technology in the competitive world market, 3GPP commenced a study about the Long Term Evolution of UMTS in December 2004. The main considerations for the evolution of LTE consisted in providing higher data rates, reducing latency, enhancing system coverage and capacity, and lowering the cost of the operators. [3]. In Figure 1.3, the evolution trend from 2G towards 4G systems is depicted. Some key aspects of LTE are discussed in the following subsections:

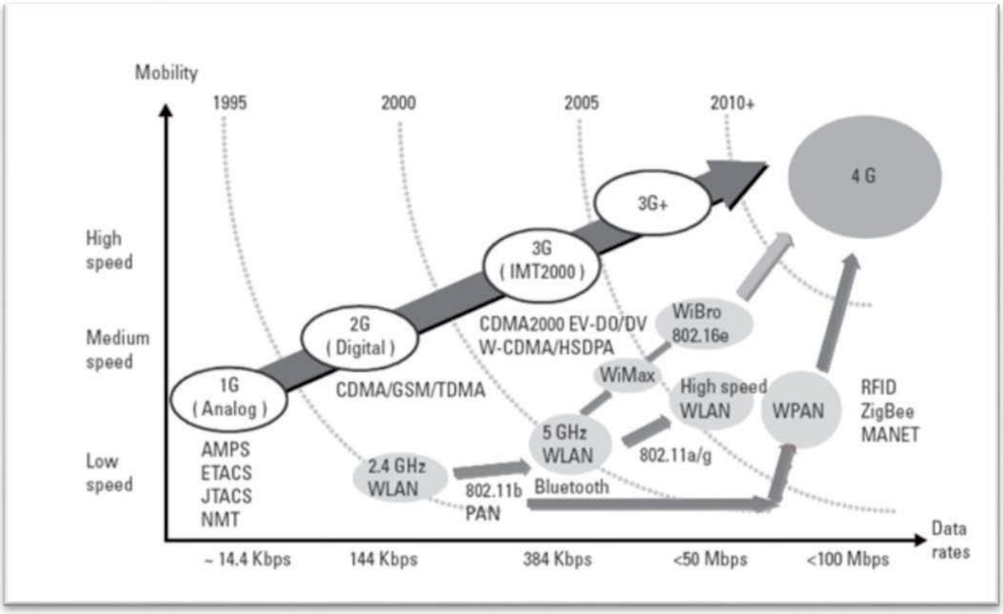

Figure 1.3: Progress in the mobile cellular system's technology [7]

LTE Network Architecture of Network

The Evolved Universal Terrestrial Radio Access Network (E-UTRAN) is the simplified architecture of the access network used in LTE where Node Bs known as enhanced NodeBs (eNBs), connect User Equipment (UEs) with the core network to carry out Radio Resource Management (RRM) functions. Overall, the LTE architecture consists of E-UTRAN and Evolved Packet Core (EPC) that are depicted in Figure 1.4.

There are four EPC network elements namely Mobility Management Entity (MME), Service Gateway (SGW), Packet Data Network (PDN), Packet Gateway (PGW) and the Policy and Charging Rule Function (PCRF). All LTE interfaces are IP protocol based. The eNBs are connected internally by X2 interfaces. The S1 interface is used to connect the MMEs of EPC to the eNodeBs.

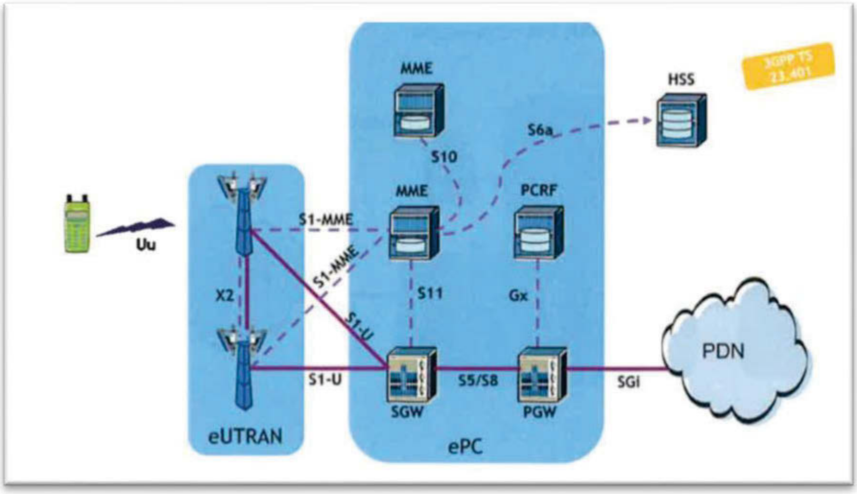

Figure 1.4: LTE general architecture and interfaces

1.2 LTE Air-Interface

In this section, a brief description of the LTE air-interface will be provided.

1.2.1 Flexibility of Spectrum

A wide range of spectrum flexibility is offered in LTE systems which allows them to be deployed in different bandwidths and duplexities. From Figure 1.5, it can be seen that considering the bandwidth availability, the LTE's range of transmission bandwidth extends from 1.25MHz up to 20 MHz [8]. One of the main reasons for migration to LTE is its high degree of bandwidth availability.

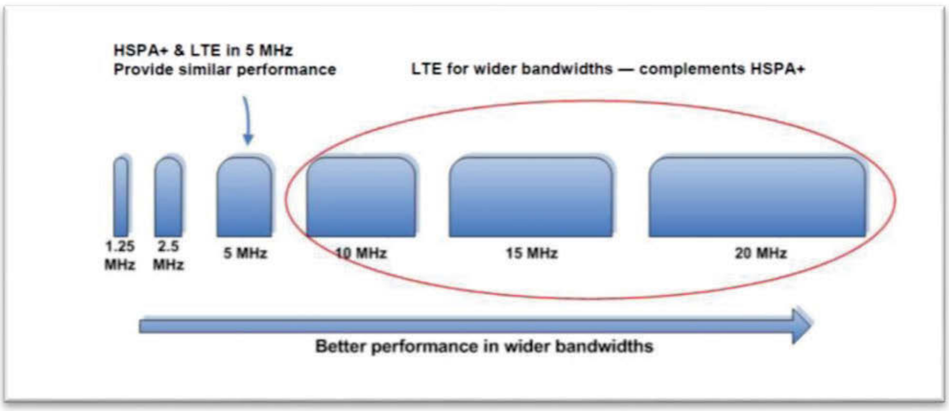

Figure 1.5: Scalable bandwidth in LTE [9]

Smaller bandwidths allow the LTE systems to use the GSM bands. When traffic increases on an LTE network, larger bandwidth is deployed to increase the data rate. The bandwidth scalability of LTE allows its coexistence with systems that work in 900MHz, 2.1GHz and 2.6GHz range spectrums. The LTE system can work in Frequency Division Duplex (FDD) and Time Division Duplex (TDD). In FDD, the uplink and downlink transmissions are separated by frequency. In TDD, uplink and downlink transmissions are separated by time while only one bandwidth is used. The two modes are represented in Figure 1.6 where the majority of systems deploy FDD due to market preference.

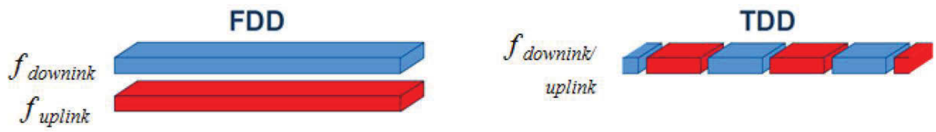

Figure 1.6: Downlink and uplink in FDD and TDD [10]

1.2.2 Multiple Access Schemes

LTE uses Orthogonal Frequency Division Multiple Access (OFDMA) for downlink transmission which is an OFDM (Orthogonal Frequency Division Multiplexing) based multiple access technique. OFDMA is a multiple access technology that consists of a desirable method to control and share resources. In this method, the available spectrum is divided into multiple carriers known as subcarriers. It employs a number of closely spaced orthogonal parallel subcarriers where each subcarrier is modulated by a low data rate flow in the frequency domain. The following three conventional modulation schemes (namely QPSK, 16 QAM and 64 QAM) can be used for each subcarrier. In Figure 1.7 the main aspect of OFDM in the frequency and time domains is presented. To prevent inter-symbol interference (caused by multi-path delay) at the receiver, guard intervals are embedded between symbols in the time domain.

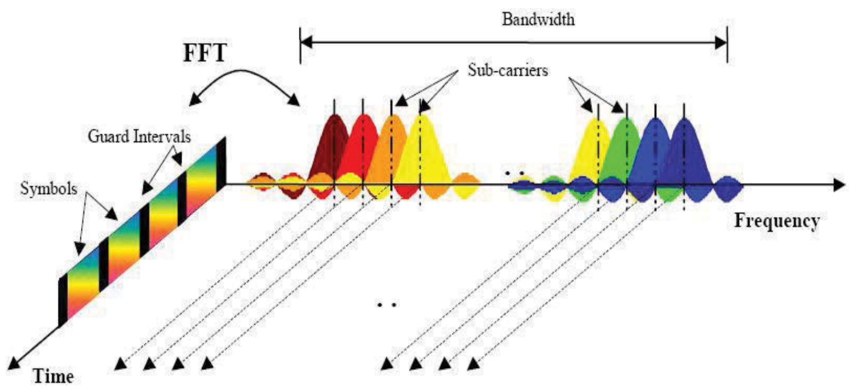

Figure 1.7: OFDM in frequency and time domain [11]

Very tight UE-transmissions can experience fading and interference which is the main reason behind the use of OFDMA in downlink LTE. OFDMA deploys Time Division Multiple Access (TDMA) which allows dynamic allocation of its subcarriers to different users (as shown in Figure1.8). The advantages of OFDMA are its robustness against narrow-band co-channel interference, Inter-Symbol Interference (ISI) and fading, high spectral efficiency and effective implementation using Fast Fourier Transform (FFT). The OFDMA capabilities consist of scheduling users by frequency and multiplexing low rate users enable it to provide robustness towards frequency-selective fading. The drawback of this access is the high Peak-to-Average Power Ratio (PAPR), which results in using a modified form of OFDMA process called Single Carrier-Frequency Division Multiple Access (SC-FDMA) in Uplink (UL). SC-FDMA improves PAPR and reduces the cost of power amplifiers for the mobiles.

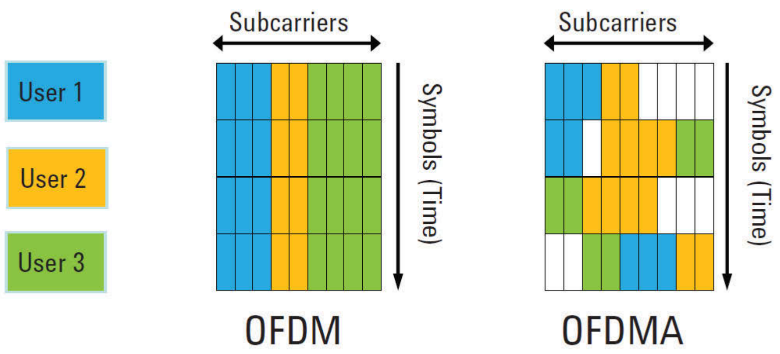

Figure 1.8: Difference between OFDM and OFDMA subcarrier [11]

1.2.3 Resource Block (RB) and Physical Resource Block (PRB)

The units allocated to users in 3GPP LTE systems are called Resource Blocks (RBs) that consist of frequency and time domain resources. Each RB has 180 kHz bandwidth in the frequency domain including 12 consecutive subcarriers. The interval of the time domain for each RB is 0.5 ms. The minimum allocation unit in downlink LTE is called Physical Resource Block (PRB) which consists of two RBs where each PRB is assigned to a user in 1ms time duration.

In order to minimize ISI, a guard time is added between every two symbols. The Cyclic Prefix (CP) is a copy of part of a symbol at the end that has been attached to the beginning of the symbols which facilitates the demodulation. Two types of CPs have been defined in LTE systems: normal and extended CP. In each time interval, 14 and 12 symbols are used in a normal and extended CP respectively. In Figure1.9, an RB with normal CP has been presented in both frequency and time domain.

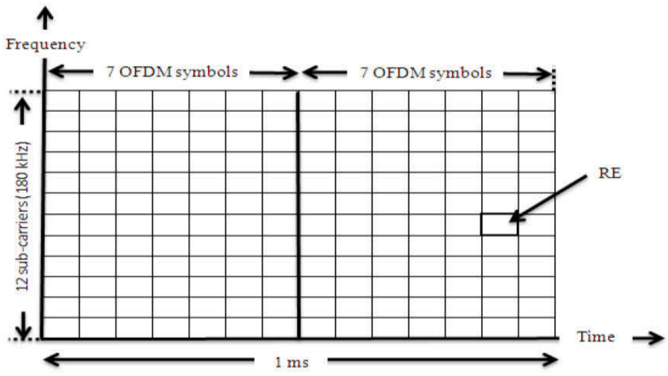

Figure 1.9: PRB in frequency and time domains with normal Cyclic Prefix [12]

An RB contains individual Resource Elements (REs) that include one OFDM subcarrier in one OFDM symbol interval [3]. Therefore, if a normal CP is used, the PRB consists of 168 REs. Most REs are in charge of transmitting data while the rest are used for signaling and controlling tasks. The number of available RBs in the LTE bandwidth is shown in Table 1.2.

Table 1.2: LTE bandwidth and number of RBs [12]

Bandwidth (MHz)	***1.25***	***3***	***5***	***10***	***15***	***20***
Number of available RBs	6	15	25	50	75	100

1.2.4 QoS (Quality of Service) in LTE

One purpose of packet scheduling is to satisfy high Quality of Service (QoS) requirements for different users with different types of applications. In the LTE 3GPP systems, Evolved Packet System (EPS) bearers are responsible for providing different QoS for different users. The QoS requirements of the traffic determine which type of EPS to use. Guaranteed Bit Rate (GBR) and Non-Guaranteed Bit Rate (Non-GBR) are two forms of EPS bearers. QoS factors like QoS Class Identifier (QCIs), the GBR, maximum data rate and allocation retention priority are parameters that define each EPS bearer. QCI and the threshold of performance metrics such as Packet Delay Budget (PDB) and Packet Loss Ratio (PLR) are shown in the table below.

Table 1.3: The standard QCI for LTE systems [13]

QCI	*Service type*	*Priority*	*Packet delay budget (ms)*	*Packet error loss rate*	*Example applications*
1	GBR	2	100	10^{-2}	Conversational voice
2		4	150	10^{-3}	Conversational video (live streaming)
3		5	300	10^{-6}	Non-conversational video (buffered streaming)
4		3	50	10^{-3}	Real time gaming
5	Non-GBR	1	100	10^{-6}	IMS signalling
6		7	100	10^{-3}	Voice, video (live streaming),interactive gaming
7		6	300	10^{-6}	Video (buffered streaming), TCP based i.e. www, e-mail, chat, ftp, p2pfile sharing, progressive video, etc.)
8		8			
9		9			

1.3 Introduction to Packet Scheduling

Packet scheduling is responsible for satisfying the high demand for reliable and high speed data transmission (with acceptable QoS). In the LTE system, scheduling decides when and in what priority packets are permitted to transmit. The decision implements fairness and QoS in utilizing radio resources. A general LTE packet scheduler model is presented in Figure 1.10. The data of each user comes from the network to be stored in their associated buffer at eNB. Every 1ms duration referred to as a Transmit Time Interval (TTI), packet scheduling is carried out at the eNB. The selection of the most appropriate RBs to be allocated to available users is performed by the packet scheduler located in the Medium Access Control (MAC) layer. A packet scheduling algorithm is deployed to do this assignment. The scheduler does the allocation based on Channel Quality Information (CQI), packet delay information, average throughput etc. to achieve satisfactory QoS, effective performance and the guarantee of fairness.

A mapping function is deployed to find the CQI for each PRB and each user sends the related feedback to the related eNB in the time interval. This report helps the eNB to choose the proper Modulation and Coding Scheme (MCS) for its downlink transmissions. One of the essential metrics for evaluating the performance of packet scheduling algorithms is the maximum throughput. The maximum throughput of each user on each PRB is calculated based on MCS. In order to support high data rates QPSK (Quadrature Phase Shift Keying), 16 QAM (Quadrature Amplitude Modulation) and 64 QAM are used in 3GPP LTE systems [14]. Transport Block (TB) is the number of packets that are transmitted to a user in each TTI. The Modulation and Coding Scheme (MCS) that is used on each RB determines the size of TB and therefore the rate of the transmitted data for each user.

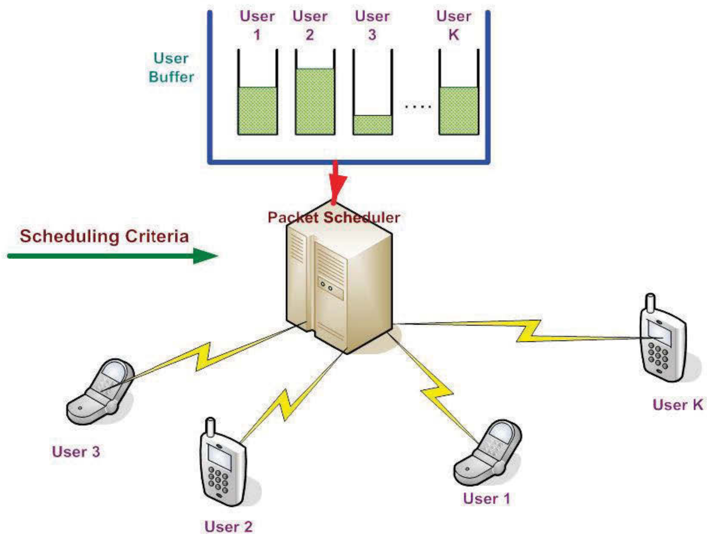

Figure 1.10: General packet scheduler model in downlink LTE system [4]

1.4. **Objective and Motivation**

1.4.1 Research Questions

The purpose of packet scheduling is to maximize throughput, minimize packet loss ratio and delay and achieve the QoS requirements for the network. Therefore, designing a novel packet scheduling algorithm with less computational complexity and better performance is one of the most important purposes of this thesis.

The diverse QoS needs of multimedia services results in complex challenges. The GBR services need a certain number of discarded packets but are more sensitive to delay while the Non-GBR services are more sensitive to packet loss and more tolerant about the delay.

The objective of my thesis is based on the above challenges and research gaps and it will seek to answer the following questions:

- Is it possible to develop a new packet scheduling algorithm with less computational complexity to enhance the throughput, QoS and fairness with different traffic like GBR, NGBR and NRT data streaming? If yes, can the improvement between the previous and the new algorithms be computed and found to be significant?

- Is it possible to design a new packet scheduling algorithm that offers acceptable QoS, better fairness and throughput compared to previous algorithms?

1.4.2. Thesis Overview

A brief explanation of the chapters that remain in this thesis is outlined below:

Chapter2: Downlink LTE Model and Simulation

Modeling of downlink LTE system, Channel Quality Information (CQI), packet scheduling and Hybrid Automatic Repeat Request (HARQ) are introduced in this chapter. The traffic modeling and performance metrics are also discussed.

Chapter3: Packet Scheduling Algorithm

In this chapter a number of time domain algorithms in the literatures are discussed then a more detailed study of four specific packet scheduling algorithms in multicarrier mobile systems with related simulation is provided.

Chapter 4: Scheduling Algorithms Based on QoS Requirement

This chapter considers the frequency-time domain scheduling algorithms based on a QoS algorithm. A new resource allocation algorithm (LOG-MLWDF) for LTE downlink systems is proposed in this chapter which is an effort to maximize the QoS requirement of the system. The new proposed algorithm and its comparison with two other famous algorithms including simulation results are also presented in this chapter.

Chapter 5: Conclusion and Future Research Direction

In this chapter the contribution of this work and a number of recommendations for future research directions are summarized.

Chapter 2 Downlink LTE System Model and Its Parameters

Modeling a large and complex mobile system is feasible through computer simulations. In this chapter, a general model of downlink LTE system is introduced and MATLAB software has been used to simulate packet scheduling in the LTE system. The sections in this chapter are organized as follows. In Section 2.1, a general system model for downlink LTE system has been provided. Section 2.2, 2.3 and 2.4, cover a model of the Channel Quality Information (CQI), packet scheduling and Hybrid Automatic Repeat Request (HARQ), respectively. The performance metrics for traffic evaluation are presented in Section 2.5 and 2.6 and all the assumptions used in the simulations are summarized in Section 2.7.

2.1 Model for Downlink LTE System

A simple model of a downlink LTE system uses a hexagonal cell with of an LTE e-Node B at the centre. Users are uniformly distributed in the cell while an eNB is located at the center of the cell (Figure 2.1). As for simulation parameters, the 5 MHz transmission bandwidth with 900MHz carrier frequency (using the FDD mode) and 25 PRBs ready to be allocated to the users have been considered.

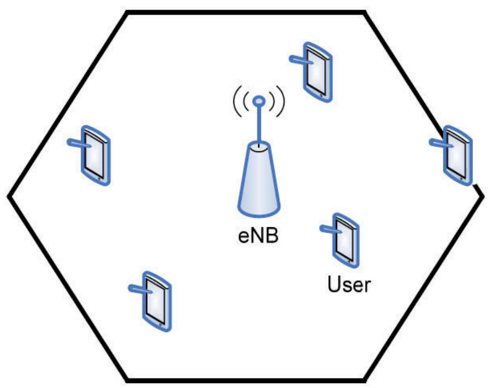

Figure 2.1: Model of simulation consists of one eNB and a number of users

2.1.1. Modeling of Mobility

The speed and the direction of movement for each user will be discussed in this Section. It is assumed that each user is moving with a constant speed ranging from 1km/h to 100km/h in random direction within the cell. At time t, the user *i* position is computed by Equation (2.1).

$$loc_i(t) = loc_i(t-1) + \big(v_i(t-1) * dir_i(t-1)\big) \tag{2.1}$$

Table 2.1: Parameters of 3GPP downlink LTE system

Simulation Parameters	**Values**
Topology of cell	1 hexagonal cell
LTE System Bandwidth (BW)	5 MHz
Carrier frequency	900MHz
System Duplexity	FDD
PRBs for above BW	25
Number of sub-carriers per PRB	12
Total number of Sub-carriers	300
Scheduling interval(TTI)	1ms
Number of OFDMA symbols per TTI	14(Normal CP)
Transmit power eNB transmit power	43.01 dBm

where, the location of user i at time t is depicted by $loc_i(t)$, the speed of user i at time *t-1* is $v_i(t-1)$ and the direction of user i at time *t-1* is $dir_i(t-1)$.

The main downlink specifications have been presented in Table.2.1. The wrap-around model is used to keep users within the simulation area (Figure 2.2). In this procedure when the user reaches the cell boundaries (Figure 2.2(a)), it re-enters the cell from the opposite edge(Figure2.2(b)).

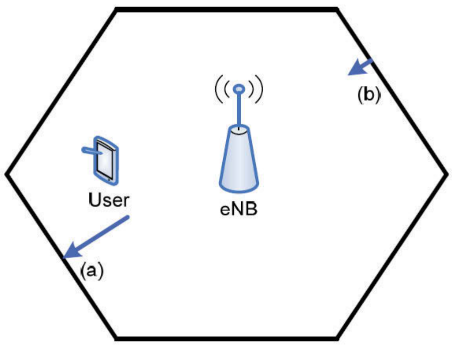

Figure 2.2: Example of a wrapped-round procedure

2.1.2. Model of Radio Propagation

Radio signals are prone to be affected during propagation in a wireless channel. Fading is referred to the variation of signal's strength over time and frequency in a wireless channel. There are two important sources of signal degradation in a wireless channel: *additive noise* and *fading* (also considered as a non-additive signal disturbance). The fading phenomena can be widely categorized in two different types: large and small scale fading. Large scale fading is characterized by average path loss and shadowing. Path loss of a signal is a function of distance and shadowing is a slow fading process characterized by the deviation of the median path loss between the transmitter and receiver. Path loss is the result of power dissipation of the radio signal in the transmission channel. Shadowing is caused by obstacles on the path, so the received signal strength may be different even with the same distance. The obstacles between transmitter and receiver absorb power and cause shadowing. The signal is blocked when the obstacle absorbs all the power. Path loss divergence occurs over large distances (100-1000 meters) while shadowing occurs over distances proportional to the abstracting object (10-100 meters) in an outdoor environment. Because shadowing and path loss occur over large distances, they are sometimes referred to as large-scale propagation effects, whereas multipath variation occurs over very short distances (about the signal wavelength) and therefore they are sometimes referred to as small-scale propagation effects or multipath fading which is a rapid variation of signal levels due to the interference arising from multiple signal paths (multi paths). Figure 2.3 shows the relationship between large-scale fading and small-scale fading.

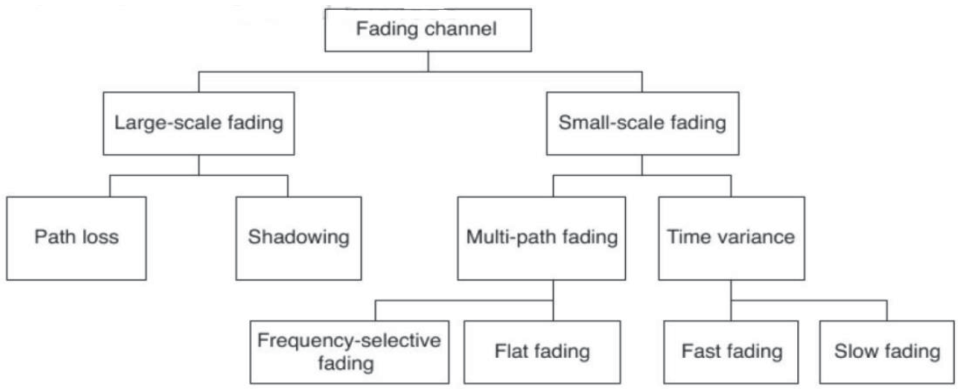

Figure 2.3: Classification of the fading channel [15]

Figure 2.4 shows the ratio of received power to transmitted power in dB versus the combined effect of path loss, shadowing and multipath in log-distance [15].

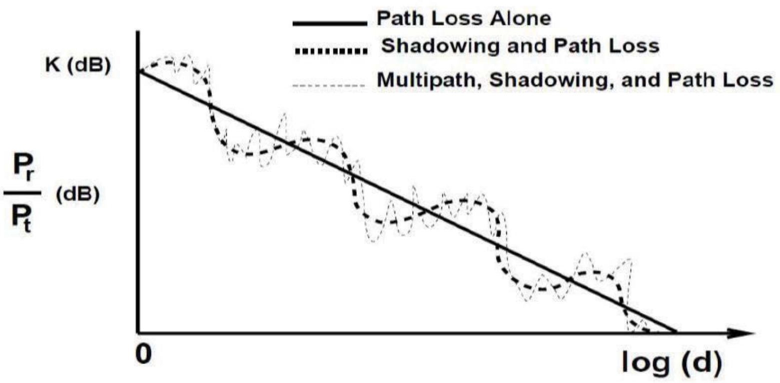

Figure 2.4: Path loss, shadowing and multipath versus distance [15]

2.1.2.1. Simplified Path Loss Model

It is hard to achieve a single model that defines path loss accurately in different environments. Accurate path loss models range from complex ray tracing models to empirical measurements based models when tight system specification must be satisfied or the best location for base stations must be determined. In system design, the simplified path loss model, which is a function of distance, is commonly used as below:

$$P_r = P_t K \left[\frac{d_0}{d}\right]^{\gamma} \tag{2.2}$$

Then:

$$P_r(dBm) = P_t(dBm) + K(dB) - 10\gamma \, log_{10}\left[\frac{d}{d_0}\right]. \tag{2.3}$$

Where K is a unit-less constant which is calculated on the basis of antenna characteristics and the average channel attenuation, $d0$ represents a reference distance for the far-field antenna, and γ is the path loss exponent. Because of the scattering phenomena in the near-field antenna, the model is only valid at transmission distances $d > d0$, where $d0$ is typically assumed to be 1-10 m indoors and 10-100 m outdoors. The value of $K < 1$ is sometimes assumed to be equal to free space path loss at distance $d0$[16]:

$$K(dB) = -20 \, log_{10} {}^{4\pi d_0}/_{\lambda} \tag{2.4}$$

2.1.2.2. Shadowing

In addition to path loss, an obstacle usually blocks the signal and causes random variation at a given distance in the path loss. Changes in reflecting surfaces and scattering objects can also result in random variation in the path loss. Therefore, an adequate model for the random attenuation related to these effects is needed. Since the location, size, dielectric characteristic and scattering obstacles are generally unknown, statistical models are usually employed to describe this attenuation which have been confirmed practically to model the variation in path loss or received power precisely, in both outdoor and indoor radio propagation environments. In the log-normal shadowing model, the path loss ψ is assumed to be random with a log-normal distribution given by

$$p(\psi) = \frac{\xi}{\sqrt{2\pi}\sigma_{\psi dB}\psi} exp\left[-\frac{(10\, log_{10}\psi - \mu_{\psi dB})^2}{2\sigma^2_{\psi dB}}\right], \psi > 0 \quad (2.5)$$

Where $\xi = {}^{10}/_{ln10}$, μ_{dB} is the mean of $\psi_{dB} = 10\, log_{10}\psi$ in dB, and $\sigma_{\psi dB}$ is the standard deviation of ψ_{dB}. If the path loss is log-normal, then the received power and receiver SNR will also be log-normal since these are constant multiples of ψ . The mean of ψ (the linear average path loss) can be obtained from (2.6) as follows:

$$\mu_{\psi} = E[\psi] = exp\left[\frac{\mu_{\psi dB}}{\xi} + \frac{\sigma^2_{\varphi dB}}{2\xi^2}\right] \quad (2.6)$$

The change from the linear mean (in dB) to the log mean (in dB) is derived from (2.6) as follows:

$$10\, log_{10}\mu_{\psi} = \mu_{\psi dB} + \frac{\sigma^2_{\psi dB}}{2\xi} \quad (2.7)$$

The log-normal shadowing's performance is typically parameterized by the log mean $\mu_{\psi dB}$, which is referred to as the average dB path loss. The linear mean path loss, $10\, log_{10}\mu_{\psi}$ in dB, is referred to as the average path loss. With a change of variables, we see that the distribution of ψ (in dB value) is Gaussian with a mean of $\mu_{\psi dB}$ and standard deviation of $\sigma_{\psi dB}$.The log-normal distribution is defined by two parameters: $\mu_{\psi dB}$ and $\sigma_{\psi dB}$ $where$ $\mu_{\psi dB}$ is always nonnegative as blocking obstacles results in attenuation. Although, in some occasions the average attenuation related to both path loss and shadowing is included in the path loss model. For instance, linear path loss models on the basis of empirical data will contain the average shadowing associated with the measurements into the path loss model. In this circumstance, the shadowing model based on the simplified path loss model should have $\mu_{\psi dB} = 0$. Although, if the path loss model does not include average attenuation due to shadowing or if the shadowing model includes path loss via its mean, then $\mu_{\psi dB}$ and $\sigma_{\psi dB}$ will be positive, and must be derived from an analytical model, simulation, or practical measurements[16].

2.1.2.3. Combined Path Loss and Shadowing

To represent power decrease versus distance along with the random attenuation of path loss from shadowing, the combination of two models is developed. The path loss model and shadow fading take the average path loss μ_{dB} and create variations in this mean, as shown by the path loss and shadowing curve in Figure 2.4. This curve represents the combination of the simplified path loss model and the log-normal shadowing random process set for this combined model. The ratio of received power to transmitted power (in dB) is given by:

$$\frac{P_r}{P_t}\,(dB) = 10\,log_{10}\,K - 10\gamma\,log_{10}\frac{d}{d_0} + \psi_{dB}\ . \tag{2.8}$$

Where ψ_{dB} is a Gauss-distributed random variable with a mean of zero and a variance of $\sigma^2_{\psi_{dB}}$. From Eq. (2.8) and Figure 2.4, it can be seen that, the path loss decreases linearly relative to $log_{10}\,d$ with a slope of 10 γ dB/decade, where γ is the path loss exponent; although, the variations due to shadowing change more rapidly.

Throughout the simulations, we set $K = -31.54$ dB, the path loss exponent $\gamma = 6$ and shadowing is considered to follow the log normal model with mean obtained from path loss model and the standard deviation of $\sigma_{\psi dB}$ = 3.65 dB [16].

2.1.2.4. Rayleigh multipath channel model

In this section, a quick overview of a simple statistical multipath channel model called Rayleigh fading channel model is discussed.

2.1.2.4.1. Multipath Environment

We can reasonably characterize a multipath environment by an impulse which transmits from the transmitter to the receiver as a train of impulses.

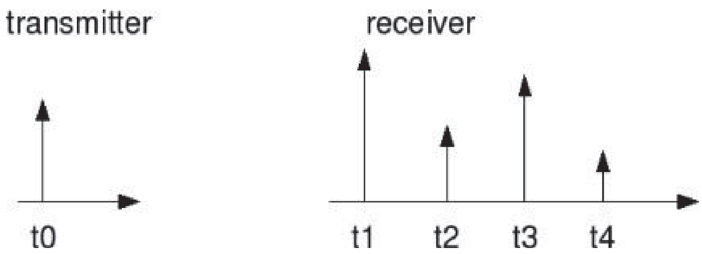

Figure 2.5: Impulse response of a multipath channel

The transmitted band pass signal is assumed to be as follows:

$$x(t) = \Re\{x_b(t)e^{j2\pi f_c t}\}. \tag{2.9}$$

Where $x_b(t)$ is the baseband signal, f_c is the carrier frequency and t is the time.

As shown above, the transmitted signal arrives at the receiver through multiple paths where the n^{th}path has an attenuation of $\alpha_n(t)$ and a delay $\tau_n(t)$.The received signal is

$$r(t) = \sum_n \alpha_n(t)\, x[t - \tau_n(t)]. \tag{2.10}$$

Putting the value of *x(t)* from Eq.2.9 in Eq.2.10, we have the following formula:

$$r(t) = \Re\{\sum_n \alpha_n(t) x_b\, [t - \tau_n(t)] e^{j2\pi f_c[t-\tau_n(t)]}\} \tag{2.11}$$

The baseband equivalent of the received signal is,

$$r_b(t) = \sum_n \alpha_n\, (t) e^{-j2\pi f_c \tau_n(t)} x_b[t - \tau_n(t)]$$

$$= \sum_n \alpha_n\, (t) e^{-j\theta_n(t)} x_b[t - \tau_n(t)]. \tag{2.12}$$

Where $\theta n\ (t) = 2\pi f c\ \tau n\ (t)$is the phase of the n^{th} path with an impulse response of:

$$h_b(t) = \sum_n \alpha_n\, (t) e^{-j\theta_n(t)} \tag{2.13}$$

2.1.2.4.2. Rayleigh Fading Model

The phase of each path can vary by 2π radian when the delay $\tau_n(t)$ varies by $\frac{1}{f_c}$. If f_c is large, relatively small motions in the medium can cause a change of 2π radians. It is assumed that each phase has been uniformly distributed between 0 and 2π radians and is independent from the other paths because the distance between the devices is larger than the wavelength of the carrier frequency.

Each path can be modeled as circularly symmetric complex Gaussian random variable with time as the variable. One form of the circularly symmetric complex Gaussian random variable is presented below:

$$Z = X + J.$$

When there is significant number of paths we can employ the Central Limit Theorem then the model referred to as the Rayleigh Fading Channel model.

2.1.3. Signal to Interference plus Noise Ratio (SINR)

The ratio of signal strength to the sum of the background noise and interference is named after Signal to Interference plus Noise Ratio (SINR). Signal to Noise Ratio for user i, can be calculated as follows:

$$\rho_i(f) = \frac{|H_i(f)|^2}{N_{0_i}(f)}. \tag{2.14}$$

Where $N_{0_i}(f)$ is the noise power density for user *i*. Each RB consists of subcarriers with 15 KHz bandwidth and in this work, the minimum fading variation has been assumed to be between the subcarriers of an RB. The *SINR* for each user is computed at the central frequency of the PRB using the following equation [17]:

$$\gamma_{i,j}(t) = \frac{P_{total} * gain_{i,j}(t)}{PRB_{max}(ICI + N_0)} \tag{2.15}$$

$$gain_{i,j}(t) = 10^{\frac{\rho_{l_i}(t)}{10}} * 10^{\frac{\xi_i(t)}{10}} * 10^{\left(\frac{mpath_{i,j}(t)}{10}\right)} \tag{2.16}$$

where $\gamma_{i,j}(t)$ is the instantaneous SINR for user *i* on RB_j(in dB) at time *t*, $PRB_{max}$is the maximum available number of RBs, P_{total} is the total power in dBm transmitted from eNB, ICI is the inter cell interference in watt and N_0 is the thermal noise in watts. $\rho_{l_i}(t)$ is the path loss of user *i* at time *t* in dB, $mpath_{i,j}(t)$ is the multipath fading gain in dB and $\xi_i(t)$ is the shadow fading gain in dB. Inter-cell interference has been assumed to be constant throughout this work as only one cell has been considered.

2.2 CQI (Channel Quality Indicator)

SINR which is computed in UE and CQI, found through CQI mapping process from SINR, is sent within uplink channel to eNB. The value of CQI corresponds to MCS for a specific Block Error Rate (BLER). The BLER should not be more than a threshold. The recommended threshold is 10% for LTE [18]. Figure 2.6 represents the mapping of SINR to CQI with 10% BLER threshold.

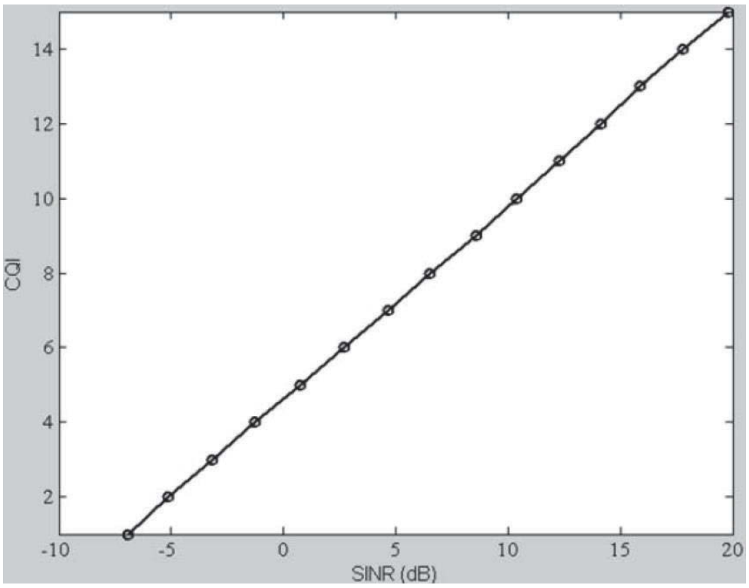

Figure 2.6: Mapping of SINR to CQI (10% BLER threshold) [12]

The CQI which is reported to eNB on each RB can be assumed to be perfect, without delay or error. We consider delay-free and error-free CQI report throughout this work for simplicity. In a real downlink system, the following assumptions need to be taken into consideration:

1. When an updated CQI report is not available, the eNB utilizes the latest accurate report.

2. The eNB can detect and find the errors in CQI report. The CQI report which is not correct and is erroneous will be discarded and the latest report will be employed for packet scheduling.

3. The outdated report of a user is directly used by the eNB. CQI report is used by the packet scheduler to make scheduling decisions and calculate the data rate of the user (which RE's efficiency can also be computed on that basis). Table 2.2 represents the related MCS and efficiency (in bits/RE) of each CQI for 10% threshold.

Table 2.2: CQI table for 10% threshold[19]

CQI index	modulation	code rate x 1024	efficiency
0	out of range		
1	QPSK	78	0.1523
2	QPSK	120	0.2344
3	QPSK	193	0.3770
4	QPSK	308	0.6016
5	QPSK	449	0.8770
6	QPSK	602	1.1758
7	16QAM	378	1.4766
8	16QAM	490	1.9141
9	16QAM	616	2.4063
10	64QAM	466	2.7305
11	64QAM	567	3.3223
12	64QAM	666	3.9023
13	64QAM	772	4.5234
14	64QAM	873	5.1152
15	64QAM	948	5.5547

2.3 Packet Scheduling

Data of users at eNB is divided into small time-stamped, fixed-sized packets and are queued in their user buffer at eNB to be scheduled on a FIFO (First-In-First-Out) basis. In this work, infinite capacity is considered for the buffer at eNB. The timer starts to calculate the time when a packet arrives in its buffer at eNB. The packet delay is the time difference between the arrival time of the packet at eNB until current time. The formula is given as below:

$$DP_{l,i}(t) = t - TOA_{l,i} \qquad l \in packets\ in\ eNB\ buffer \tag{2.17}$$

where $DP_{l,i}(t)$ presents the delay of the l_{th} packet of user i at time t and $TOA_{l,i}$ is the time that the l_{th} packet of user i arrives at eNB buffer. If the packet delay exceeds the delay threshold in the buffer, it will be discarded. The delay threshold of a buffer is assumed to be the maximum permitted waiting time of a packet at eNB. It is dependent on the type of service and application. On each PRB and in the scheduling time interval, the packet scheduler selects the highest priority user to transmit its data packets. Each PRB can be assigned only to one user although more than one PRB may be allocated to one user in each TTI. The packet scheduler always considers retransmitting the packets before the ones which wait for the first transmission.

A TB (Transport Block) is a group of packets that are transmitted to a user in a scheduling interval. A specific Transmission Sequence Number (TSN) is applied for sequential delivery of packets to the application layer [20] by the user. CRC (Cyclic Redundancy Check) bits are inserted in each

TB to do error detection. Based on the MCS of each user on each PRB, the size of TB and the related data rate is determined. Better transmission capacity is the result of higher MCS. A TB architecture with packets and CRC bits has been represented in Figure.2.7.

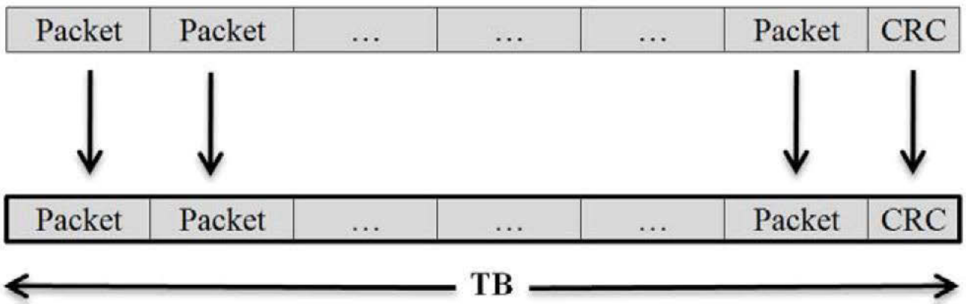

Figure 2.7: TB architecture diagram [12]

All Packets of each user wait in a queue in the eNB buffer up to the time of transmission [21-23] and are removed from the queue when they receive a positive acknowledgement (ACK) feedback related to each TB or the expiry of the re-sequencing timer in the Radio Link Control (RLC) feedback has been received or when more than the maximum number of re-transmissions are reached if the HARQ (Hybrid Automatic Repeat Request) procedure is enabled. If some of the packets in each TB have a delay more than the delay threshold, all packets of TB will be removed from the buffer.

2.4 HARQ (Hybrid Automatic Repeat Request)

ARQ (Automatic Repeat Request) and FEC (Forward Error Correction) have been combined in HARQ technique [24]. Before (re)transmission of each TB it is encoded and then decoded by considering the CRC bits. Once the TB is decoded accurately the user sends an ACK feedback to eNB, otherwise a Negative Acknowledgement (NACK) is sent.

NACK shows a failed decoding and a request for re-transmission will be sent to the eNB. There are two types of HARQ. In type I of HARQ, the incorrect TB is discarded by the user. Type II is a more practical version and in that method, the incorrect TB is stored in the buffer of the user and stays there to be compared with subsequent retransmission(s). Moreover, Type II HARQ includes two famous techniques: (i) CC (Chase Combining) and (ii) Incremental Redundancy (IR) [26].

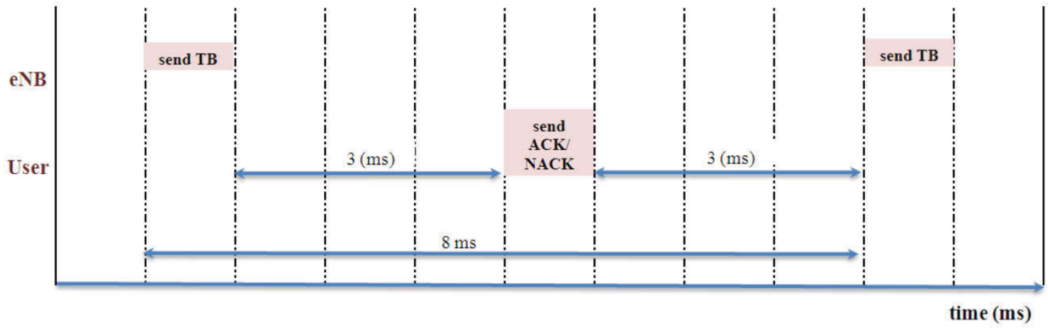

Figure 2.8: Cycle of SAW protocol [25]

In the CC method, the re-transmission of TB is done by identical MCS and the same number of PRBs. The re-transmitted TB is then united with the later received TB(s), which have the identical TSN number at the end of receiver. In the IR HARQ method, several versions of a TB with unlike collection of systematic and parity bits are developed. A new version of TB is sent to the user in each retransmission. In the IR method, we have different MCSs for every retransmission. Combining different re-transmissions can help to improve the efficiency of decoding and transmission.

In this work, we consider a Stop-and-Wait (SAW) protocol in LTE with 8 ms time duration. In that duration the user decodes a TB, implements a CRC, encodes and sends HARQ feedback (ACK/NACK) and the eNB decodes the HARQ feedback, makes and encodes a TB (on the base of HARQ feedback). In Figure 2.8 a complete cycle of the SAW protocol has been presented.

2.5 Characteristics of Traffic

In this work it was assumed that an active data session user runs conversational voice, web browsing and best effort services. The conversational voice is a GBR application while web browsing is a Non-GBR application. For best effort users the network does not guarantee any certain quality of service level. Detailed description of the traffic models is presented as follow:

2.5.1. Web Browsing Traffic Model

Web-page downloads show a user request for data. The reading time is the required time to have the web page. In Table 2.3 the parameters of HTTP (Hypertext Transfer Protocol) traffic model have been presented. The web browsing traffic consists of the following parameters: size of the main object (S_M), size of the embedded object (S_E), number of embedded objects (N_d), reading time (D_{pc}) and the parsing time for the main page (T_p).

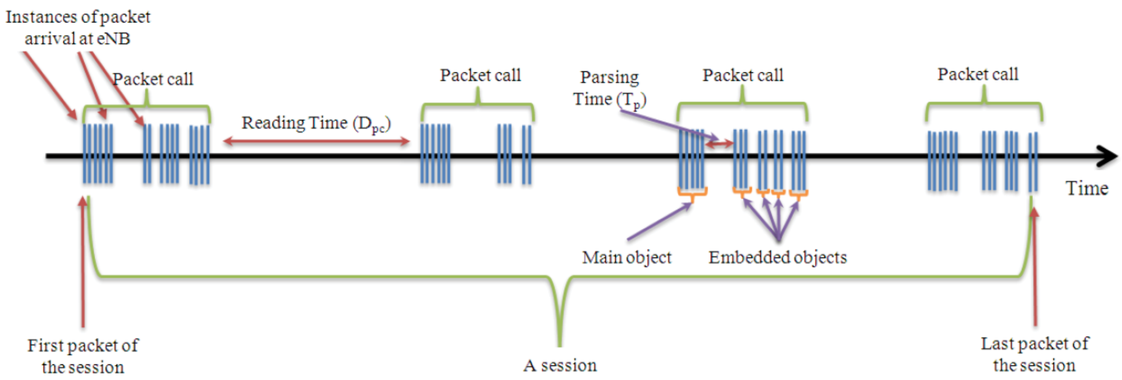

Figure 2.9: A typical web browsing session [27]

The distribution model for S_M and S_E is a truncated log-normal, and a truncated Pareto distribution is used for modeling the number of embedded objects (N_d). The parsing time (T_p) refers to the inter-arrival time between the main object and the first embedded object. To represent T_p and D_{pc} (inter-arrival time between two sequential packet calls), an exponential distribution is used.

Table 2.3: Parameters of HTTP traffic model [27]

Component	Distribution	Parameters	PDF
Main object size (S_M)	Truncated Lognormal	Mean = 10710 bytes Std. dev. = 25032 bytes Minimum = 100 bytes Maximum = 2 Mbytes	$f_x = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left[\frac{-(\ln x-\mu)^2}{2\sigma^2}\right], x \ge 0$ $\sigma = 1.37, \mu = 8.35$
Embedded object size (S_E)	Truncated Lognormal	Mean = 7758 bytes Std. dev. = 126168 bytes Minimum = 50 bytes Maximum = 2 Mbytes	$f_x = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left[\frac{-(\ln x-\mu)^2}{2\sigma^2}\right], x \ge 0$ $\sigma = 2.36, \mu = 6.17$
Number of embedded objects per page (N_d)	Truncated Pareto	Mean = 5.64 Max. = 53	$f_x = \frac{\alpha k^{\alpha}}{x^{\alpha+1}}, k \le x < m$ $f_x = \left(\frac{k}{m}\right)^{\alpha}, x = m$ (NOTE) $\alpha = 1.1, k = 2, m = 55$
Reading time (D_{pc})	Exponential	Mean = 30 sec	$f_x = \lambda e^{-\lambda x}, x \ge 0$ $\lambda = 0.033$
Parsing time (T_p)	Exponential	Mean = 0.13 sec	$f_x = \lambda e^{-\lambda x}, x \ge 0$ $\lambda = 7.69$
Note: Subtract k from the generated random value to obtain N_d			

2.5.2. Video Streaming Traffic Model

Video streaming traffic can be modeled as a number of frames where each frame arrives at a regular time interval (F_a). Each video frame has a fixed number of packets (N_p). To model the size of packets (S_p) and the delay interval between them in a frame (P_a), the truncated Pareto distribution is used. Parameters of video streaming traffic are given in Table 2.4.

Table 2.4: Parameters of video streaming traffic model with 128kbps data rate [27]

Information types	***Distribution and parameters***	***PDF***
Inter-arrival time between the beginning of successive frames (F_a)	*Deterministic* (Based on **20 fps**) 50 ms	--
Number of packets (slices) in a frame (N_p)	*Deterministic* 8 packets	--
Packet (slice) size (S_p)	*Truncated Pareto* **Mean = 100 bytes** **Maximum = 270bytes** **Minimum = 50 bytes**	$f_x = \left[\frac{\alpha k^\alpha}{x^{\alpha+1}}\right]$ ***K*=40 bytes**, α=1.2
Inter-arrival time between packets (slices) in a frame (P_a)	*Truncated Pareto* **Mean = 3.65 ms** **Maximum = 6.25 ms**	$f_x = \left[\frac{\alpha k^\alpha}{x^{\alpha+1}}\right]$ K=2.5ms, α=1.2

2.5.3. Voice Traffic Model

We model voice traffic similar to the Voice over LTE (VoLTE) with a 20ms packet generation interval and a 100ms delay threshold. The rate of voice data generation is 12.65 kbps.

2.5.4. Best Effort Traffic Model

FTP (File Transfer Protocol) traffic is considered to be the model of best effort traffic in this thesis. The parameters of the FTP model are a number of transferred files transmitted separated by reading times. Two main parameters of FTP sessions are a file size in S and the reading time D (Table 2.5). The time difference between the time the download of the earlier file ends and the time that user asks for the next file is considered to be the reading time.

Table 2.5: Parameters of the File Transfer Protocol (FTP)[28]

Parameter	Statistical characterization
File size S	Truncated lognormal distribution mean = 2 Mbytes, standard deviation = 0.722 Mbytes, maximum size = 5 Mbytes (before truncation) PDF: $f_x = \frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(\ln x - \mu)^2}{2\sigma^2}}$ $x > 0$ $\sigma = 0.35, \mu = 14.45$
Reading time D	Exponential distribution with mean = 180 seconds PDF: $f_x = \lambda e^{-\lambda x}$ $x \geq 0$ $\lambda = 0.006$

2.6 Performance Metric

The performance of packet scheduling algorithms is evaluated by system throughput, fairness, service Packet Loss Ratio (PLR) and average delay of the system as metrics in this thesis. The total data rate of all users that have successfully delivered their packets in a cellular mobile system is called system throughput. In this work, the total summation of correctly received packets of users in downlink LTE system divided by the simulation time is defined as average throughput. The mathematical formula for this definition is as below:

$$averagethroughput_i = \frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} trpac_i(t). \tag{2.18}$$

Where T is total simulation time for user i, N is number of users in service time and $trpac_i$ is total size of correctly received packets.

PLR is defined as the total number of discarded packet divided by total number of all packets that were received in the buffer of eNB during the interval. To cover the QoS needs of a service, PLR should be below the threshold value and computed according to following equation:

$$PLR = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} pdiscard_i(t)}{\sum_{i=1}^{N}\sum_{t=1}^{T} nptotal_i(t)} \tag{2.19}$$

where, $pdiscard_i(t)$ is total number of discarded packets (in bits) in the simulation time T (for user i), $nptotal_i(t)$ is the total number of all packets (in bits) that reach the eNB buffer of user i at simulation time T where T is the simulation interval and N is the total number of users.

Fairness is defined as the difference between two users that have the most and least performance metrics (as number of transmitted packets, throughput, delay.) divided by the total number of all packets that have arrived at the buffer of eNB subtracted from 1. In this work we consider the number of transmitted packets as metric; therefore, if the difference between users according to the number of transmitted packets is high the fairness will be less. The formula is as below:

$$fairness = 1 - \frac{ptotaltransmit_{max} - ptotaltransmit_{min}}{\sum_{i=1}^{N}\sum_{t=1}^{T} nptotal_i(t)} \tag{2.20}$$

Where $ptotal\,transmit_{max}$ and $ptotaltransmit_{min}$ are the total number of correctly received packets of the most and the least served users at time t, respectively. $\mathrm{nptotal}_i(t)$ is the total number of all packets (in bits) arrived at eNB buffer of user i in simulation time T.

The total average delay of packets is the average queuing delay of the system in eNB. To have better performance, the average system delay should be kept at a minimum level. In this work, the average delay of Head of Line (HoL) packets is considered as the average delay. The packet that has the longest waiting time in each user's buffer at eNB is defined as HoL packet.

$$W_i(t) = max\{DP_{l,i}(t)\}\ l \in packets\ in\ eNB\ buffer \tag{2.21}$$

where HoL packet's delay of user i at time t is represented as $W_i(t)$ and $DP_{l,i}(t)$ is the delay of l_{th}packet of user *i* at time *t*. The mathematical formula for the total average delay of packets is defined as below:

$$total\ average\ delay\ of\ packets\ in\ the\ system = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N} W_i(t). \tag{2.22}$$

Where $W_i(t)$ is Head of Line packet delay of user *i* at time *t*, *T* is the total simulation duration and *N* is the total number of users.

2.7 Summary

In this chapter, a general model of the downlink LTE system and the simulation model consisting of channel and mobility model have been described. The simulation modeling of the CQI, packet scheduling and HARQ have been discussed. The GBR and Non- GBR services that have been used in traffic simulation have been described. Finally, the performance evaluation metrics have been discussed.

Chapter 3 Scheduling Algorithm for Packet Cellular Networks

One of the most important aspects of RRM (Radio Resource Management) in LTE systems is packet scheduling which refers to the process of selecting user's packets effectively to guarantee fairness, improve system performance and satisfy Quality of Service (QoS) requirements. The packet scheduler, which is located at the eNB in the MAC layer, has become the outstanding feature in LTE because it uses the packet-switched method to deliver the multimedia data services.

Various packet scheduling algorithms have been introduced in single-carrier mobile systems to satisfy QoS of different multimedia services. They considered the time domain in packet scheduling. The LTE system deploys resource allocation in both frequency and time domains (as described in Chapter1).

We first introduce a number of time domain packet scheduling algorithms in mobile cellular system. The performance of Max-Rate (Maximum Rate) [29], RR (Round Robin) [30], PF (Proportional Fair) [31], BET (Blind Equal Throughput), Delay Prioritized Scheduling (DPS), Modified-Largest Weighted Delay First (M-LWDF), Channel-Dependent Earliest Due Deadline (CD-EDD) and Exponential Rule (EXP) algorithms will be studied throughout this chapter. These algorithms prioritize the packets based on Channel State Information (CSI), although some of them consider buffer state information in their resource allocations as well. Then, the performance of four famous time domain packet scheduling algorithms is investigated in a multi-carrier LTE system to prove the validation of the simulator in Chaper 2. The results are also shown to be in line with theoretical results in the literature.

The chapter is organized as follows. Time domain packet scheduling algorithms are illustrated in Section 3.1. The performance of three famous single carrier algorithms namely Max-Rate, RR, DPS and PF in LTE system is described with simulations in Section 3.2 followed by a summary at the end.

3.1 Packet Scheduling Algorithms in Time Domain

Some of the well-known packets scheduling algorithms in time domain are discussed in the following subsections.

3.1.1 Max-Rate (Maximum Rate) Algorithm

The aim of the Maximum Rate algorithm [29] is to maximize the throughput by using the time-varying wireless channel condition. However, this algorithm does not guarantee a fair allocation among users. In this packet scheduling algorithm, a user with poor channel condition has less chance for transmission. In each assignment interval, the algorithm chooses a user that maximizes the metric below:

$$\mu_i = r_i(t). \tag{3.1}$$

Where $\mu_i(t)$ is the scheduling formula for user *i* in time interval *t* and $r_i(t)$ is the instantaneous data rate of user *i* in whole bandwidth at time interval *t*.

3.1.2 RR (Round Robin) Algorithm

To allocate equal transmission resources among users, the Round Robin algorithm was developed. It provides equal opportunity to users in turn to transmit the packets and has shown to improve fairness considerably. The algorithm provides great fairness performance while it may cause degradation in throughput. Since channel quality is not considered in the selection of the RR scheduling algorithm, the throughput of the system cannot be optimized.

3.1.3 PF (Proportional Fair) Algorithm

To achieve a good trade-off between fairness and throughput, the PF algorithm has been developed for the Non-GBR services in CDMA networks [31]. The aim of the PF algorithm is to increase the throughput of the users whose instantaneous achievable data rate is better compared with their average throughput. This algorithm does not consider the buffer information of the user including delay of Head of Line (HoL) packets, therefore, it cannot provide good service for real time services. The scheduling formula is presented in Equation (3.2), Equation (3.3) and (3.4).

$$\mu_i(t) = \frac{r_i(t)}{R_i(t)} \tag{3.2}$$

$$R_i(t+1) = \left(1 - \frac{1}{t_c}\right) R_i(t) + I_i(t+1) * \frac{1}{t_c} * r_i(t+1) \tag{3.3}$$

$$I_i(t+1) = \begin{cases} 1 & when\ packets\ of\ user\ i\ are\ allocated\ at\ interval\ t+1 \\ 0 & when\ packets\ of\ user\ i\ are\ not\ allocated\ at\ interval\ t+1. \end{cases} \tag{3.4}$$

Where $\mu_i(t)$ is the PF formula, $r_i(t)$ indicates the instantaneous data rate of user i calculated in the whole bandwidth in interval *t*, $R_i(t)$ represents the average throughput of user *i* at time interval *t*, $I_i(t+1)$ is the index function that shows whether the packet is chosen for transmission at time interval $t+1$ or not and t_c is a time constant. This time constant provides a control over maximizing throughput and guaranteeing fairness in the PF algorithm. If we choose a greater t_c, the algorithm is more similar to the Max-Rate algorithm, and with choosing lower t_c, it is comparable with the RR algorithm.

3.1.4 BET (Blind Equal Throughput) Algorithm

To provide efficient fairness, Blind Equal Throughput has been developed in [32] for LTE systems, which does not consider channel state in making scheduling decisions. This algorithm stores a record of the previous average throughputs of each user and tries to guarantee a fair assignment among the users. The BET algorithm allocates a channel to a user to maximize $\mu_i(t)$ in the following formula:

$$\mu_i(t) = \frac{1}{R_i(t)}. \tag{3.5}$$

Where, $\mu_i(t)$ defines the preference of each user at each time interval, and $R_i(t)$ shows the average throughput at scheduling time *t* for user *i*.

This algorithm is not a good candidate for increasing the throughput as compared with PF and Max-Rate because it does not consider the channel quality in making decisions.

3.1.5 DPS (Delay Prioritized Scheduling) Algorithm

To satisfy the needs of real time applications, packet scheduling algorithms should consider the delay deadlines of packets to prioritize them. The Delay Prioritized Scheduling (DPS) algorithm is a simple scheduling algorithm that considers packet delay information in its equation. This algorithm prioritizes the users which their delays are about the deadline threshold to satisfy the QoS requirements of GBR services in Downlink LTE systems.

$$d_i(t) = T_i - W_i(t). \tag{3.6}$$

Where $W_i(t)$ is Head of Line (HoL) packet delay of user *i* at TTI *t*, T_i is the delay threshold of the buffer, which is dependent on the type of service, and $d_i(t)$ is the live time of the HoL packet of user *i* at TTI *t*.

3.1.6 M-LWDF (Modified-Largest Weighted Delay First) Algorithm

In order to consider fairness, throughput and also the real time requirements, M-LWDF has been investigated in [33]. This algorithm will increase the Quality of Service (QoS) of several real time users which use a wireless channel. The packet delay, average throughput and instantaneous data rate along all bandwidths have been included in the M-LWDF packet scheduling algorithm. This algorithm was proposed for a Code Division Multiple Access-High Data Rate (CDMA-HDR) system. M-LWDF prioritizes users based on maximizing $\mu_i(t)$ in Equation (3.7).

$$\mu_i(t) = a_i * W_i(t) * \frac{r_i(t)}{R_i(t)} \tag{3.7}$$

$$a_i = -\frac{(\log \delta_i)}{T_i}. \tag{3.8}$$

Where a_i is the QoS requirement of user i, $W_i(t)$ is HoL packet delay of user i at TTI t, $r_i(t)$ shows the instantaneous data rate along all bandwidth and $R_i(t)$ is the average throughput of user i at TTI t (Equation (3.3)). In Equation (3.8), δ_i and T_i represent PLR and the buffer delay threshold of user i respectively, which are dependent on the type of services.

3.1.7 EXP (Exponential Rule) Algorithm

To satisfy the requirements of real time and non-real time services in HDR/CDMA systems, a scheduling algorithm named the Exponential Rule (EXP) algorithm was proposed in [34]. The formula below represents the EXP algorithm:

$$\mu_i(t) = \alpha_i * W_i(t) * \frac{r_i(t)}{R_i(t)} * exp\left(\frac{\alpha_i * W_i(t) - \alpha W_{-avg}}{1 + \sqrt{\alpha W_{-avg}}}\right) \tag{3.9}$$

$$\alpha W_{-avg} = \frac{1}{N} \sum_{i=1}^{i=N} \alpha_i * W_i(t). \tag{3.10}$$

Where $\mu_i(t)$ is the priority of user i to receive the packets in each TTI t, a_i is the QoS requirement of user i (see Equation (3.8)), $W_i(t)$ is HoL packet delay of user i at TTI t, $r_i(t)$ shows the instantaneous data rate along all bandwidths, $R_i(t)$ represents the average throughput of user i at TTI t (Equation (3.3)) and N is the total number of users.

3.1.8 Channel-Dependent Earliest Due Deadline (CD-EDD) Algorithm

Channel-Dependent Earliest Due Deadline (CD-EDD) is a packet scheduling algorithm introduced to provide QoS for a variety of time sensitive traffic in a mobile cellular system [35]. Same as MLWDF and EXP algorithms, the CD-EDD algorithm allocates the packets of data considering average throughput, instantaneous data rate and packet delay information. In the scheduling, if the average throughput and instantaneous data rate of each user are the same, the CD-EDD gives the transmission priority to a user with most urgent HoL delay, whereas MLWDF and EXP algorithms are most likely to provide transmission priority to the longest delay in the buffer of the base station. Equation (3.11) represents the formula of CD-EDD algorithm as below:

$$\mu_i(t) = \alpha_i * \frac{r_i(t)}{R_i(t)} * \frac{W_i(t)}{T_i - W_i(t)}. \tag{3.11}$$

Where $\mu_i(t)$ is the priority of user i at TTI t, α_i is the QoS requisite of user i (same as equation (3.8)), $r_i(t)$ is the instantaneous data rate in whole bandwidth, $R_i(t)$ is the average throughput of user i at TTI t, $W_i(t)$ is HoL packet delay of user i at TTI t and T_i represents the buffer delay threshold of user i.

3.2 Performance evaluation of the DPS, PF, Max-Rate and RR algorithms in LTE System

In this subsection, the performance of the DPS, RR and PF scheduling algorithms is studied in downlink LTE system. Their performance is measured in terms of fairness, throughput and PLR metrics. Single carrier scheduling algorithms in time domain are designed to assign all resources to a user in each time interval. DPS, RR and PF algorithms have good results in a single carrier mobile system. Since downlink LTE system is a multicarrier cellular system developed for multimedia transmission, some modifications are needed to adapt single carrier algorithms in LTE scheduling systems. Packet scheduling in downlink LTE systems is performed in both time and frequency domains. Several users compete on a number of available PRBs in each 1 ms TTI. In the next chapter, we will discuss the assignment methods in detail.

A number of PRBs may be allocated to a user while a PRB may be assigned to only one user in each time interval (as described in the previous chapter). In the following subsections, multi carrier time frequency Max-Rate, RR and PF algorithms which are adapted from single carrier algorithms are discussed.

In Section 3.2.1 the modifications to change the single carrier RR, PF and DPS algorithms to the downlink LTE system are described. Simulation results are provided in Section 3.2.2.

3.2.1 Modifications to the Famous Scheduling Algorithms in Downlink LTE Systems

In a Downlink LTE System, a number of PRBs are available for sharing among users in each time interval. We modified the single carrier Max-Rate algorithm as below to be adaptable in multicarrier system.

$$\mu_{ij}(t) = r_{ij}(t) \tag{3.17}$$

The aim of this algorithm is to maximize $\mu_{ij}(t)$ with selecting the most suitable candidate among users to receive packets on PRB_j at time interval t. Where $\mu_{ij}(t)$ is the priority of user i on $PRB\,j$ at TTI t and $r_{ij}(t)$ is the instantaneous data rate of user i on $PRBj$ at TTI t. The multicarrier PF algorithm chose the most suitable user from available users in each time interval to receive packets on PRB_j in order to maximize $\mu_{ij}(t)$ in Equation (3.18):

$$\mu_{ij}(t) = \frac{r_{ij}(t)}{R_i(t)} \tag{3.18}$$

$$R_i(t+1) = \left(1 - \frac{1}{t_c}\right) R_i(t) + I_i(t+1) * \frac{1}{t_c} * rtot_i(t+1) \tag{3.19}$$

$$rtot_i(t+1) = \sum_{j=1}^{RB_{max}} I_{i,j}(t+1) * r_{ij}(t+1) \tag{3.20}$$

$$I_{ij}(t+1) = \begin{cases} 1 & if\ packets\ of\ user\ i\ are\ scheduled\ at\ scheduling\ interval\ t+1 \\ 0 & if\ packets\ of\ user\ i\ are\ not\ scheduled\ at\ scheduling\ interval\ t+1 \end{cases} \tag{3.21}$$

where $r_{ij}(t)$ is the instantaneous data rate of user i on RB_j at TTI t, $R_i(t)$ is the average throughput of user i at scheduling interval t, t_c is a time constant, $rtot_i(t+1)$ is the total data rate across selected PRBs for user i at TTI $t+1$, $I_{ij}(t+1)$ is the indicator which shows whether user i is selected to transmit packets on RB_j at TTI $t+1$ or not, RB_{max} is the total available RBs and $\mu_{ij}(t)$ is the priority of user i on PRB j at TTI t.

Finally, the RR algorithm users transmit their packets on allocated RBs in a cyclic trend.

3.2.2 Performance of Four Famous Algorithms in LTE system

The performance of DPS, Max-Rate, PF, RR algorithms in LTE cellular system is investigated in this section. The downlink LTE system model described in the previous chapter has been used for simulation. Four additional assumptions that we have made are described as follows: i) The CQI report in the eNB collected from the users is non-erroneous and perfect. ii) Users receive all packets which are correctly decoded. iii) The buffer of data packets is assumed to have infinite capacity. iv)The MCS is assumed to be fixed, without considering the report of Channel Quality Information.

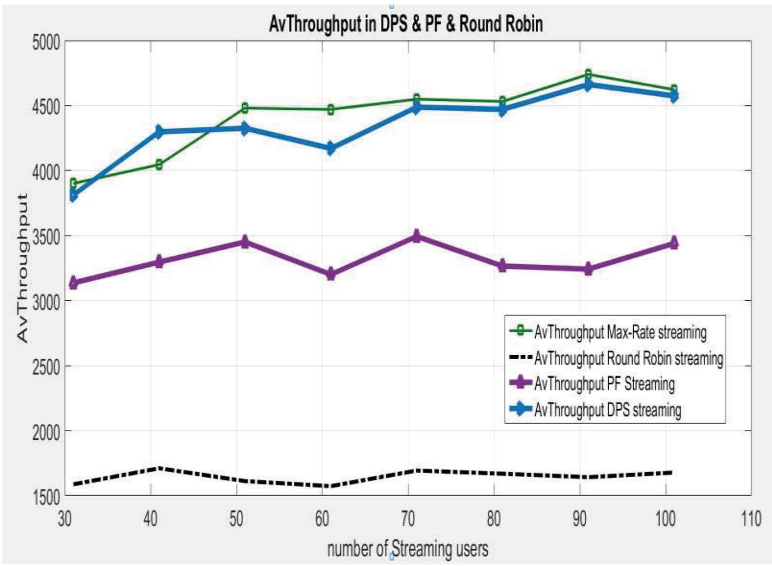

Figure 3.1: Average throughput vs. number of users

The average throughput of the system for multi-carrier Max-Rate, RR, PF and DPS algorithms has been shown in Figure 3.1 with increasing number of users. It can be seen that the Max-Rate and DPS algorithms have the best throughput results, because they focus on achieving Max-Rate and minimum delay, respectively. On the other hand, the RR algorithm has the least average throughput of the four algorithms because it only considers the fairness in its scheduling algorithm. In Figure 3.2, the fairness of multi-carrier Max-Rate, RR, PF and DPS algorithms has been presented. RR has the best fairness performance among these four algorithms and Max-Rate has the worst fairness among all because it does not consider the average throughput of the users or packet delays to control fair allocation amongst them. It only uses maximizing the rate which results in increasing the average throughput of users. All four algorithms mentioned above have an increasing trend in fairness with increasing the number of users. PF and DPS come in the second and third places respectively after the RR algorithm.

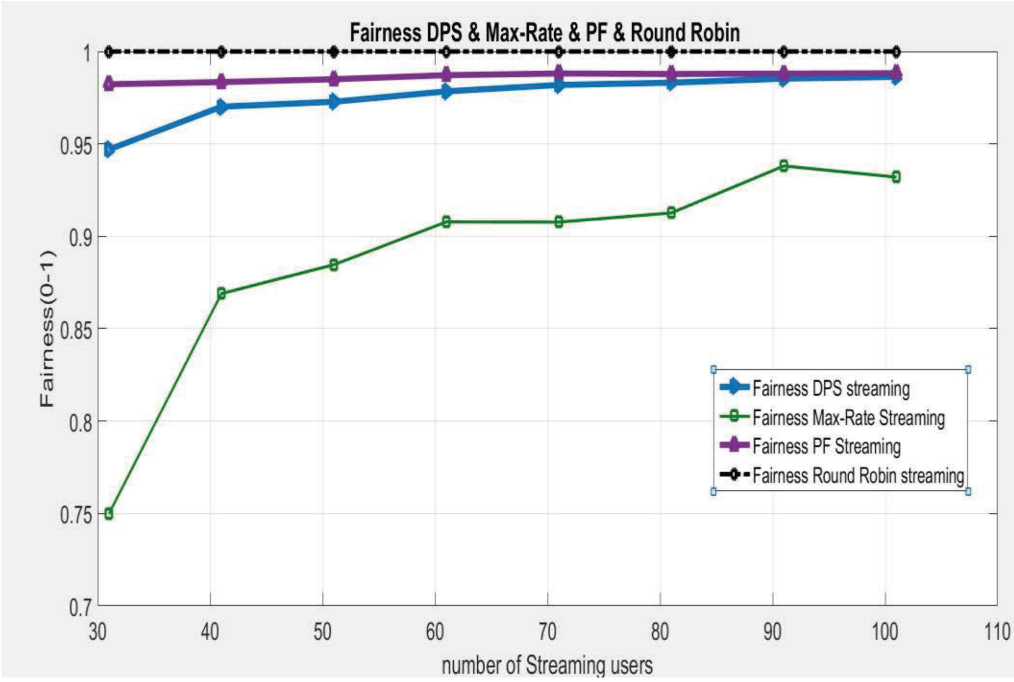

Figure 3.2: Fairness vs. number of users

By looking at Figure 3.3 which illustrates the PLR (Packet Loss Ratio), it can be seen that the most PLR belongs to Max-Rate algorithm because it does not consider the fairness and the delay of the packets in the scheduling. The Second most PLR belongs to the PF algorithm that is considerably smaller than Max-Rate because it uses the average throughput of the users as a fairness term in RB allocation. DPS and RR have zero PLR because in these scheduling algorithms, delay and turn of the users have been included respectively.

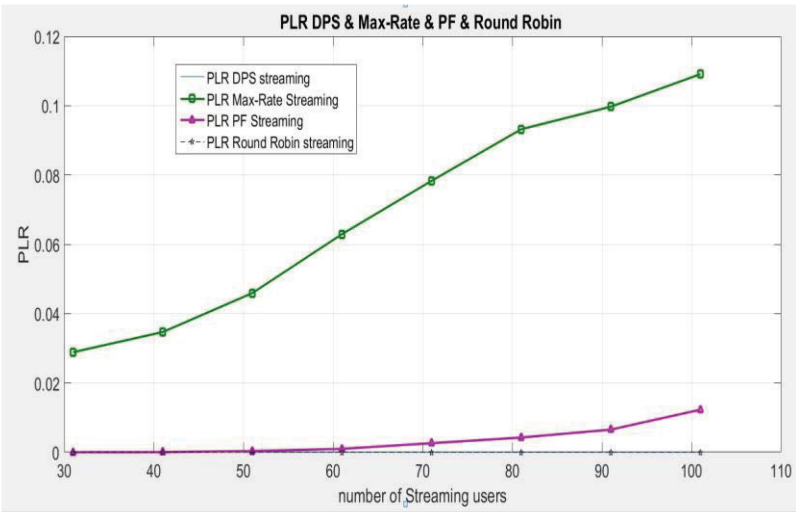

Figure 3.3: PLR vs. number of users

3.3 Summary

This chapter investigated a number of famous scheduling algorithms in multicarrier downlink LTE systems. The performance of well-known algorithms Max-Rate, RR, PF and DPS were simulated in the MATLAB environment. To use the four famous algorithms for LTE systems, a number of adaptations were proposed. Results demonstrated that the DPS and PF algorithms are the best algorithms for providing the trade-off among average throughput, fairness and packet loss ratio.

Chapter 4 Scheduling Algorithms Based on QoS Requirements

One of the most important advancements in mobile cellular systems is the improvement in data rate which is due to the increasing demand for higher Quality of Service in multimedia services. Multipath fading, shadowing, inter-cell interference and mobility of users result in unpredictable wireless channels which affect resource allocation and QoS requirements. Several popular scheduling algorithms were investigated in Chapter 3. LTE systems exploit frequency and multiuser diversity to overcome the frequency-selective effect on wireless channels. In frequency diversity algorithms, multiple subcarriers are used to transmit data simultaneously. Multi-user variety in LTE resource allocation indicates that an inappropriate PRB for one user could be acceptable for other users. In [36-39], the performance of packet scheduling algorithms with frequency diversity over time domain scheduling algorithms has been investigated.

In this chapter, the frequency-time domain scheduling algorithms based on QoS needs are considered. These algorithms are categorized as follows: RAA (Resource Allocation and Assignment), Matrix-Based Scheduling and JTFDS (Joint Time and Frequency Domain Scheduling).

This chapter consists of the following sections: Section 4.1 describes Resource Allocation and Assignment algorithms and presents some examples. In Section 4.2, Matrix-Based Scheduling Algorithm is discussed. In Section 4.3, QoS-Oriented Joint Time and Frequency Domain Scheduling is described. Some relevant works in LTE packet scheduling have been discussed in Section 4.4. In Section 4.5, Utility Based Resource Allocation Algorithms and the new proposed algorithm LOG-MLWDF in LTE system has been investigated. The performance evaluation and comparison of simulation results with three other popular scheduling methods are discussed in the final section.

4.1. RAA (Resource Allocation and Assignment) Algorithm

The RAA (Resource Allocation and Assignment) scheme was proposed for efficient resource allocation with the goal to deploy a trade-off between throughput and fairness. This algorithm consists of two stages: Stage 1 (resource allocation) and Stage 2 (resource assignment). In Stage 1, the numbers of RUs which are allocated to each user are calculated on the basis of transmission requirements in each TTI. If the total number of allocated RUs is more than the available number of RUs in one TTI, the difference will be removed from the users' allocation list until the allocation number is equal to the available number of RUs. The priority of choosing users for each RU is computed based on the scheduling algorithm in Stage 2. The calculated number of RUs for every user in Stage 1 is assigned to users based on the scheduling algorithm and its priority. The scheduling algorithm is based on transmission needs to maximize the usage of RUs. Figure 4.1 presents the flow chart of the Resource Allocation and Assignment algorithm.

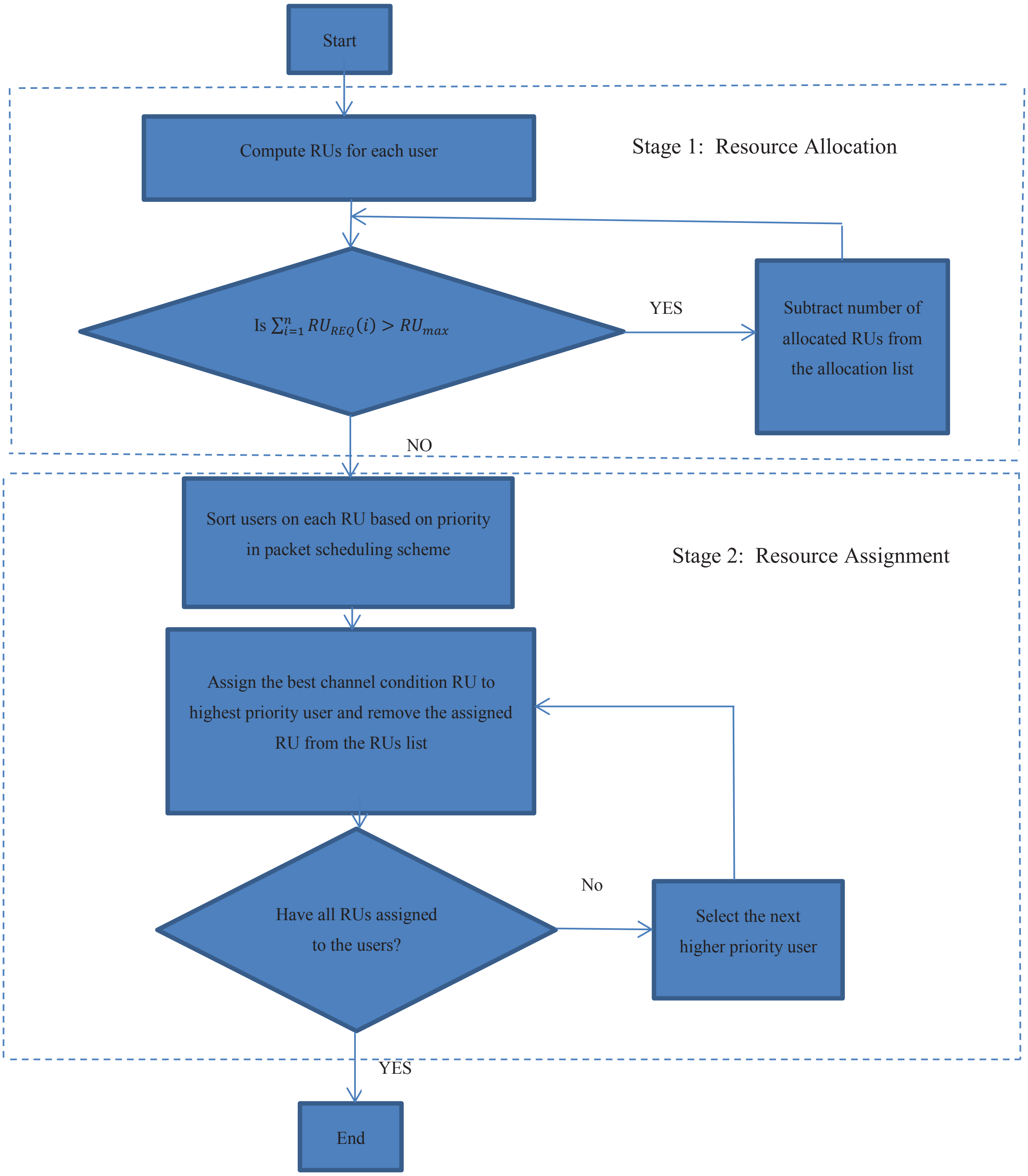

Figure 4.1 : Flow chart of RAA algorithm in each time interval

4.1.1. Queue Aware Scheduling (QAS) Algorithm

In [40, 41], an RAA algorithm has been introduced, which is referred to as Queue Aware Scheduling (QAS) in this work. In each TTI, the scheduling scheme of QAS has two phases. In Phase 1, users who have data packets in their buffer get one RU (Resource Unit) in each TTI in turn. If the total number of allocated RUs is less than the available RUs in each time interval, then the remaining RUs are allocated to users according to the equation below:

$$n_i(t) = n_i(t) + RU_{rem} * \left\lfloor \frac{B_i(t)}{\sum_{k=1}^{k=N} B_k(t)} + 0.5 \right\rfloor. \tag{4.1}$$

Where RU_{rem} is the remaining number of RUs, $B_i(t)$ is the length of the queue in eNB buffer of user *i* at time interval *t*, $n_i(t)$ is the total number of RUs that are allocated to user *i* at time interval *t*, and *N* is the total number of users. Phase 2 is the resource assignment phase: $n_i(t)$ RUs are assigned to user *i* based on channel state information and the priority of the user *i* ; then the assigned RUs will be removed from the list of available RUs. Phase 2 will be completed after all users have been assigned to the number of allocated RUs in Phase 1. In the next TTI, the user priorities are reassigned. The highest priority user in the current TTI will be scheduled as the lowest priority user in the next interval. This way, the algorithm tries to achieve good fairness amongst its users.

4.1.2 Opportunistic and Delay Sensitive (ODS) Algorithm

In [42], a packet scheduling algorithm consisting of two stages of resource allocation and assignment techniques has been introduced. The algorithm is named as ODS (Opportunistic and Delay Sensitive) algorithm is in this work. The flow chart of this algorithm has been given in Figure 4.2. The resource allocation of ODS consists of three phases in each interval. Phase1 calculates the number of RUs as $n_i(t)$for user *i* at scheduling interval *t* as follows:

$$n_i(t) = \left\lceil \frac{\bar{\mu}_i(t)}{\frac{1}{N}\sum_{k=1}^{k=N} \bar{\mu}_k(t)} * \frac{R_i(t)}{\frac{1}{N}\sum_{k=1}^{k=N} R_k(t)} \right\rceil. \tag{4.2}$$

Where, the number of RUs that are allocated to user *i* at scheduling interval *t* is $n_i(t)$. $\mu_i(t)$ is the maximum possible data rate (channel capacity) on all RUs for user *i* at TTI of *t*, $\bar{\mu}_i(t)$ is the average subcarrier capacity of i^{th} user, $R_i(t)$ is the average traffic rate of user *i* at TTI *t*, and *N* is total number of users. After the allocation, there are two possible phases. In the first phase, if the number of allocated RUs is less than the available RUs, the remaining RUs should be allocated to the users according to the equations given below:

$$S' = S - \sum_{N_t} n'_i(t) \tag{4.3}$$

$$n_i(t) = n_i'(t) + \left\lceil S' \frac{\frac{max\{1, V_i(t)\}}{d_i(t)}}{\sum_{j \epsilon N_t} \frac{max\{1, V_j(t)\}}{d_j(t)}} \right\rceil \tag{4.4}$$

$$d_i(t) = T_i - W_i(t) \tag{4.5}$$

where S' is the remaining unused RUs, $V_i(t)$ is the number of lost packets by the i^{th} user up to the time t and $d_i(t)$ is the remaining time to expire for i^{th}user's Head of Line packet which is the difference between the deadline T_i and HoL packet delay $W_i(t)$ up to time t.

The second phase is when the number of allocated RRUs is more than the available RUs ($RRU_{allocated} > RRU_{max}$). Therefore, we should reduce the number of allocated RRUs. In this step, at first users are sorted based on the decreasing Head of Line packet delay, where a user with HoL packets that have more time to be expired are more probable to reduce their number of allocated resource blocks. Then, the number of RRUs is reduced from the sorted users in turn and an updated number of allocated RRUs are calculated. The iteration is continued until $RRU_{allocated} = RRU_{max}$. In the resource assignment technique, the users are first sorted based on the number of packets with overdue deadlines; and then the user with higher number of Packet Loss Ratio (PLR) is given higher priority. The proposed scheduling algorithm assigns the RRUs with best channel conditions to the user with highest priority and removes the assigned RUs from the available RRUs list. This assignment repeats until all RUs are allocated to the users.

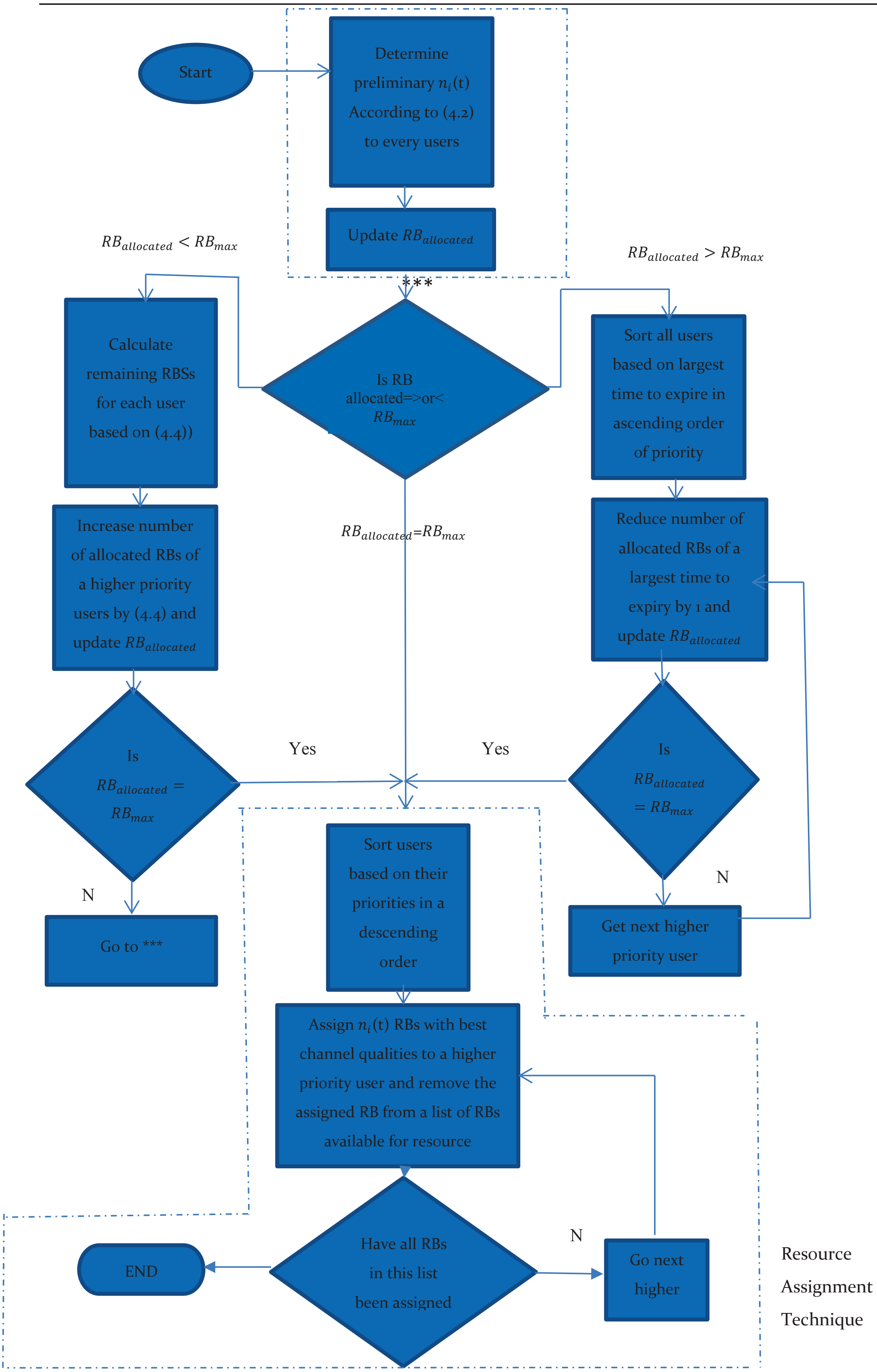

Figure 4.2 : The Opportunistic and Delay Sensitive (ODS) Algorithm

4.1.3. DFS (Delay First Scheduling) Algorithm

Delay First Scheduling (DFS) algorithm is introduced in [43] based on the MLWDF algorithm (studied in Chapter3) to support services with different QoS requirements. This scheduling is based on giving the highest priority to HoL packets and consists of two Steps: resource allocation and assignment. In the first Step, a fixed number of RBs calculated on the basis of user's minimum data rate referred to as step-resource-number is given to each user. The resource assignment technique consists of 4 Steps. In the first Step, the RRU priority list is determined for each user. The highest priority for user i's RRU list consists of RRUs with best channel quality. In Step 2, the priority of each user (without empty buffer) is calculated on the basis of maximizing the equation (4.6).

$$\mu_i(t) = \alpha_i * W_i(t) * \frac{r_{avg_i}(t)}{R_i(t)} * \frac{PLR_i(t)}{\delta_i}. \tag{4.6}$$

Where α_i is the QoS requirement of user i (Equation (3.8)), $W_i(t)$ is the HoL packet delay of user i at time interval t, R_i is the average throughput of user i at time interval t (Equation (3.3)), δ_i is the PLR threshold of user i and $PLR_i(t)$ is the user i's PLR at time t. Then in Step 3, the fixed calculated number of RUs is assigned to each user by priority based on delay of HoL packets and channel condition calculated in Step1 and Step2. In Step 4, the HoL packet delay of users is updated and the allocated RUs are removed from the available RRU's list. Steps 2 to 4 are repeated until all available RRUs have been assigned to users or there are no more data packets in the buffers of users.

4.1.4. Quality of Service-Driven Resource Allocation (QRA) Algorithm

To assign radio resource units to multimedia data traffic with different Quality of Service requirements, an RAA algorithm was developed in [44]. This algorithm is called QRA (Quality of Service -Driven Resource Allocation) in this work. To assign scheduling weight proportions to multimedia data traffic the Hebbian learning mechanism has been used. This algorithm is able to decrease the PLR and average delay; therefore, it can guarantee the Quality of Service requirements of real time traffic and can provide good QoS for non-real time traffic by setting the trade-off between increasing fairness and throughput.

The First Step in this algorithm is computing the number of Resource Units for real time traffic. Computing and comparing the change in the rate of packet drops in each interval and saving the variation of data in RAP (Resource units Allocation Priority) vectors are done according to the

Hebbian learning mechanism. If the packet drop rate variation is positive, the weight related to the real time traffic is recalculated by the equation below [44]:

$$We_{RT}(t) = \begin{cases} We_{RT}(t-1) + \gamma & if\ PDR_{RT}(t)\ > PDR_{RT}(t-1) \\ We_{RT}(t-1) - \gamma & else \end{cases} . \quad (4.7)$$

Where, $We_{RT}(t)$ is the weight of allocation for real time users at time interval *t* and γ is the rate of learning. $n_{RT}(t)$ is the number of Resource Units calculated based on Equation (4.8)

$$n_{RT}(t) = RU_{total}(t) * \left(\alpha(t-1) + We_{RT}(t-1)\right) \quad (4.8)$$

where $RU_{total}(t)$ is the number of all available Resource Units at time interval t, $\alpha(t-1)$ is the ratio of the total Resource Units allocated to real time users at time interval *t* and $We_{RT}(t-1)$ is the weight of allocation for real time users at simulation time *t*.

Therefore, the number of RUs allocated to non-real time users at time interval *t* referred to as $n_{NRT}(t)$, is calculated according to following formula:

$$n_{NRT}(t) = RU_{total}(t) - n_{RT}(t) \quad (4.9)$$

The Second Step is the resource assignment phase which consists of 3 Stages in each simulation interval. Stage1 computes the priority of each user. The priorities of real time and non-real time users are calculated according to Equation (4.10) and (4.11):

$$\mu_{RTi}(t) = \frac{w_{wait\ i}(t)}{T_i} * \left(H_i(t)\right)^2 + \left(B_i(t)\right)^2 \quad (4.10)$$

$$\mu_{RTi}(t) = \frac{w_{wait\ i}(t)}{T_i} * \frac{Thrnrt_i}{R_i(t)} * \left(H_i(t)\right)^2 \quad (4.11)$$

where $w_{wait\ i}(t)$ is user *i*'s waiting time duration from the previous scheduling time until now, T_i is the delay threshold of real time and non-real time users buffer in Equation (4.10) and (4.11), $H_i(t)$ is the average channel gain for user *i* at time interval *t*, $B_i(t)$ is the buffer length of user i at TTI t, $Thrnrt_i$ is the throughput needs of non-real time user i, and $R_i(t)$ is the average throughput of user *i* at time interval *t*.

In Stage 2, the priority of users for each traffic type is sorted and in Stage 3, RUs with best channel state information is assigned to users of each traffic type on the basis of sorted users in Stage 2; then the allocated Resource Unit is removed from the available RU list. Stages 2 and 3 are repeated until there are no data packets in the buffer or no Resource Units are available.

4.2. Scheduling Algorithm Based on Matrix

In [44-47], an efficient scheduling algorithm based on channel matrix computation is proposed. Matrix based scheduling algorithms consist of two stages: channel matrix computation and scheduling on the basis of calculated matrix. The dimension of channel matrix consists of rows equal to the number of Radio Resource Units and columns equal to the number of active users. The elements of the matrix include the priority of each user on each Radio Resource Unit at TTI *t* are determined based on the scheduling algorithm. The channel matrix with a dimension of $K{*}RRU_{max}$ used in the scheduling algorithm is shown below:

$$\begin{bmatrix} E_{1,1}(t) & \cdots & E_{1,RRU_{max}}(t) \\ \vdots & \ddots & \vdots \\ E_{k,1}(t) & \cdots & E_{kRRU_{max}}(t) \end{bmatrix} \tag{4.12}$$

Stage1 consists of 3 steps: The maximum element of the matrix is chosen in the First step. This indicates that the highest priority belongs to user i on RRU_j at time interval *t.* Based on every user's requirement, the number of Radio Resource Units is calculated for each user in order to serve the packets of the user in step 2. Column *j* is removed from the matrix when RRU_j is allocated to one user (when it is used by user *i* at time *t*), and if no data packets exist in the buffer of user *i*, the *ith* row in the matrix would be removed in step 3. Then the channel matrix will be updated based on the scheduling information. These three steps are repeated until the matrix is empty.

In [45], the authors introduced an extension of the PF algorithm based on a Matrix-based method, namely OFPF (OFDMA Frame-Based Proportional Fairness). This algorithm proposed a hybrid resource allocation and assignment based on a matrix-based scheme and modified the average throughput in the PF scheduling algorithm with its scheduling technique. This scheduling technique tries to provide an acceptable trade-off between fairness and throughput amongst users. The OFDMA Frame-Based Proportional Fairness algorithm includes four Steps: In Step1 the algorithm computes a $RRU_{max} * K$ channel matrix for a system consisting of K users and Radio Resource Units (RRU_{max}) as shown in (4.12). Each element of the matrix is calculated based on the modified PF algorithm (Equation (4.13)), which determines the priority of every user on each RRU at time interval *t*. The modified PF algorithm formula is presented below:

$$\mu_{ij}(t) = \frac{r_{i,j}(t)}{R_{ofpfi}(t)} \tag{4.13}$$

$$R_{ofpfi}(t+1)\begin{cases} \left(1-\frac{1}{t_c}\right)R_{ofpfi}(t) + \frac{1}{t_c} * r_{i,j}(t+1) & if\ B_i(t+1) > 0. \\ R_{ofpfi}(t) & if\ B_i\ (t+1) = 0 \end{cases} \tag{4.14}$$

Where $r_{ij}(t)$ is the instant data rate of user *i* on RRU_j at time interval *t*, $R_{ofpfi}(t)$ is the modified average throughput of user *i* at time interval *t*, $B_i\ (t+1)$ is the buffer length of user *i* on RRU_j at time interval t in bits and t_c is the invariable time constant. In the Second Step, the algorithm chooses the maximum element in the channel matrix that has higher priority referred to as $E_{ij}(t)$. In the next Step, (Step 3) the jth column is removed from the matrix because RRU_j has been allocated to user i. If the buffer of user i become empty, the ith row is deleted. In the last step the whole matrix is updated according to Equation (4.13). Steps 2 to 4 are repeated until there is no available RRU or the channel matrix is empty.

4.3 QoS-Oriented Joint Time and Frequency Domain Scheduling

To decrease the complexity of computation and improve the scheduling performance, joint frequency and time domain algorithms have been studied in [48, 39, 49, 50]. These algorithms consist of two phases:

1) A time domain scheduler prioritizes users to a subset of candidates based on a time domain scheduling algorithm. These are used as an input for the frequency domain scheduler.

2) The frequency domain packet scheduler selects the priority of the users on each RRU and allocates the RRU according to the selection.

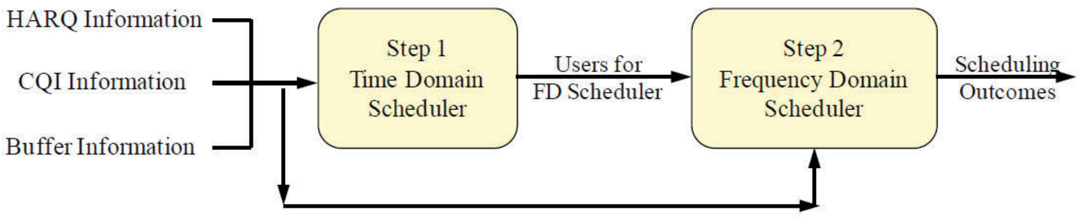

Figure 4.3 : JTFDS scheme Structure

The JTFDS structure is presented in Figure 4.3. To reduce the scheduling complexity and improve the performance, Time Domain (TD) and Frequency Domain (FD) schedulers are used respectively. Using the TD scheduler at the beginning reduces the complexity. In [48, 39], it is proven that the complement of two different schedulers with different aims is the reason for better performance.

In [48], an algorithm (named the Oriented Time-Frequency Scheduling (OTFS) algorithm) with the purpose of controlling the fairness based on JTFDS scheme is introduced. This algorithm categorizes the users in two groups based on the channel quality and uses several scheduling policies in one interval to provide optimal trade-off between fairness and throughput. In time domain, the OTFS algorithm consists of two stages: In Stage 1 the users are categorized into two groups pre-defined by the Target Bit Rate (TBR). The users who are below the Target Bit Rate are categorized as Group 1 and prioritized over users from Group 2 who are above the Target Bit Rate. In Stage 2, the users in Group 1 and Group 2 are sorted according to the BET algorithm (as discussed in Section 3.1.4) and PF algorithm, respectively. When the TBR is near zero the scheduling technique is equivalent to the PF algorithm. Using the PF algorithm results in better throughput (because of considering the channel quality information in the scheduling approach). On the other hand, if the TBR tends to infinite numbers, the scheduler would use the BET algorithm which results in better fairness among all users. So the amount of predefined TBR controls the use of the PF and BET algorithms. The PFS (Proportional Fair Scheduled) algorithm was used in the frequency domain to collaborate with the time domain scheduler. By using Equation (4.15) the scheduler gives the highest priority to users who maximize the $\mu_{ij(t)}$ at TTI *t*.

$$\mu_i(t) = \frac{r_i(t)}{R_{-sch_i}(t)} \tag{4.15}$$

$$R_{-sch_i}(t+1) = \left(1 - \frac{I_i(t+1)}{t_{-sch_c}}\right) R_{-sch_i}(t) + \frac{I_i(t+1)}{t_{-sch_c}} * r_i(t+1) \tag{4.16}$$

$$I_i(t+1) = \begin{cases} 1 \ \ if\ packets\ of\ user\ i\ are\ scheduled\ at\ TTI \quad\ t+1 \\ 0\ if\ packets\ of\ user\ i\ are\ not\ scheduled\ at\ TTI\ t+1 \end{cases} \tag{4.17}$$

where $r_i(t)$ presents as instantaneous data rate along the whole bandwidth of user *i* at time interval *t*, $R_{-sch_i}(t)$ is the average calculated throughput of user *i* at simulation time *t*, $I_i(t+1)$ is a number that shows if user *i* is scheduled at time interval *t* and t_{-sch_c} is a time invariable equal to 30 [51]. The frequency domain scheduler chooses a user who has a maximum $\mu_{i,j}(t)$ by the Wei_i for transmission on each PRB_j at time interval *t*.

$$Wei_i = max\left(1, \frac{TBR}{R_i}\right) \tag{4.18}$$

where TBR is the target bit rate and $R_i(t)$ is the average throughput of user *i* at time interval *t*.

4.4. Relevant works on LTE packet scheduling

4.4.1. Coupled Throughput-Fairness Delay Scheduler

The key idea of [52] is to satisfy the Quality of Service requirements of real-time services and to provide throughput and fairness for non-real-time traffic applications. A new classification for packets as “urgent” and “non-urgent” has been introduced in order to find a solution for optimization between the QoS of RT users and the throughput and fairness of NRT users. The key point of the article is that RT traffic should not be scheduled with a superior priority to the

NRT traffic unless they are in the "urgent" state to be transmitted. The urgent state defines when the delay of packets is going to be more than the permissible threshold. The authors compare their methods with four other algorithms in terms of packet delay, packet dropped rate and throughput. Simulation results show that the proposed algorithm increases the QoS for real-time users and the throughput and fairness for non-real-time users have also been improved.

4.4.2. Joint Real-time and Non-real-time Packet scheduling

In [53], the authors consider joint Real Time (RT) and Non-Real Time (NRT) packet flows and resource block allocation in OFDMA wireless networks. In the conventional methods the data traffic categorized to RT and NRT packets. The authors claim this way is too conservative therefore it should be reconsidered and reengineered. They propose a novel joint flow in a common pool of RBs. Mean bit-rate, mean queue-length, instantaneous queuing delay and channel information have been used to satisfy the requirements. An input-output approach for bit-rate performance of mixed RT and NRT queues is also proposed. The simulation results show that the unified method has a higher bit-rate compared to several baselines.

4.4.3. Performance Comparison of scheduling algorithms in ns3

A performance comparison analysis of several scheduling algorithms named Maximum Throughput Scheduler (MTS), Blind Equal Throughput Scheduler (BET), Proportional Fair Scheduler (PFS), Channel-QoS Aware (CQA) and Priority Set Scheduler (PSS) for UDP and TCP traffic sources is investigated in [54]. Both the time and frequency domains that include both flat and frequency selective channels are considered in this work. The analysis has been performed in terms of gain, fairness, packet service time, cell capacity and throughput. Time and frequency domain versions of Fair Throughput Guarantees Scheduler (FTGS) have been implemented in ns-3 in this work. It is evident that the FD version of the FTGS in comparison with the TD version can significantly decrease the time of inter-scheduling. The analysis presents a good tradeoff between cell capacity and fairness for two traffic sources (TCP and UDP) whilst the proposed algorithm FTGS also has good performance to some extent in the wrong estimation of the SINR.

4.4.4. MIMO Packet Scheduling in LTE

In [55], the Frequency Domain Packet Scheduling (FDPS) problem with the multiple input-multiple output (MIMO) approach has been investigated. To maximize the Proportional Fair (PF) rule which is extended to the time and frequency domain for each user in each transmission time interval (TTI), transmission diversity or special multiplexing on the MIMO mode is considered.

It has been proven that the Single-User MIMO (SU-MIMO) FDPS problem deployed in the LTE system is an NP-hard problem. Approximations of the two algorithms have been developed to solve the problem with verified bounds. These two algorithms mentioned as Alg1 and Alg2 outperformed 1×2SIMO FDPS in the 24-35% range. To optimally allocate time-frequency RBs to users based on the PF scheduling algorithm in MIMO Orthogonal Frequency Division Multiple Access a new algorithm is proposed in [56]. The FDPS problem has been modeled as a unique binary linear programming problem in this work and has been proven to be uni-modular. The problem has then been solved by linear problem solvers. The non-similarity of MCS for different RBs which are allocated to a user, are considered in this work.

4.4.5. Carrier Aggregation in LTE-Advanced Networks

The carrier aggregation technique has been incorporated by Long Term Evolution-Advanced to increase the transmission rate. In [57], the authors have considered that the assigned Component Carriers (CCs) in each UE can be changed and CCs can be reassigned to each UE at each TTI. The resource allocation problem in the LTE-A system has been formulated under the coding and modulation scheme constraints as an NP hard problem. A scheme called greedy-based has been proposed to increase the throughput of the system considering the proportional fairness among all UEs. Simulation results show that the proposed method outperforms the previously studied methods in terms of throughput. Almost 50% of the optimum answer can be guaranteed by this method. In [58] the authors studied how to optimize resource allocation in LTE-Advanced system with aggregation of multiple Component Carriers including assigning the CCs to each user and multiplexing multiple users for each CC. Different carrier load balancing methods in layer-3 have been considered and their Component Carriers assignment to each user have been investigated in layer-2. The Round Robin (RR) load balancing method outperforms the Mobile Hashing (MH) method for a low percentage of LTE-Advanced users and a low number of users. A cross-CC packet scheduling (PS) algorithm at Layer2 has been proposed in this article which improves the coverage and fairness of the system in comparison with the independent Packet Scheduling (PS) per CC. A 90 % gain ratio in coverage without losing the cell throughput has been provided by this method compared to the independent scheduling algorithm per CC.

4.4.6. Smart Downlink Scheduling Algorithm with Hard Handoff

When a mobile user moves from one cell to another, it is referred to as hard handoff (HO) procedure. It is one of the important issues in delivering multimedia streaming traffic in LTE systems. The problem consists of the deadline violation, fairness and service degradation caused by it. To improve multimedia traffic transmission, a smart QoS-driven downlink scheduling scheme has been proposed in [59] which considers HO. The design considers the following QoS metrics:

1) the delay deadline of voice-over internet 2) deadline of video traffic packets 3) service degradation caused the by hard HO procedure.

The design objectives have been achieved by considering the following three control modules:

1) a control module to control the transmission delay has been used to ensure the arrival of different types of traffic.

2) a module to control the handoff when the user moves from one cell to another.

3) a module for resource allocation to assign the traffic data to RBs.

From the simulation results in [59], it can be seen that the proposed scheme has satisfied results for multimedia traffic services.

4.5 Utility Based Resource Allocation Algorithm

4.5.1 Utility Function

In economics, the idea of using Utility Functions has been well investigated and used to solve the concerns about efficient and fair resource allocation. Utility functions have been used to balance efficiency and fairness for cross layer optimization. The radio frequency resources (used by a user) are mapped to a real number by utility function. In cross layer optimization, utility functions have been used to balance throughput and fairness. Utility functions act as an optimization goal for the adaption of MAC and physical layer techniques. Providing a trustable data rate is one of the essential factors to satisfy QoS requirements of users. Therefore, the utility function is a function of data rate *(r)*. The scheduling algorithms, which use CSI and utility functions based on rate, are not suitable for delay sensitive users. To provide satisfactory QoS for users with latency requirements, average delay-based utility functions are developed. In this work, four different scheduling algorithms based on utility functions for different types of traffic in 3GPP LTE network have been investigated.

4.5.1.1 Algorithms based on Utility Function

4.5.1.1.1 LOG-MLWDF Algorithm

LOG-MLWDF [61] is an extension of the MLWDF algorithm in which the logarithmic weight of the average delay has been added to MLWDF. The extended LOG-MLWDF is given by:

$$\mu_i(t) = \gamma_i r_i(t)\left(W_i + log\left(1 + Wave_i(t)\right)\right)^{1.5} \tag{4.19}$$

Where:

$$Wave_i(t) = \frac{1}{T}\sum_{i\epsilon N} W_i\,(t). \tag{4.20}$$

4.5.1.1.2 Max Delay Unit (MDU) Algorithm

MDU [62] deploys the idea that the scheduling metric can be calculated based on the average waiting time to improve the quality of service. The formula of the MDU scheme formula is given below:

$$\mu_i(t) =_{RB^n, i\in A^n, s\in B^n} \sum_{i\in A^n, s\in B^n} \frac{|U'_{i,s}(w_i)[n]|}{\lambda_{i,s}} r_i[n], \tag{4.21}$$

which is subject to the following conditions:

$$\cup_{i\in A^n} RB_i^n \subseteq K,$$

$$RB_i^n \cap RB_j^n = \emptyset, i \neq j\ , \forall i,j \in A^n.$$

$\lambda_{i,s}$: Average arrival data rate of user i with the traffic type s,

$r_i[n]$: Data transmission rate of user i at time slot n,

A^n: Set of queues at time slot n,

B^n: Set of the service types.

$\mu_i(t)$: The priority of user i at time interval t

Three types of traffic such as voice, streaming and best effort have been considered in this work. A number of consecutive functions are used in MDU scheduling which are provided in the following formulas as:

marginal function for voice users:

$$\left|U'_{i,v}(w_i)\right| = \begin{cases} w_i, w_i \leq 2.5ms \\ w_i^{1.5} - 2.5^{1.5} + 2.5, w_i \geq 2.5ms \end{cases} \tag{4.22}$$

marginal function for streaming users:

$$\left|U'_{i,s}(w_i)\right| = \begin{cases} w_i^{0.6}, w_i \leq 5ms \\ w_i - 5 + 5^{0.6}, w_i \geq 5ms \end{cases} \tag{4.23}$$

marginal function for best effort users:

$$\left|U'_{i,s}(w_i)\right| = \begin{cases} w_i^{0.5}, w_i \leq 5ms \\ 5^{0.5}, w_i \geq 5ms \end{cases} \tag{4.24}$$

4.5.2 Dynamic Subcarrier Allocation

The DSA (Dynamic Subcarrier Allocation) has been considered in this work as the resource allocation scheme to dynamically allocate RBs to the users in order to maximize the average utility function as below:

$$Max \frac{1}{M} \sum_{i=1}^{M} U_i(r_i) \tag{4.25}$$

Subject to:

$$\bigcup_{i \in M} RB_i = [0, \beta],$$

and $RB_i \cap RB_j = \emptyset,\ i \neq j\ , \forall i, j \in M.$

The assignment of one RB does not have any impact on the assignments of other RBs. The utility functions can be linear or nonlinear functions of rate. The DSA method can be used in the following formula when the utility functions are linear:

$$m(k, n) = argmax_{i \in \mu} \{U'_i C_i[k, n]\}. \tag{4.26}$$

Where, $m(k, n)$indicates that $RB_k$is assigned to user m at time slot n, and $C_i[k, n]$ is the feasible data rate for RB_k at time slot n, and it is totally calculated by CQI. With nonlinear utility function, RBs cannot be allocated independently and in that case, DSA method would be very complicated. When the utility functions are concave, iterative algorithms are used [60].

4.5.3 Packet Scheduling Algorithms based on Utility Functions

4.5.3.1 MDU, M-LWDF and LOG-MLWDF in 3GPP LTE

In this section MDU, M-LWDF and LOG-MLWDF scheduling algorithms (described in 4.5.1.1.1 and 4.5.1.1.2) are investigated in 3GPP LTE networks. Model of the system includes one BS and M users. A multipath fading channel is assumed to be the model of wireless simulated channel. The formula given below explains the channel impulse model:

$$h_i(t, \tau) = \sum_k \gamma_{k,i}(t) \delta(\tau - \tau_{k,i}). \tag{4.27}$$

where $\tau_{k,i}$ is the delay of *kth* path and $\gamma_{k,i}(t)$ is the complex amplitude that is a wide-sense stationary, narrow band and Gaussian random process. The paths are independent from each other. The channel frequency response is:

$$H_i(f,t) = \sum_k \gamma_{k,i}(t)e^{-j2\pi f\tau_{k,i}} \tag{4.28}$$

The Signal to Noise Ratio (channel gain) for user i, is as follows:

$$\rho_i(f) = \frac{|H_i(f)|^2}{N_i(f)}, \tag{4.29}$$

where $N_i(f)$ is the noise power density.

The achievable rate of user *i* at frequency *f*, $C_i(f)$, (assuming BER and power density to be constant) is presented as follows:

$$C_i(f) = \log_2\left(1 + \frac{\beta p(f)|H_i(f)|^2}{N_i(f)}\right) = \log_2(1 + \beta p(f)\rho_i(f))\ \frac{\frac{bits}{sec}}{Hz}. \tag{4.30}$$

Where:

$$\beta = \frac{1.5}{-ln(5BER)}. \tag{4.31}$$

The throughput of user *i* is calculated as follows:

$$r_i = \int_{D_i} C_i\,(f)\mathrm{d}f \tag{4.32}$$

and has been used in DSA (Dynamic Subcarrier Allocation) as Equation (4.25), which makes the calculation simpler compared to RR algorithms. Figure.4.4 shows the flowchart of packet scheduling algorithm using utility function and DSA allocation scheme to assign RBs to users. The MDU, MLWDF and LOG-MLWDF scheduling algorithms have been adapted to use the utility function idea as follow. The MLWDF utility function is given below:

$$U_i = \gamma_i W_i(t) r_i(t)\,. \tag{4.33}$$

The LOG-MLWDF utility function is presented as below:

$$U_i = \gamma_i r_i(t)\big(W_i + log\big(Wave_i(t)\big)\big)^{1.5}$$

It has been modified in this thesis as follows:

$$U_i = \gamma_i r_i(t)\big(W_i + ln\big(1 + Wave_i(t)\big)\big)^{1.5}, \tag{4.34}$$

where:

$$Wave(t) = \frac{1}{T}\sum_{i\epsilon N} W_i\,(t).$$

And the MDU utility function is shown in the following equation:

$$max_{RB^n, i\in A^n, s\in B^n} \sum_{i\in A^n, s\in B^n} \frac{\left|U'_{i,s}(w_i)[n]\right|}{\lambda_{i,s}} r_i[n].$$

Where the marginal utility function for voice users is as below:

$$\left|U'_{i,v}(w_i)\right| = \begin{cases} w_i, w_i \leq 2.5ms \\ w_i^{1.5} - 2.5^{1.5} + 2.5, w_i \geq 2.5ms. \end{cases}$$

The marginal utility function for streaming users is given by:

$$\left|U'_{i,s}(w_i)\right| = \begin{cases} w_i^{0.6}, w_i \leq 5ms \\ w_i - 5 + 5^{0.6}, w_i \geq 5ms. \end{cases}$$

The marginal utility function for best effort users is as follows:

$$\left|U'_{i,s}(w_i)\right| = \begin{cases} w_i^{0.5}, w_i \leq 5ms \\ 5^{0.5}, w_i \geq 5ms \end{cases}.$$

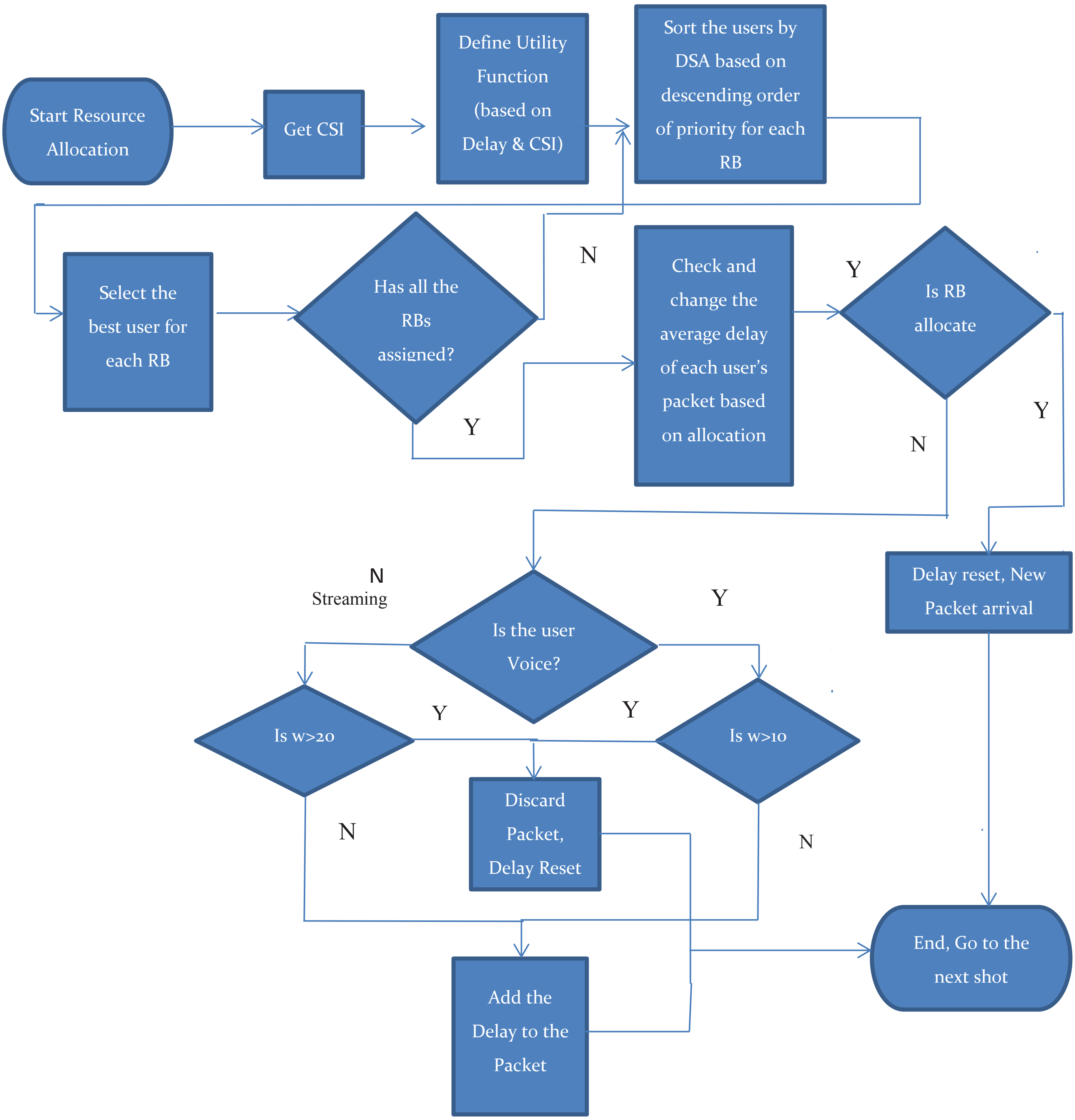

Figure 4.4: Packet Scheduling Algorithm for Voice and Streaming Users Based on DSA and Utility Functions

4.5.4 Results and discussion

The performance evaluation is based on the modeling assumption in Section 2.1, 2.3, 2.5, 2.6. Two scenarios are considered in this work. We also have some relevant assumptions as follows. In the first scenario, there are 40 voice and 61 to 190 streaming users distributed uniformly within the simulation area. In the second scenario, 1 to 36 best effort users have been added to the network. Users are constantly moving between [1-100] km/h speed in random directions. The SNR reports to the serving eNodeB are assumed to be instantaneous and free of delay and error. The buffer size for all streaming and best effort users are assumed to be infinite. The MLWDF, LOG-MLWDF and MDU methods have been simulated for 10000 time slots (1ms each) using the MATLAB software. When a packet delay is more than the waiting time threshold, it is considered to be discarded. The permissible waiting time of a packet in the eNodeB buffer (considered as the threshold of HOL packet delay), is set to be 20ms, 10ms and 100ms for streaming, voice and best effort users respectively. BER is assumed to be 10^{-6} in this work. The performance evaluation of these simulations is based on system throughput, fairness and delay which have been described in Section 2.6. In this section, simulation results are analysed. Figures 4.5 to 4.9 are related to the first scenario in which there are no best effort users in the network. Figure 4.5 and Figure 4.6 show the system throughput graphs for streaming and voice users of the three packet scheduling algorithms. The throughput of streaming users for the LOG-MLWDF algorithm outperforms MDU and MLWDF (as seen in Figure 4.5), whereas for voice users, MDU has better throughput performance compared to the LOG-MLWDF and MDU methods (as shown in Figure 4.6). Table 4.1, shows that LOG-MLWDF outperforms MLWDF in order of 13.8 to 28.7 percentages in terms of throughput of the streaming user.

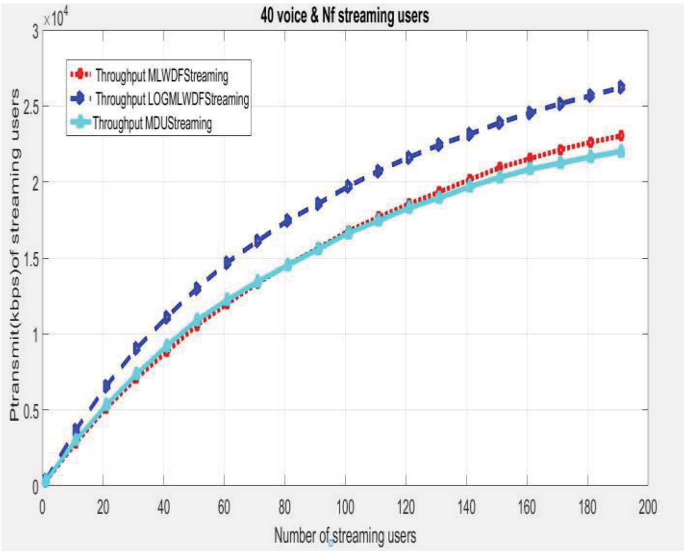

Figure 4.5 : System throughput for streaming users vs. number of steaming users

Table 4.1: Throughput comparison for streaming users in kbps (first scenario)

	20	60	100	140	160	190
LOG-MLWDF	6569	14698	19708	23137	24545	26217
MLWDF	5103	11989	16775	20151	21539	23028
MDU	5296	12231	16640	19685	20841	22007
Improvement of LOG-MLWDF over MLWDF(%)	28.7	22.5	17.4	14.8	13.9	13.8

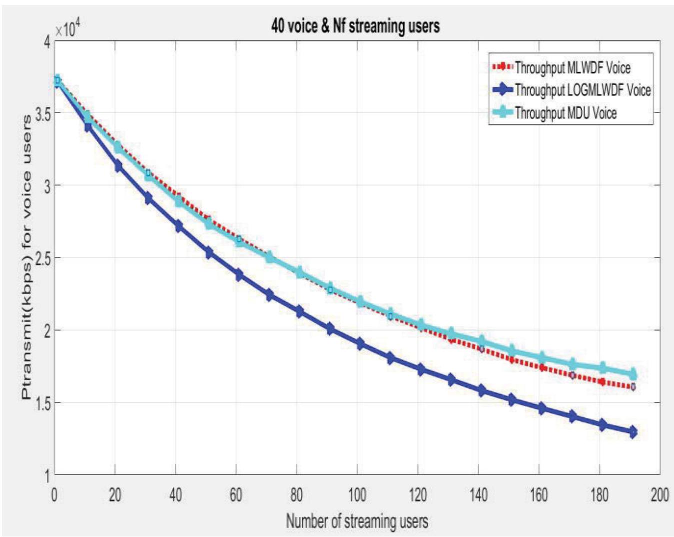

Figure 4.6: System throughput for streaming & voice users vs. number of streaming users

Table 4.2: Throughput comparison for voice users in Mbps (first scenario)

	20	60	100	140	160	190
LOG-MLWDF	31.36	23.77	19.09	15.9	14.58	13.01
MLWDF	32.73	26.63	21.83	18.67	17.39	16.01
MDU	32.55	26.08	21.95	19.12	18.03	16.96
Improvement of LOG-MLWDF over (%)MDU	-3.6	-8.8	-13	-16.84	-19.13	-23.29

Table 4.2 shows that MDU has better results in voice throughput compared to LOG-MLWDF in the order of 3.6 to 23.29%. Figure 4.7 and 4.8 show the PLR performance for the three algorithms. It can be observed from Figure 4.6 and Table 4.3 that for all streaming users, LOG-MLWDF has satisfied the PLR threshold of 10^{-3}[63]. With an increasing number of streaming users, MLWDF and MDU have greater PLR than LOG-MLWDF for streaming users. Table 4.3 shows that LOG-MLWDF improves the PLR performance of MLWDF by the range of 93 to100% for best effort users. For voice users, LOG-MLWDF has greater PLR than the other two algorithms (Figure 4.8 and Table 4.4) even though the PLR is still below the permissible threshold of conversational voice which is 10^{-2}[63]. The MDU algorithm followed by LOG-MLWDF has the highest PLR for streaming and voice users.

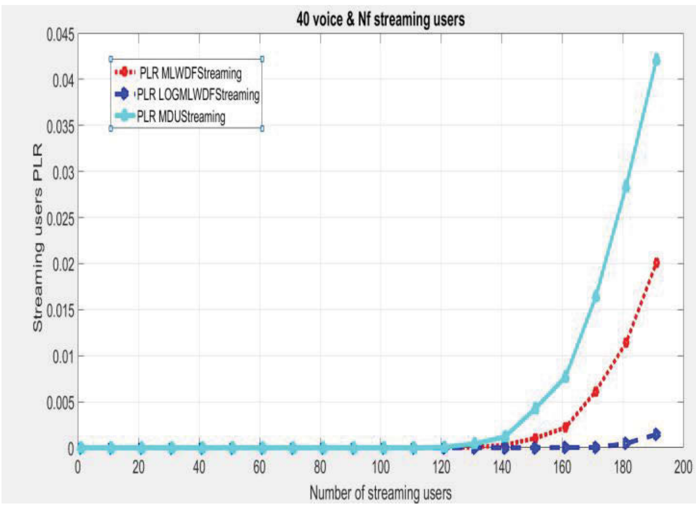

Figure 4.7 : Packet Loss Ratio of streaming users vs. number of streaming users

Table 4.3 : PLR comparison for streaming users (first scenario)

	20	60	100	140	160	190
LOG-MLWDF	0	0	0	0	9.58e-6	0.001
MLWDF	0	0	0	0.0012	0.0022	0.0211
MDU	0	0	0	0.0012	0.0088	0.0425
Improvement of LOG-MLWDF over MLWDF(%)	0	0	0	-100	-99.5	-93.36

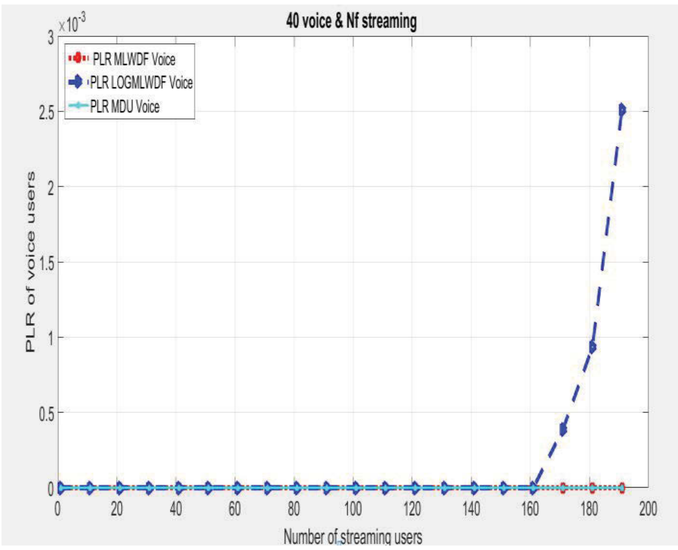

Figure 4.8: Packet Loss Ratio of voice users vs. number of streaming users

Table 4.4: PLR comparison for voice users (first scenario)

	20	60	100	140	160	190
LOG-MLWDF	0	0	0	0	3.95e-5	0.0022
MLWDF	0	0	0	0	0	0
MDU	0	0	0	0	0	0

In Figure 4.9, the fairness of the three algorithms has been compared. MDU has the worst fairness performance among the three methods. The fairness performance of LOG-MLWDF is better than the MDU and MLWDF algorithm, and MLWDF shows better fairness performance compared to MDU. Table 4.5 gives the improvement ratio of LOG-MLWDF compared to MLWDF which is within the range of 0.5 to 0.32%.

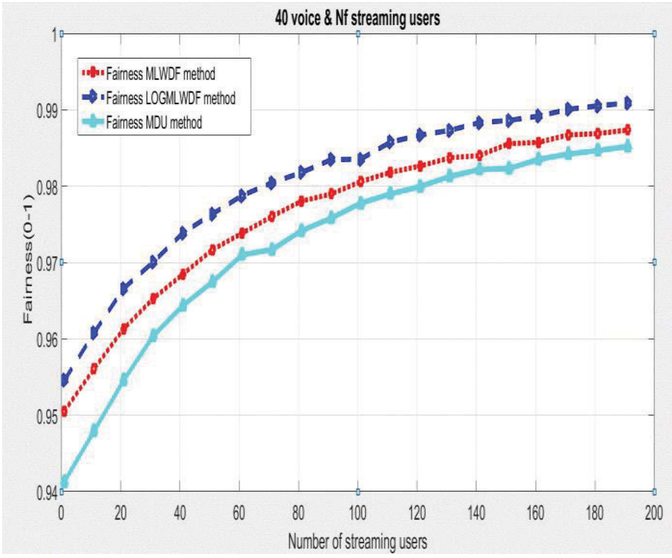

Figure 4.9: Fairness vs. number of streaming users

Table 4.5: Fairness comparison (first scenario)

	20	60	100	140	160	190
LOG-MLWDF	0.9660	0.9789	0.9847	0.9880	0.9893	0.9911
MLWDF	0.9613	0.9738	0.9805	0.9843	0.9861	0.9875
MDU	0.9546	0.9705	0.9777	0.9825	0.9837	0.9851
Improvement of LOG-MLWDF over MLWDF(%)	0.4	0.5	0.42	0.37	0.32	0.36

Considering the second scenario in which best effort users are added to the previous network environment, simulation results are given in Figure 4.10 to Figure 4.12. From the percentile delay and throughput perspective of the best effort users, it can be seen that LOG-MLWDF method outperforms the other two algorithms. LOG-MLWDF outperforms the throughput of best effort user's compared to MLWDF method by the range of 155.8% to 253.5% (Table 4.6). The 95th percentile delay for best effort users decrease by a percentage range of 38 to 41% comparing LOG-MLWDF over MLWDF performance in Table 4.7.

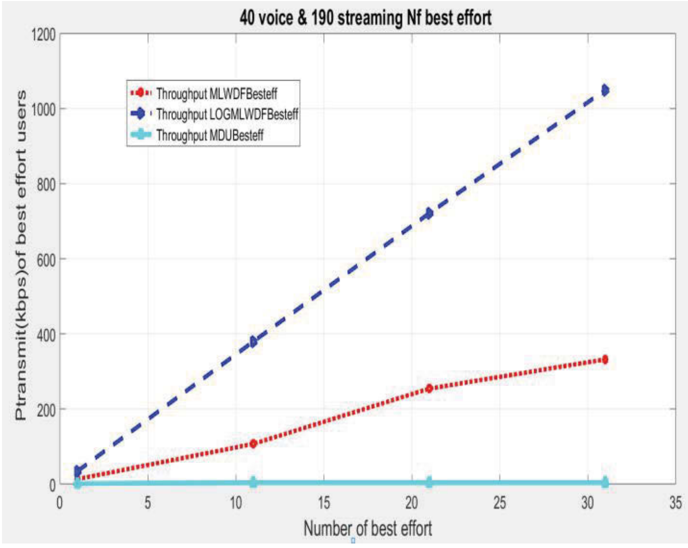

Figure 4.10 : Best effort throughput vs. number of users

Table 4.6: Throughput comparison for best effort users (second scenario)

	1	10	20	30
LOG-MLWDF	34869	379842	721275	1049079
MLWDF	13629	107439	254526	332406
MDU	1770	4425	4425	4425
Improvement of LOG-MLWDF over MLWDF(%)	155.8	253.5	183.3	215.6

In Figure 4.12, the three algorithms have been compared according to the fairness metric. It can be seen that LOG-MLWDF outperforms the other two algorithms and MDU has the least fairness compared to the other algorithms.

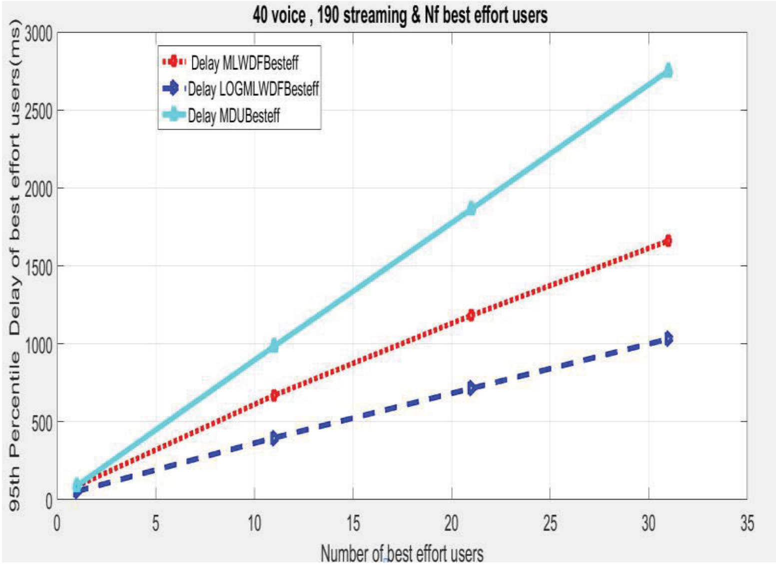

Figure 4.11 : Percentile delay of best effort users vs. number of users

Table 4.7: 95th Percentile delay for best effort users (second scenario)

	1	10	20	30
LOG-MLWDF	53	394.50	727	1040
MLWDF	86	655	1215	1759
MDU	89.5	975.50	1862	2754
Improvement of LOG-MLWDF over MLWDF(%)	-38	-39	-40	-41

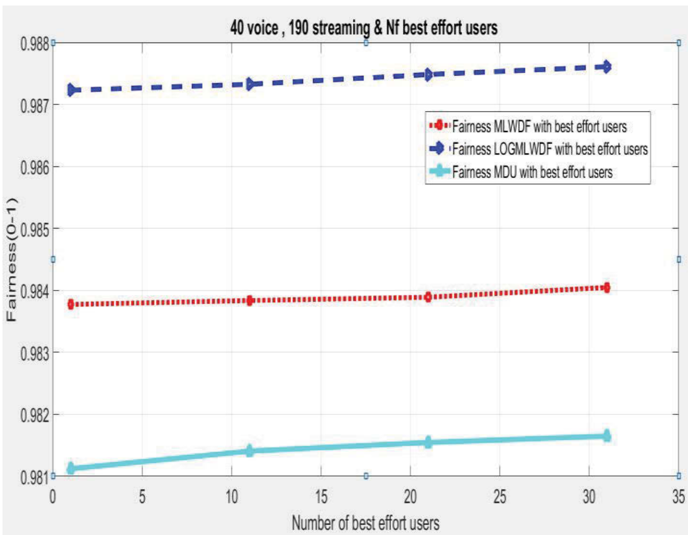

Figure 4.12 : Fairness vs. number of users

4.6 Summary

In this chapter different scheduling algorithms based on QoS requirements have been reviewed and the LOG-MLWDF scheduling algorithm for LTE wireless networks has been introduced. It is based on dynamic resource allocation as cross layer optimization solution of downlink 3GPP LTE systems. A comparison between three algorithms MDU, MLWDF, and the proposed extended LOG-MLWDF has been given. It considers the strengths and weaknesses of LOG-MLWDF compared to the other two well-known algorithms. The proposed method outperforms the other two methods in terms of PLR and throughput for streaming users and 95^{th} delay and throughput for best effort users. It also has the best fairness performance amongst the mentioned algorithms.

Chapter 5 Conclusion and Future Research Direction

Based on the research questions initially provided in the introduction, some contributions were made to address the existing research gaps throughout this thesis. Section 5.1 looked into the research methods that were used throughout the thesis. A brief summary of the contributions in this thesis is considered in Section 5.2. Future research directions are listed in Section 5.3.

5.1. Research Methods

The applicable research method that is used in this work is a combination of a mathematical modeling and simulation in a MATLAB environment. My research process is given as follows:

1- Study LTE Network and modeling it.

2- Study different methods of packet scheduling and resource allocation in OFDM and LTE networks

3- Optimize existing packet scheduling methods

4- Evaluate the proposed packet scheduling methods in heterogeneous traffic and different scenarios.

5.2 Summary of the Contribution

5.2.1 Low Computational Complexity Algorithm

The proposed packet scheduling algorithm in this work used the dynamic resource allocation as an optimal solution for the cross-layer optimisation problem in downlink 3GPP LTE systems. Cross-layer optimisation based on utility functions indicates the method by which wireless resources (RBs in LTE) are allocated to users in order to maximise the utility function. It is presented in this thesis as the main resource allocation algorithm which considers three scheduling algorithms as utility functions. This algorithm addresses the following research question: Is it possible to develop a new packet scheduling algorithm with less computational complexity to enhance the throughput, QoS and fairness with different traffic like GBR, NGBR and NRT data streaming? If yes, how much improvement does the new algorithm offer compared to previous algorithms? Dynamic Subcarrier Allocation (DSA), here used as dynamic resource allocation, provides less computational complexity compared to the scheduling algorithm based on the QoS requirement described in Section 4.1, Section 4.2 and Section 4.3 as RAA (Resource Allocation and Assignment) Algorithm, Scheduling Algorithm Based on Matrix and QoS-Oriented Joint Time and Frequency Domain Scheduling. The proposed algorithm in this thesis is based on DSA formula for linear utility functions, the computational complexity in this method (consisting of calculating $U'C$ for users and related RBs) in each time interval is less than calculating the elements of the matrix in which the priority of each user on each Radio Resource Unit at TTI t is determined based on the scheduling algorithm(Utility Function). In this algorithm Sorting-Search Algorithm has been used for subcarrier assignment that the computational complexity is nearly $(M-1)^2(K+1)\log_2 K$ which is still efficient compared to the number of the choices of combinatorial optimization problem K^M. Starting from the first RB, $U'C$ is calculated for all users and the maximum value is obtained. The related user for this obtained value is selected to be the user which the RB is allocated to. In the second step, the allocated RB is removed and the previous step is repeated for the purposes of choosing the next RB. The next steps are to continue as step one and two, till the last RB is allocated. The allocation and assignment have been merged in this algorithm which makes it much simpler compared to the RAA algorithm which has separate steps for resource allocation and assignment or the QoS-Oriented Joint Time and Frequency Domain Scheduling which has separate phases for the frequency and time domain scheduler.

5.2.2. Providing good Quality of Service (QoS) performance

LOG-MLWDF as the proposed algorithm in LTE in Chapter 4 has been shown to have the better performance in the PLR throughput for streaming users compared to the MDU and MLWDF methods. It has been proved that LOG-MLWDF outperforms MLWDF by 13.8 to 28.7 percentages from the throughput of streaming user aspect. It also significantly improves the PLR performance of MLWDF for streaming users by the range of 93 to100%. The performance improvement in fairness for the LOG-MLWDF method over the MLWDF is within the range of 0.5 to 0.32%. In the second scenario, when we add the best effort users to the network, the results show that the LOG-MLWDF method significantly improves the best effort user's throughput by the range of 155.8% to 253.5% over the MLWDF method. It also decreases the 95th percentile delay of best effort users by the percentage range of 38 to 41 over the MLWDF method.

5.3 Future Research Direction

Although mobile cellular systems have improved greatly over recent years, packet scheduling over the MAC layer has experienced a number of challenges [64-72]. Some of the important problems are: inadequate CQI report, a radio transmission environment that can result in error, and different QoS requirements for diverse data services. The reported CQI could be imperfect because of old or erroneous CQI and wrong channel estimation. The delay between the time of the estimated CQI by the user and the time when the CQI report is used by the packet scheduler may result in wrong scheduling. The CRR (CQR Reporting Rate) may cause another challenge for packet scheduling and this rate is controlled by the eNB. It is obvious that the accuracy of the MCS (Modulation and Coding Scheme) is dependent on CRR, with higher CRR the CQI information at eNB is more recent and accurate. On the other hand, the CQI report would be more inaccurate if a lower CRR is used. For high speed users, the CQI report is even less accurate.

HARQ, consists of finding errors, correcting errors and retransmission requests, is extensively used by current mobile services to provide reliable retransmission of TBs. Based on the challenges which have been detailed, some future research directions can be summarized as follows:

- Is it possible to design a new algorithm based on impairment and instability in the wireless channel environment and imperfect CQI? And how much improvement in performance can the proposed algorithms obtain over previous algorithms?
- Since countries have agreed to develop a new generation of 5G mobile technologies, it opens the door for innovation that will determine the way of the future research in communication. Packet scheduling will be one of the challenges in 5G that will need to be considered in more detail in future work.

REFERENCES

[1] ITU, “World Telecommunication/ICT Indicators Database,” in http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2014-e.

[2] R. Ramachandran, “Evolution to 3G Mobile Communication: Second Generation Cellular Systems,” *RESONANCE,* vol. 8, no. 9, pp. 60-72, 2003.

[3] R. E. Sheriff, Y. Fun Hu, “Mobile Satellite Communication Network,” Wiley Publication, November 26, 2001.

[4] S. C. Nguyen, "Packet Scheduling for LTE-Advanced, the True 4G Technology," in Faculty of Engineering and Information Technology: Master Thesis, University of Technology, Sydney, 2011.

[5] A. Samukic, "UMTS universal mobile telecommunications system: development of standards for the third generation," in *Global Telecommunication Conference, 1998. GLOBECOM 1998. The Bridge to Global Integration. IEEE*, 1998, pp. 1976-1983.

[6] A. Al-Kandari, M. Al-Nasheet and A. R. Abdulgafer,"WiMAX vs. LTE: an Analytic Comparison," *IEEE Digital Information and Communication Technology and it's Applications (DICTAP) Conference*, Bangkok, Thailand, May 2014

[7] T. I. (2011), "Short and Long-Term Visions of 4G," *in http://trends-intelecoms. blogspot.com.au/2011/07/short-and-long-term-visions-of-4g.html,* accessed: 23 September 2013.

[8] A. Hadden, "Mobile broadband-where the next generation leads us [industry perspectives],"*Wireless Communications, IEEE, vol. 16, pp. 6-9, 2009.*

[9] Qualcomm, "3GPP Long-Term Evolution," *Qualcomm Incorporated,* January 2008.

[10] L. Wieweg, "UMTS/LTE Flexible Capacity within the Harmonised Bands," in Ericsson LTE Warsaw, June 2007.

[11] Agilent. "3GPP Long Term Evolution: System Overview, Product Development, and Test Challenges, Application Note*, " Literature Number,* 2009.

[12] H. A. M. Ramli, "Performance Analysis of Packet Scheduling in Long Term Evolution (LTE)," *Doctor of Philosophy, Faculty of Engineering and Information Technology, University of Technology*, Sydney, Australia, 2011.

[13] 3GPP, "Policy and Charging Control Architecture (Release 9)," *TS 23.203, version 9.3.0*, December 2009.

[14] M. R. Souryal and R. L. Pickholtz, "Adaptive modulation with imperfect channel information in OFDM," in Communications*, 2001, ICC 2001, IEEE International Conference on vol.6, 2001, pp. 1861-1865.*

[15] Y. Soo Cho, J. Kim, W. Young Yang, C. G. Kang, “MIMO-OFDM Wireless Communications with MATLAB”, Wiley, 2010.

[16] G. Smith,” Wireless Communications”, Cambridge university press, 2005.

[17] H. A. M. Ramli, K. Sandrasegaran, R. Basukala, and W. Leijia, "Modelling and Simulation of Packet Scheduling in the Downlink Long Term Evolution System," in 15th Asia-Pacific Conference on Communications, 2009, pp. 68-71.

[18] C. Mehlführer, M. Wrulich, J. C. Ikuno, D. Bosanska, and M. Rupp, "Simulating the long term evolution physical layer," in Proc. of the 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, 2009, p. 124.

[19] 3GPP, "Physical Layer Procedures (Release 10)," TR 36.213, version 10.1.0, June 2011.

[20] K. Dongmyoung, C. Youngkyu, J. Sunggeun, H. Kwanghun, and C. Sunghyun, "A MAC/PHY Cross-LayerDesign for Efficient ARQ Protocols," IEEE Communications Letters, vol. 12, no.12, pp. 909-911, 2008.

[21] H. Shirani-Mehr, H. Papadopoulos, S. Ramprashad, and G. Caire, "Joint Scheduling and ARQ for MU- MIMO Downlink in the Presence of Inter-Cell Interference," IEEE Transactions on Communications, vol. PP, no. 99, pp. 1-12, 2010.

[22] X. Liu, Z. Huiling, and W. Jiangzhou, "Packet Retransmission using Frequency Diversity in OFDMA," in IEEE 21st International Symposium on Personal Indoor Mobile Radio Communications, 2010, pp. 1190-1194.

[23] C. Yao-Liang and T. Zsehong, "Performance Analysis of Two Multichannel Fast Retransmission Schemes for Delay-Sensitive Flows " IEEE Transactions on Vehicular Technology, vol. 59, no. 7, pp. 3468-3479, 2010.

[24] B. Kian Chung, A. Doufexi, and S. Armour, "Performance Evaluation of Hybrid ARQ Schemes of 3GPP LTE OFDMA System," in Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007, IEEE 18th International Symposium on, 2007, pp. 1-5.

[25] M. Stambaugh, "HARQ Process Boosts LTE Communications," *Agilent Technologies Sep,* 2008.

[26] M. B. Pursley and S. D. Sandberg, "Incremental-redundancy transmission for meteor-burst communications," Communications, IEEE Transactions on, vol. 39, pp. 689-702, 1991.

[27] 3GPP, "Feasibility Study for Orthogonal Frequency Division Multiplexing (OFDM) for UTRAN Enhancement (Release 6)," TR25.892, version 6.0.0, June2004.

[28] Farooq Khan, "LTE for 4G Mobile Broad Cast: Air Interface Technology and Performance," Cambridge University Press. New York, NY, USA, March 2009.

[29] B. S. Tsybakov, "File Transmission over Wireless Fast Fading Downlink," IEEE Transactions on Information Theory, vol. 48, no. 8, pp. 2323-2337, 2002.

[30] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, 3G Evolution: HSPA and LTE for Mobile Broadband, First ed.: Elsevier Ltd., 2007.

[31] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Communication Wireless System," in IEEE51st Vehicular Technology Conference Proceedings. vol. 3, 2000, pp. 1854-1858.

[32] G. Monghal, K. I. Pedersen, I. Z. Kovacs, and P. E. Mogensen, "QoS Oriented Time and Frequency Domain Packet Schedulers for The UTRAN Long Term Evolution," in Vehicular Technology Conference, 2008. VTC Spring , IEEE, 2008, pp. 2532-2536.

[33] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing Quality of Service over a Shared Wireless Link," IEEE Communications Magazine, vol. 39, no. 2, pp. 150-154, 2001.

[34] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real time and non-real-time data in HDR," in in Proceedings of 17th International Teletraffic Congress (ITC-17, 2000).

[35] A. K. F. Khattab and K. M. F. Elsayed, "Channel-quality dependent earliest deadline due fair scheduling schemes for wireless multimedia networks," in 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems, 2004, pp. 1-8.

[36] Y. Ofuji, A. Morimoto, H. Atarashi, and M. Sawahashi, "Sector throughput using frequency-and-time domain channel-dependent packet scheduling with channel prediction in OFDMA downlink packet radio access," in Vehicular Technology Conference, 2005. VTC-2005-Fall. 2005 IEEE 62nd, 2005, pp.1589-1593.

[37] C. Wengerter, J. Ohlhorst, and A. G. E. von Elbwart, "Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA," in Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st, 2005, pp. 1903-1907 Vol. 3.

[38] A. Pokhariyal, T. E. Kolding, and P. E. Mogensen, "Performance of Downlink Frequency Domain Packet Scheduling for the UTRAN Long Term Evolution,"in Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on, 2006, pp. 1-5.

[39] B. Kian Chung, S. Armour, and A. Doufexi, "Joint Time-Frequency Domain Proportional Fair Scheduler with HARQ for 3GPP LTE Systems," in Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th, 2008, pp. 1-5.

[40] J. Gross, H. Karl, F. Fitzek, and A. Wolisz, "Comparison of Heuristic and Optimal Subcarrier Assignment Algorithms," in Proceedings of International Conference on Wireless Networks, 2003, pp. 1-7.

[41] J. Gross, J. Klaue, H. Karl, and A. Wolisz, "Subcarrier Allocation for Variable Bit Rate Video Streams in Wireless OFDM Systems," in IEEE 58th Vehicular Technology Conference. vol. 4, 2003, pp. 2481-2485.

[42] A. K. F. Khattab and K. M. F. Elsayed, "Opportunistic Scheduling of Delay Sensitive Traffic in OFDMA-based Wireless Networks," in International Symposium on a World of Wireless, Mobile and Multimedia Networks, 2006, pp.288-297.

[43] X. Liu, L. Guangyi, Y. Wang, and P. Zhang, "Downlink Packet Scheduling for Real-Time Traffic in Multi-User OFDMA System," in Vehicular Technology Conference, 2006. VTC-2006 Fall. 2006 IEEE 64th, 2006, pp. 1-5

[44] R. Kausar, Y. Chen, and K. K. Chai, "Adaptive Time Domain Scheduling Algorithm for OFDMA based LTE-Advanced networks," in Wireless and Mobile Computing, Networking and Communications (WiMob), 2011 IEEE 7th International Conference on, 2011, pp. 476-482.

[45] N. Ruangchaijatupon and J. Yusheng, "Simple Proportional Fairness Scheduling for OFDMA Frame-Based Wireless Systems," in Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE, 2008, pp. 1593-1597.

[46] N. Ruangchaijatupon and J. Yusheng, "Resource Allocation for Guaranteed Service in OFDMA Based Systems," in Wireless Communications and Networking Conference, 2009. WCNC 2009. IEEE, 2009, pp. 1-6.

[47] I. S. Hwang, C. Chien-Yao, and H. Bor-Jiunn, "Channel-Aware Slot Assignment by Ant Colony in OFDMA-Based Mobile WiMAX Networks," in Parallel Processing Workshops (ICPPW), 2011 40th International Conference on, 2011,pp. 119-126.

[48] G. Monghal, K. I. Pedersen, I. Z. Kovacs, and P. E. Mogensen, "QoS Oriented Time and Frequency Domain Packet Schedulers for The UTRAN Long Term Evolution," in Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE, 2008, pp. 2532-2536.

[49] A. Wang, L. Xiao, S. Zhou, X. Xu, and Y. Yao, "Dynamic resource management in the fourth generation wireless systems," in Communication Technology Proceedings, 2003. ICCT 2003. International Conference on, 2003, pp. 1095-1098 vol.2.

[50] A. Pokhariyal, K. I. Pedersen, G. Monghal, I. Z. Kovacs, C. Rosa, T. E. Kolding, and P. E. Mogensen, "HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution," in Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th, 2007, pp. 2761-2765.

[51] T. E. Kolding, "QoS-aware proportional fair packet scheduling with required activity detection," in Vehicular Technology Conference, 2006. VTC-2006 Fall. 2006 IEEE 64th, 2006, pp. 1-5.

[52] C. Wang, Y. Huang, "Delay-scheduler coupled throughput-fairness resource allocation algorithm in the long-term evolution wireless networks," IET Communications journal, November 2014, Volume: 8, Issue: 17, pp. 3105 – 3112.

[53] A. Sharifian, R. Schoenen, *and* H. Yanikomeroglu, "Joint Real time and Non real time Flows Packet Scheduling and Resource Block Allocation in Wireless OFDMA Networks," IEEE Transactions on Vehicular Technology, April 2016 ,Volume: 65, Issue: 4, pp. 2589-2607.

[54] O. Grondalen, A. Zanella, K. Mahmood, M. Carpin, J. Rasool, O.N. Osterbo, "Scheduling Policies in Time and Frequency Domains for LTE Downlink Channel: a Performance Comparison," IEEE Transactions on Vehicular Technology, July 2016.

[55] S. -B. Lee, S. Choudhury, A. Khoshnevis, S. Xu, S. Lu, " Downlink MIMO with Frequency-Domain Packet Scheduling for 3GPP LTE," April 2009, INFOCOM 2009, IEEE Conference, pp.1269-1277.

[56] S. Niafar, Z. Huang, D. H. K. Tsang, "An Optimal Standard-Compliant MIMO Scheduler for LTE Downlink," IEEE Transactions on Wireless Communications, 2014, vol. 13, issue. 5, pp. 2412 – 2421

[57] H-S. Liao, P. Chen, and W-T. Chen, "An Efficient Downlink Radio Resource Allocation with Carrier Aggregation in LTE-Advanced Networks, " IEEE Transactions on Mobile Computing , Oct. 2014, vol. 13, issue. 10, pp. 2229-2239.

[58] Y. Wang , K. I. Pedersen , T. B. Sorensen , P. E. Mogensen, "Carrier Load Balancing and Packet Scheduling for Multi-Carrier Systems," May 2010, IEEE Transactions on Wireless Communications ,vol. 9, issue. 5, pp. 1780-1789.

[59] "Smart Downlink Scheduling for Multimedia Streaming over LTE Networks with Hard Handoff," IEEE Transactions on Circuits and Systems for Video Technology, Nov. 2015, vol.25, issue.11, pp. 1815 – 1829.

[60] G. Song and Y. (G). Li, "Cross Layer Optimization for OFDM Wireless Network-part I and part II," IEEE Trans. Wireless Communication, vol. 4, March 2005.

[61] R. Heidari, M. Mehrjoo,"Delay and Rate based Multichannel Scheduling for Heterogeneous Traffic", IEEE CNDS Conference, Feb. 2011.

[62] G. Song, Y.(G). Li, Leonard J. Cimini, Jr. Haitare Zheng, "Joint Channel Aware and Queue Aware Data Scheduling in Multiple Shared Wireless Channels," WCNC 2004/IEEE Communication Society.

[63] 3GPP (2009b) TS 23.203 (V9.3.0) "Policy and Charging Control Architecture", Release 9.

[64] T. Bejaoui, A. Masmoudi, and N. Nasser, "Radio Resource Optimization and Scheduling Techniques for HSPA and Advanced LTE Technologies," Evolved Cellular Network Planning and Optimization, pp. 256-296, 2011.

[65] K. Zhen, W. Jiangzhou, and K. Yu-Kwong, "A New Cross Layer Approach to QoS-Aware Proportional Fairness Packet Scheduling in the Downlink of OFDM Wireless Systems," in IEEE International Conference on Communications, 2007,pp. 5695-5700.

[66]K. Gunaseelan, R. Venkateswari, and A. Kandaswamy, "A Novel Efficient Resource Allocation Algorithm for Multiuser OFDM Systems," IETE Technical Review, vol. 25, no. 4, pp. 201-208, 2008.

[67] H. Lei, L. Zhang, X. Zhang, and D. Yang, "A Packet Scheduling Algorithm Using Utility Function for Mixed Services in the Downlink of OFDMA Systems," in IEEE 66th Vehicular Technology Conference, 2007, pp. 1664-1668

[68] P. Jeongsik, H. Sungho, and C. Ho-Shin, "A Packet Scheduling Scheme to Support Real-Time Traffic in OFDMA Systems," in IEEE 65th Vehicular Technology Conference, 2007, pp. 2766-2770.

[69] P. Svedman, S. K. Wilson, and B. Ottersten, "A QoS-Aware Proportional Fair Scheduler for Opportunistic OFDM," in IEEE 60th Vehicular Technology Conference. vol. 1, 2004, pp. 558 - 562.

[70] Y. J. Zhang and K. B. Letaif, "Adaptive Resource Allocation and Scheduling for Multiuser Packet-based OFDM Networks," in Proceedings of the IEEE International Conference on Communications, 2004, pp. 2949-2953.

[71] W. Lihua, M. Wenchao, and G. Zihua, "A Cross-Layer Packet Scheduling and Subchannel Allocation Scheme in 802.16e OFDMA System," in IEEE Wireless Communications and Networking Conference, 2007, pp. 1865-1870.

[72] J. Sang Soo, J. Dong Geun, and J. Wha Sook, "Cross-Layer Design of Packet Scheduling and Resource Allocation in OFDMA Wireless Multimedia Networks," in IEEE 63rd Vehicular Technology Conference. vol. 1, 2006, pp.309-313.