UNIVERSITI MALAYSIA SARAWAK • UNIMAS •

Faculty of Engineering

Utility-Based Scheduling for Wireless Network using Modified Sigmoidal Function

Olalekan Bello

Doctor of Philosophy
2016

Utility-Based Scheduling for Wireless Network using Modified Sigmoidal Function

Olalekan Bello

A thesis submitted
In fulfillment of the requirements for the degree of Doctor of Philosophy

Faculty of Engineering
UNIVERSITI MALAYSIA, SARAWAK
2016

Acknowledgements

First and foremost, I would like to thank God Almighty for His prophesies that have come true and His protections and provisions in the journey towards the fulfillment of this doctoral degree. My special gratitude goes to my beloved wife Grace Ebunola, who has been my moral support and encouragement. Due to her love and patience, this dream has come true. I give many thanks to my sister Biola Adeyemi and my wonderful friend Tunji Alabi who have been both morally and financially supportive.

I would like to thank very much my main supervisor, Assoc. Prof. Dr. Hushairi Hj Zen, for receiving me with open arms, and for his commitment, patience and generosity in guiding me. He has been an inspiration in both personal and professional aspects. My acknowledgments go to my co-supervisors, Assoc. Prof. Al-Khalid Othman and Prof. Dr. Datuk khairruddin Ab Hamid for their rich technical discussions and friendship.

I am very grateful to the federal government of Nigeria for the sponsorship. I also like to thank Universiti Malaysia, Sarawak for the Zamallah scholarship. Last, but not least, many thanks to the friends and colleagues I have known here in Malaysia for their help and friendship.

Abstract

In order to face the challenges of wireless network operators and increase the number of satisfied users with different quality-of-service (QoS) requirements, an efficient utility-based scheduling and resource allocation technique is needed. Therefore, three utility-based scheduling frameworks, called maximum QoS satisfaction (MQS), was developed to maximize user satisfaction while maintaining efficient trade-off between resource distribution fairness and system efficiency among users in a broadband wireless network. The MQS was compared with maximum sum rate (MSR), proportional (PF) and delay-based satisfaction maximization/throughput-based satisfaction maximization (DSM/TSM) scheduling schemes which were designed for various optimization objectives. In the cell, a user only has one traffic flow which can be chosen from Voice over Internet Protocol (VoIP), video streaming, File Transfer Protocol (FTP) and Hypertext Transfer Protocol (HTTP) services. The simulation of the multiplexed VoIP, video, FTP and HTTP services, show that the MSR obtained 2.8Mbps, 52% and 0.35; PF achieved 2Mbps, 39% and 0.4; MQS achieved 1.7Mbps, 77% and 0.72; and DSM/TSM recorded 0.96Mbps, 54% and 0.52 , respectively, in average system throughput, user call satisfaction and fairness. The analysis of the results, show that the MSR algorithm maximizes the system efficiency as it obtained the highest average system throughput, but which results in the lowest fairness index among users. The PF algorithm compromises between efficiency and fairness in resource distribution, thus brings down the average system throughput in exchange for the higher call fairness compared with MSR. The MQS sacrificed the average system throughput further to achieve the highest user call satisfaction and fairness, while the DSM/TSM traded off higher amount of throughput; but still achieve fairly lower user call satisfaction and fairness compared with MQS. Therefore, with the MQS, the network operators can guarantee user satisfaction to

maintain a high number of subscribers, decrease churn, and attract new subscribers. Resource allocation in relaying systems has focused on maximizing the system capacity without consideration for resource distribution fairness or user satisfaction. To solve this problem, the traditional amplify-and-forward (AF) scaling coefficients are equalized and applied in subcarrier allocation. The MQS and DSM; and MQS-sc and DSM-sc scheduling algorithms are used to analyze the performances of the traditional and equalized methods, respectively. The simulation results, in the multiplexed services consisting of VoIP, video, FTP and HTTP, show that the MQS achieved 1.42Mbps, 93% and 0.77, respectively, in average system throughput, user call satisfaction and fairness; which were increased by the MQS-sc to 1.52Mbps, 95% and 0.84, respectively. The DSM-sc also increased the average system throughput, user call satisfaction and fairness of DSM from 0.96Mbps, 50% and 0.44 to 1.1Mbps, 78% and 0.44, respectively. This implies that the equalized method can indeed be used by relay-based network operators to improve call satisfaction and fairness levels among users, while also increasing the average system capacity.

Penjadualan Berasaskan Utiliti bagi Rangkaian Tanpa Wayar yang Menggunakan Fungsi Sigmoidal Kali

Abstrak

Untuk menghadapi cabaran pengendali rangkaian tanpa wayar dan peningkatan bilangan pengguna yang berpuas hati dengan kualiti perkhidmatan (QoS) keperluan yang berbeza, yang cekap berasaskan utiliti penjadualan dan sumber peruntukan teknik yang diperlukan. Oleh itu, tiga berasaskan utiliti penjadualan rangka kerja dipanggil kepuasan QoS maksimum (MQS), telah dibangunkan untuk memaksimumkan kepuasan pengguna sambil mengekalkan off-perdagangan yang cekap antara keadilan pengagihan sumber dan kecekapan sistem antara pengguna dalam rangkaian jalur lebar tanpa wayar. MQS itu adalah berbanding dengan kadar jumlah maksima (MSR), berkadar (PF) dan pemaksimuman kepuasan pemaksimuman/kendalian-berdasarkan kepuasan berasaskan kelewatan (DSM/TSM) penjadualan skim yang telah direka untuk pelbagai pengoptimuman objektif. Di dalam bilik penjara, pengguna hanya mempunyai satu aliran trafik yang boleh dipilih dari suara melalui protokol Internet (VoIP), video streaming, perkhidmatan protokol pemindahan fail (FTP) dan protokol pemindahan hiperteks (HTTP). Simulasi VoIP multiplexed, video, Perkhidmatan HTTP dan FTP, Papar MSR yang diperolehi 2.8Mbps, 52% dan 0.35; PF dicapai 2Mbps, 39% dan 0.4; MQS mencapai 1.7Mbps, 77% dan 0.72; dan DSM/TSM mencatat 0.96Mbps, 54% dan 0.52, masing-masing dalam kendalian sistem purata, pengguna panggilan kepuasan dan keadilan. Analisis keputusan, menunjukkan bahawa algoritma MSR yang memaksimumkan kecekapan sistem kerana ia diperolehi kendalian sistem purata yang tertinggi, tetapi yang mengakibatkan Indeks keadilan terendah antara pengguna. Kompromi algoritma PF antara kecekapan dan keadilan dalam pengagihan sumber, sekali gus

membawa turun kendalian sistem purata Pertukaran keadilan panggilan lebih tinggi berbanding dengan MSR. MQS yang mengorbankan kendalian sistem purata tambahan untuk mencapai kepuasan panggilan pengguna tertinggi dan keadilan, manakala DSM/TSM diniagakan lebih tinggi jumlah kendalian kargo; tetapi masih mencapai kepuasan panggilan pengguna agak rendah dan keadilan berbanding MQS. Oleh yang demikian, dengan MQS, pengendalipengendali rangkaian boleh menjamin kepuasan pengguna untuk mengekalkan bilangan pelanggan yang tinggi, mengurangkan churn dan menarik pelanggan baru. Peruntukan sumber dalam menyampaikan sistem tumpuan memaksimumkan kapasiti sistem tanpa pertimbangan untuk sumber pengagihan keadilan atau pengguna kepuasan. Untuk menyelesaikan masalah ini, pekali skala tradisional menguatkan dan-MARA (AF) Wacana dan digunakan dalam pengagihan subcarrier. MQS dan DSM; dan algoritma penjadualan MQS-sc dan DSM-sc akan digunakan untuk menganalisis prestasi daripada kaedah tradisional dan Wacana, masing-masing. Keputusan simulasi, Perkhidmatan multiplexed terdiri daripada VoIP, video, FTP dan HTTP, menunjukkan bahawa MQS tersebut mencapai 1.42Mbps, 93% dan 0.77, masing-masing dalam kendalian sistem purata, pengguna panggilan kepuasan dan keadilan; yang telah meningkat sebanyak MQS-sc 1.52Mbps, 95% dan 0.84, masing-masing. DSM-sc juga meningkat kendalian sistem purata, pengguna panggilan kepuasan dan keadilan DSM dari 0.96Mbps, 50% dan 0.44 untuk 1.1Mbps, 78% dan 0.44, masing-masing. Ini menunjukkan bahawa kaedah wacana boleh memang digunakan oleh pengendali rangkaian berasaskan penyampai untuk mempertingkatkan tahap kepuasan dan keadilan panggilan antara pengguna, manakala juga meningkatkan kapasiti sistem purata.

Table of Contents

List of Tables

List of Figures

List of Symbols

Symbols	Meaning
$r_{m.k}$	Achievable transmission rate for user *m* at subcarrier *k*
q	Accumulated loop interference signal vector
d	Actual distance between the base station and user
D	AF amplification factor
$A_m(t)$	Arrival bit rate for user *m*
$\bar{R}_m$	Average data rate (throughput) for user *m*
$\bar{\Upsilon}_{SD}$	Average SNR of S-D channel
$\bar{\Upsilon}_{SR}$	Average SNR of S-R channel
$\bar{\Upsilon}_{RD}$	Average SNR of R-D channel
z_{SD}	AWGN between Source and Destination
z_{SR}	AWGN between Source and Relay
z_{RD}	AWGN between Relay and Destination
$\mathcal{C}$	Cardinality. Number of users in each service class
τ_c	Channel coherence time
$\|h_{SD}\|^2$	Channel gain between Source and Destination
$\|h_{SR}\|^2$	Channel gain between Source and Relay
$\|h_{RD}\|^2$	Channel gain between Relay and Destination
T_s	Duration of a time slot
$\hat{G}$	Equalized AF relay gain
$\hat{a}$	Equalized scaling coefficient

β	Forgetting constant for low-pass filtering
D_m^{hol}	Instantaneous head-of-line packet delay for user *m*
Υ_{SD}	Instantaneous SNR of S-D channel
Υ_{SR}	Instantaneous SNR of S-R channel
Υ_{RD}	Instantaneous SNR of R-D channel
χ_σ	Log-normally distributed variance in shadowing
C	Loop interference cancellation
h_{LI}	Loop interference channel
Υ_{LI}	Loop interference SNR in Full-duplex mode
D_m^{Req}	Maximum packet delay required by user *m*
$\bar{\omega}_m$	Mean bit arrival rate for user *m* at time slot *t*
R_m^{Req}	Minimum average data rate (throughput) required by user *m*
σ_n^2	Noise power (variance)
ρ	Normalizing parameter for the utility function
$PL_{\boldsymbol{m}}$	Path Loss
ψ_m	Performance metric
P_{SD}	Power allocated to subcarrier between Source and Destination
P_{SR}	Power allocated to subcarrier between Source and Relay
P_{RD}	Power allocated to subcarrier between Relay and Destination
$Q_m[k, t]$	Queue associated to user *m* on subcarrier *k* at time slot *t*
γ	Received signal
P_{R}	Relay node power
d_0	Reference distance in the cell

Δh	Residual loop interference channel in Full-duplex
a	Scaling coefficient
Γ	SNR gap
P_S	Source (base station) node power
x	Transmitted symbol
B	Total available cell bandwidth
R_m	Total data rate (throughput) for user *m*
K	Total number of subcarriers in cell
M	Total number of users in cell
P_T	Total transmits power at the source
$U(\bar{R}_m)$	Utility as a function of average data rate (throughput) for user *m*
$U(D_m^{hol})$	Utility as a function of instantaneous head-of-line packet delay for user *m*
σ_{SD}^2	Variance of S-D channel
σ_{SR}^2	Variance of S-R channel
σ_{RD}^2	Variance of R-D channel
λ	Wavelength

Abbreviations

Abbreviations	**Meaning**
3GPP	3rd. Generation Partnership Project
3GPP2	3rd. Generation Partnership Project 2
3G	Third Generation
4G	Fourth Generations
AF	Amplify-and-Forward
AMC	Adaptive Modulation and Coding
AMCS	Adaptive Modulation and Coding Schemes
APA	Adaptive Power Allocation
AWGN	Additive White Gaussian Noise
BE	Best Effort
BER	Bit Error Rate
BPSK	Binary Phase Shift Keying
BWANs	Broadband Wireless Access Networks
BWASs	Broadband Wireless Access Systems
BS	Base Station
BS-MS	Base Station-to-Mobile Station link
BS-RS	Base Station to Relay Station link
CAQA-JSPA	Channel-Aware Queue-Aware for Joint Subcarrier and Power Allocation
CBR	Constant Bit Rate
CC	Coded Cooperation

CF	Compress-and-Forward
CG	Clipped-gain protocol
CDMA	Code Division Multiple Access
CNR	Carrier-to-Noise Ratio
CRA	Continuous Rate Allocation
CSI	Channel State Information
D	Destination node
DF	Decode-and-Forward
DRA	Discrete Rate Allocation
DSA	Dynamic Sub-carrier Assignment
DSM	Delay-based Satisfaction Maximization
EDD	Earliest Due Date
EDF	Earliest Deadline First
ertPS	extended real-time Polling Service
EXP	EXPonential scheduling policy
EV-DO	Evolution-Data Optimized
FC	Frugality Constraint
FCB	Fair Class-Based
FD	Full-Duplex
FER	Frame Erasure Rate
FIFO	First In First Out
FTP	File Transfer Protocol
HD	Half Duplex
HDR	High Data Rate

HIPERLAN/2	High Performance Radio Local Area Network Type 2
HOL	Head-of-Line
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HTTP	HyperText Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IFFT	Inverse Fast Fourier Transform
JFI	Jain’s Fairness Index
JSPA	Joint Subcarrier and Power Allocation
JUES	Joint Urgency and Efficiency Scheduling
LOS	Line of Sight
LTE	Long Term Evolution
LTE-Advanced	Long Term Evolution-Advanced
MAC	Medium Access Control
MIMO	Multiple Input Multiple Output
MLWDF	Modified Largest Weighted Delay First
MS	Mobile Station
MDU	Maximum Delay Utility
MPF	Modified Proportional Fairness
MRC	Maximum Ratio Combining
MSR	Maximum Sum Rate
M-QAM	Multi-level Quadrature Amplitude Modulation
MQS	Maximum Quality-of-Service Satisfaction
MUZF	Multi-User Zero Forcing

NLOS	Non- Line of Sight
nrtPS	non- real-time Polling Service
NRT	Non-Real Time
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OTFSA	Optimal Temporal Fairness Scheduling Algorithm
PDF	Probability Density Function
PF	Proportional Fairness
PLFS	Packet Loss Fair Scheduling
QAM	Quadrature Amplitude Modulation
QG	Quality Guaranteed
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
QSI	Queuing State Information
rtPS	real-time Polling Service
RRA	Radio Resource Allocation
RRM	Radio Resource Management
RS	Relay Station
RS-MS	Relay Station to Mobile Station link
RT	Real Time
R	Relay node
S	Source node
SNR	Signal-to-Noise Ratio
STE	Shortest Time to Extinction

SUI	Stanford University Interim
TDD	Time Division Duplexing
TDMA	Time Division Multiple Access
T1/E1	T-carrier/E-carrier signaling 1
TSM	Throughput-based Satisfaction Maximization
UEPS	Urgency and Efficiency based Packet Scheduling
UGS	Un-solicited Grant Service
UG	Unlimited gain protocol
UPA	Uniform Power Allocation
VG	Variable-gain protocol
VoIP	Voice over IP
WiFi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access

Chapter 1

Introduction

1.1 Background and Motivation

The rapid growth of new services such as online video games, video conferences, and multimedia services is demanding a reliable and an efficient Internet access. Therefore, acceptable quality-of-service (QoS) for guaranteeing user satisfaction and loyalty to the cellular operator (Lima et. al, 2014), is inevitable. The worldwide interoperability for microwave access (WiMAX) (IEEE, 2011) and long-term evolution (LTE) (Dahlman et al., 2013), as broadband wireless access networks; have been considered a viable solution to provide last-mile access to the Internet (Andrews et al., 2007; Dhrona et al., 2009). The WiMAX and LTE network standards have specified mechanisms for handling QoS. The QoS mechanism defines how a service class can perform its requests of bandwidth; it does not define the traffic scheduling algorithm. In other words, the standards do not specify particular scheduling algorithms for implementing the QoS mechanism. However, the flexibility of medium access control (MAC) layer of these broadband wireless access networks (BWANs) allows separate scheduling policies (schemes) to be employed in each QoS service class. Nonetheless, researchers and equipment manufacturers are motivated to keep finding more efficient methods for QoS support at the MAC layer especially when deploying Orthogonal Frequency Division Multiple Access (OFDMA) physical layer. Even with the increased data traffic offered by OFDMA technology, cellular operators still need to guarantee satisfactory provision of the services in order to maintain a high number of subscribers, decrease churn, and attract new subscribers.

Therefore, an efficient utility-based scheduling and resource allocation algorithm needs to be designed. Applying utility functions (a network economics concept) for packet scheduling and regulation of resource allocation is essential. Utility functions capture the satisfaction level of users for a given resource assignment and can be used to establish a utility-based mechanism for resource allocation in OFDMA-based networks, in which the network utility at the level of heterogeneous mixed applications is maximized subject to the current channel quality, queuing conditions and users' QoS requirements.

On many occasions users find themselves on cell-edge areas where network coverage is too poor to permit access. To extend coverage without incurring additional cost of base station deployment or installing multiple-input-multiple-output (MIMO) antennas at the terminals, deployment of relays have been considered a more economical alternative. However, to guarantee fairness in resource allocation and QoS differentiation, efficient resource allocation scheme is required. Most of the proposed algorithms for the multi-user relaying system are yet heuristic and hard to implement as well. Moreover, they do not take advantage of the pre-existing solutions already proposed for non-cooperative multi-user systems (Rasouli, 2012).

Among the existing relay protocols, the amplify-and-forward (AF) offers the most cost-effective implementation. Its closest competitor; the decode-and-forward (DF) protocol introduces significant delay as a result of having to examine every packet header at the router. AF relays forward data without examining network layer headers, and is possible due to the synchronicity of OFDMA systems. In addition, since the AF relays do not decode the packets, channel decoder delays are eliminated, reducing its impact on higher layers. Therefore, AF relays are good candidates for enhancing the coverage in the next generation of wireless network. Providing a fair and QoS radio resource allocation framework for this kind of relay is, therefore, crucial (Sharifian, 2014).

1.2 Problem Statements

Current problem in resource allocation is the design of efficient utility-based scheduling algorithm which will maximize users' QoS and at the same time provide acceptable trade-off between system capacity and resource distribution fairness in mixed services. The need to support a wide range of multimedia applications in end-to-end transmission makes it inevitable for BWAN standards to guarantee the satisfactory provision of the diverse quality of services in the wireless links. To achieve this, an efficient scheduling algorithm that is capable of accommodating three major classes of service such as real-time (RT), non-real-time (NRT) and best-effort (BE) needs to be developed.

In cross-layer resource management architecture, utility function offers a tangible metric for pricing the benefit of making use of certain radio resource to guarantee the practical transmission services (Li et al., 2012). Resource allocation of RT and NRT using a single utility function and a single QoS element such as head-of-the-line (HOL)-delay (Ali & Zeeshan, 2012; Andrews et al., 2000; Andrews et al., 2005; Balakrishnan & Canberk, 2014; Hajjawi & Ismail, 2015; Musabe & Larijani, 2014; Mushtaq et al., 2014; Navaie et al., 2007; Shakkottai & Stolyar, 2001) , or the mean delay (Song et al., 2009; Song et al., 2004; Ukil, 2009), or bit-rate (Agarwal et al., 2007; Chen et al., 2007; Enderle & Lagrange, 2003; Kelly et al., 1998; Kuo & Liao, 2005; Kuo & Liao, 2007; Kushner & Whiting, 2004; Liu et al., 2007; Madan et al., 2010; Pantelidou & Ephremides, 2008; Rodrigues & Casadevall, 2012; Sediq et al., 2012; Sediq et al., 2013; Song & Li, 2005a; Svedman et al., 2007; Wang et al., 2008), or the traffic prioritization (Katoozian et al., 2008; Katoozian et al., 2009; Liu et al., 2007; Wang & Jia, 2010) have been developed and studied in the literature.

Delay-based-only utilities cannot guarantee QoS for NRT and as such any performances obtained for them are only relative. Higher priority connections, in priority-only utilities, starve the lower priority ones. Besides, they provide no absolute performance but only relatively better performance (Lee et al,. 2009). Bit-rate-based-only utilities are relevant to commonly called infinite-backlog scenario, where the queues are assumed to be always full, independent of service. Nevertheless, bit-rate-based utilities are blind to the requirements of RT flows where the requirements are more important.

Scheduling algorithms that are based on joint bit-rate and delay utilities (Al-Manthari et al., 2009; Dhrona et al., 2009; Rodrigues & Casadevall, 2009; Wang et al., 2007) have been designed. This category of scheduling algorithms are, although, designed to apply different utility functions to schedule homogeneous mixed services; they are sometimes found unable to capture the exact nature and characteristic of each service, thereby resulting in poor or limited QoS guarantee. Therefore, a utility-based scheduling mechanism, called the maximum QoS satisfaction (MQS) will be designed for wireless networks to mitigate the problem that current scheduling algorithms cannot have a good performance in mixed services.

Selective (opportunistic) relaying schemes (Bletsas et al., 2006; Duval et al., 2010; Emmanouil et al., 2009; Fareed & Uysal, 2009; Hasan et al., 2011; Kim & Kim, 2009) have been studied and found to aim mainly at maximizing the average system capacity (throughput), without consideration for fairness or/and QoS. The works (Ng et al., 2012) proposed a distributed resource allocation algorithm with different relaying protocols to fulfill heterogeneous QoS requirements. However, the algorithm only enables dynamic selection between AF and DF relaying.

The AF enjoys simpler and less costly implementation; however it unavoidably amplifies noise along with the transmitted signal. To improve its bit-error-rate, the unlimited-gain (UG) (Hasna & Alouini, 2002) and clipped-gain (CG) (Lee, 2009) relay protocols were developed. Although these protocols have been proven to perform a little better than the variable-gain

(VG) AF (Emamian et al., 2002) in terms of bit-error rate performances; however when tested in resource allocation, their average system throughput performances are found to be equal. Riihonen et al. (2009) proposed resource allocation in full-duplex (FD) mode, but they did not consider the delay constraints. Bi and Zhang (2010) proposed relay-assisted resource allocation which achieves same diversity gain as opportunistic scheduling in time division multiple accesses (TDMA) without incurring fairness penalty. However, the algorithm did not consider a system with heterogeneous QoS requirements.

Sharifian (2014) developed a fair packet scheduling and subcarrier allocation algorithms for relay system. However, the utility employed is based on bit-rate only and did not consider mixed traffic. Resource allocation scheme proposed by Rasouli (2012) divides cooperating users into groups and different cooperation coefficient (scaling coefficient) is assigned to each group. The scheme maintains the same level of fairness but obtains higher data rates compared with a similar algorithm without user cooperation. However, the cooperation coefficients are statically allocated to increase data rate on grouped subcarriers. Also taking advantage of a relay station parameter, equalized scaling coefficients will be derived from relay station gain of an AF and used by packet scheduling and resource allocation algorithm to provide not only increases in data rates but improved resource distribution fairness and user satisfaction for AF relaying systems.

1.3 Objectives of the Thesis

The objectives could be summarized as follows:

1. To study and analyze the performances of scheduling algorithms with respect to QoS provisioning in a multi-user heterogeneous service scenario.

2. To design novel utility-based scheduling scheme to maximize user satisfaction while providing efficient trade-off between system efficiency and resource allocation fairness.
3. To formulate and apply a relay station parameter for subcarrier allocation in order to improve user satisfaction while assuring high system efficiency and resource allocation fairness in AF relaying network.

1.4 Scope of the Research

The scope of this research is limited to developing utility-based scheduling algorithm that balances the trade-offs between fair distribution of resources and efficient allocation of the resources on one hand; and on the other hand between QoS satisfaction and efficient allocation of the resources among network users. Many scheduling algorithms exist, but the proposed MQS scheduler will be compared with only delay-based maximization satisfaction (DMS)/ throughput-based maximization satisfaction (TMS), proportional fairness (PF) and maximum sum rate (MSR) algorithms that are considered as references.

The research is also limited to developing a relay-assisted subcarrier allocation that increases system capacity at the same time that improves fair distribution of resources and QoS guarantee among users in relaying networks. The performances in relaying system will be analyzed and compared, using the proposed MQS scheduler and DSM/TSM only. Many performance parameters are available in literature for comparing and analyzing radio resource allocation algorithms. However, for our purpose we only use average system throughput, user call satisfaction and fairness as performance parameters to compare and analyze the scheduling algorithms and the subcarrier allocation schemes in AF relay.

The Stanford university interim (SUI) channel model (Erceg et al., 1999) to model the Rayleigh fading and traffic parameters (Andrews et al., 2007; Stratogiannis et al., 2010) used during subcarrier assignment are only recommended for simulating WiMAX system. However, other channel models exist as well as traffic parameters for other systems such as LTE. Although other analytical evaluation tools such as NS-2/NS-3 and OPNET are available and could be used in the performance analysis of the algorithms developed, our algorithms were developed and analyzed using MATLAB 7.9.0 (R2009b) analytical tool.

1.5 Thesis Research Outline

The remainder of this thesis is organized as follows. Chapter 2 presents detailed concepts of radio resource allocation in OFDMA systems. It also reviews recent and past works in scheduling and resource allocation techniques in network. In Chapter 3, the methodological processes which facilitate the performance analysis of scheduling and subcarrier allocation are detailed. A new scheduling and resource allocation framework is then proposed for wireless network. In order to improve resource allocation fairness and provide QoS guarantee in relay network, a relay station parameter is formulated and applied to scale the backward signal-to-noise (SNR) in the source to relay link. In Chapter 4, results are presented and discussed. Finally, in Chapter 5 the thesis is concluded and some directions for future works are provided.

Chapter 2

Literature Review

2.1 Introduction

Rapid development of wireless communication technologies and systems resulting in new hardware and standards during the last decade has provided ubiquitous high data communication to mobile users. Second-generation (2G) wireless systems were very successful as they enhance the quality of speech communications in cellular systems. Their successes prompted the development of Third-generation (3G) wireless systems. While 2G systems such as GSM, IS-95, and cdmaOne were designed to carry speech and low-rate data, 3G systems were designed to provide high-data-rate services.

During the evolution from 2G to 3G, a variety of wireless systems, including GPRS and IMT-2000 have been developed (Hui & Yeung, 2003). Increasing user demands in terms of data rate in 3G standards soon led to the standardization of the 4G mobile telecommunications systems. However, the key advantages of 4G systems depends highly on exploiting the physical layer access opportunities offered by orthogonal frequency division multiplexing (OFDM). OFDM is an efficient and robust modulation technique which is capable of combating fast variations and the frequency selectivity of the radio channel.

OFDMA builds on OFDM to provide another degree of freedom by allowing dynamic assignment of subcarriers to different users at different time instances, to take advantage of the fact that at any time instance channel responses are different for different users at different subcarrier frequencies (Ali et al., 2007). Therefore, in OFDMA both frequency and multi-user diversities are present. Different sources of diversity that can be exploited in communication include: time diversity, frequency diversity, spatial diversity and multiuser diversity.

Frequency diversity occurs when different sub-carriers of a broadband wireless system have a strongly varying attenuation (Rodrigues & Casadevall, 2009). Multiuser diversity is achieved when opportunistic transmission system intelligently utilizes the independent fading variation to serve only the users with the strongest channel; because it is highly probable that one of the user channels lies in its own peak state among a large number of users (Lee et al., 2009).

The time-varying multipath fading of the radio channel limits the ability of the networks to achieve optimum spectral efficiency and provide quality-of-service guarantee for users. Since radio resource is limited and scarce in wireless networks, resource allocation between users is a crucial issue. The channel quality of each user may vary over time. With the given available resource, there are different criteria or strategies to assign resource to each user, which leads to different benefits, such as system throughput, user fairness, or the Quality of Service (QoS) of user data flow (Chen et al., 2011). Therefore, this chapter presents concepts and techniques of scheduling and resource management for OFDMA-based wireless systems.

2.2 Wireless Channels

The allocation and management of resources are crucial for wireless networks, in which the scarce wireless spectral resources are shared by multiple users. However, wireless channels suffer from time-varying multipath fading; moreover, the statistical channel characteristics for different users usually are different (Song, 2005). This section describes the channel impairments that affect the wireless transmissions and provide the fairly general mathematical and statistical models that are used in the sequel.

Wireless channel is characterized by three different attenuating effects: multipath fading, large-scale path-loss and shadowing. The small-scale fading or fading models the fluctuations of the signal strength around the average for a particular location. The large-scale path-loss models the average received signal strength for a given transmitter-receiver (T-R) separation, while location-dependent variation of large-scale path-loss is modeled by shadowing.

2.2.1 Small Scale Fading: Multipath

In wireless mobile communication systems, a signal travels from the transmitter to the receiver over multiple paths. Multipath arises because the propagated signal is reflected, diffracted, and scattered by the objects present in the channel environment.

• Reflection occurs when a propagating signal falls on a surface with a dimension much larger than the signal wavelength, λ.

• Diffraction occurs when the electromagnetic waves encounter an impenetrable obstacle. Secondary waves are then formed and diffracted field can even reach a shadowed receiver.

• Scattering occurs when the objects in the channel environment causes the reflected energy to spread out in all directions.

2.2.2 Large Scale Fading: Path-Loss and Shadowing

If the radio transmission is propagated in an ideal free space, perfectly uniform and non-absorbing, the attenuation of RF energy between the transmitter and the receiver will behave according to an inverse square law: the received power expressed in terms of transmitted power will be attenuated only by the path-loss factor. Shadowing represents the average signal power attenuation or path-loss resulting from radio propagation over a large area (Sklar, 1997).

It occurs when a large obstruction such as a hill or building obscures the signal path between the transmitter and receiver. Shadowing causes slow fading. Measurements have established that the large-scale path-loss for a particular T-R separation is random and distributed log-normally around the mean value described by the path-loss formula. Therefore, the standard deviation of a log-normally distributed random variable representing the effect of shadowing is $\mathcal{X}_\sigma$ in dB.

The small scale fading superimpose on the large scale fading. The small scale fading or multipath fading is due to the presence of many objects in the environment that induce a fluctuation in the receiver signal's amplitude, phase, and angle of arrival. Thus, the receiver observes multiple path and time delayed versions of the transmitted signal and the received signal can be characterized by the large- and small scale fading. Furthermore, the receive signal is corrupted by additive noise and interference at the radio receiver input. The receive signal $r(t)$ is generally written in terms of a convolution between the transmitted signal $x(t)$ and the impulse response of the channel $h(t)$:

$$r(t) = x(t) * h(t) + n(t) \tag{2.1}$$

where $n(t)$ is the noise contribution at the receiver input and t is the time variable. A mobile radio signal can be written in terms of two components $e(t)$ and $h_0(t)$ (Lee, 1986):

$$r(t) = x(t) * e(t). h_0(t) + n(t) \tag{2.2}$$

where $e(t)$ is the large scale fading which is log-normally distributed and $h_0(t)$ is the small scale fading. With the relative motion of the transmitters, receivers and scattering objects, the multipath fading manifests itself in a time-spreading and time-variant phenomenon. For signal dispersion, the fading degradation can be categorized as frequency-flat or frequency selective. Similarly, the time-variant degradation can be categorized as fast- or slow fading. The following section presents the statistical model used to characterize the channel effects of Rayleigh fading.

2.2.3 Statistical Model of Rayleigh Fading Channel

The small scale fading can be modeled with the Rayleigh fading model given that the envelope of the receive signal is distributed according to a Rayleigh probability density function (PDF) can be written as:

$$P(a|\sigma) = \begin{cases} \frac{a}{\sigma^2} e^{-\frac{a^2}{2\sigma^2}} & for\ a \geq 0 \\ 0 & otherwise \end{cases} \tag{2.3}$$

where a is the envelope of the received signal and $2\sigma^2$ is the mean power of the multipath signal. This fading model applies to a fairly common situation in mobile radio environment when all multiple reflective waves are received from the surrounding and there is no line-of-sight component. The SNR (or power of the received signal) γ is exponentially distributed. The exponential PDF of γ can be expressed as follows:

$$P(\gamma) = \frac{1}{\bar{\gamma}} e^{-\frac{\gamma}{\bar{\bar{\gamma}}}} \tag{2.4}$$

where $\bar{\gamma}$ is the average of γ over all the channel realizations. Rayleigh fading is a reasonable statistical model for signal propagation over a highly built-up environment. When there are many objects on the path that scatter the radio signal and there is no line-of-sight propagation, the central limit theorem stipulates that the channel impulse response can be modeled by a Gaussian process and the Rayleigh distribution is therefore an appropriate assumption for the envelope of the response of the channel.

2.3 OFDM and OFDMA

In IEEE 802.16e standard (Fu et al., 2010), an OFDMA technique is used for multiple accesses and MIMO system is applied to increase data rate. Hence, the combination of OFDMA and the MIMO antenna solution have been found to be the key element of 4G radio telecommunication systems physical access (Chia et al., 2008; IEEE, 2009; Loa et al., 2010). The 4G systems are based on OFDMA, and some of their characteristics and requirements are: wider bandwidth, spectrum flexibility, lower latency, improved system capacity and coverage, reduced overall cost for the operator and packet-optimized radio access technology with enhanced peak data rates of 100 Mbps and 1 Gbps for high and low mobility, respectively. Therefore, current and future wireless network standards have adopted OFDM and OFDMA as the physical layer technology of choice.

OFDM is one of the most promising multi-carrier transmission schemes in wireless communication systems (Wang & Dittmann, 2010), as it provides frequency diversity and eliminates impulsive noise. Figure 2.1 illustrates the physical (PHY) layer of OFDM. In this layer, a high data-rate data stream is divided into N parallel low-rate data stream, $X_k,\ k = 1,2, \ldots, N$. The narrow-band signals are modulated by orthogonal subcarriers using IFFT. A guard interval that is greater than the delay spread is added between the OFDM symbols to eliminate inter-symbol interference (ISI). A cyclic copy of OFDM symbol is inserted in the guard interval. By adding a cyclic prefix (CP) to each OFDM symbol, the channel appears to be circular if the CP length is longer than the channel length thus removing the ISI caused by frequency-selective fading. The OFDM symbols are then modulated by a carrier of higher frequency after passing through the parallel to serial convertor. The reverse action takes place to regenerate the high-rate data stream at the receiver.

OFDM essentially splits the radio signal into a group of mutually orthogonal subcarriers, each having a much smaller bandwidth than the coherence bandwidth of the channel (Lataief & Zhang, 2006). Because, the subcarrier bandwidth is very small, the fading process can be assumed flat over each subcarrier.

Multiple-access strategies typically attempt to provide *orthogonal*, or noninterfering, communication channels for each active link. The most common way to divide the available dimensions among the multiple users is through the use of frequency, time or code. Frequency division multiple accesses (FDMA) divides the radio resource into non-overlapping frequency bands (subcarriers) (Andrews et al., 2007). In time division multiple accesses (TDMA), the radio resource is divided into non-overlapped time slots and each user is given a unique time slot, either on demand or in a fixed rotation. Wireless TDMA systems almost invariably also use FDMA in some form, since using the entire electromagnetic spectrum is not allowable. Code division multiple access (CDMA) systems allow each user to share the bandwidth and time slots with many other users and rely on orthogonal binary codes to separate out the users (Andrews et al., 2007). CDMA employs spread-spectrum technology which assigns distinct codes to different users.

OFDMA is essentially a hybrid of FDMA and TDMA whereby users are dynamically assigned orthogonal subcarriers (FDMA) in different time slots (TDMA). Orthogonality of subcarriers allows the subcarriers to overlap, and hence provides a more efficient spectrum utilization compared to FDMA. In relaying system, the most common and simplest duplexing has been achieved through TDMA for practical systems such as IEEE 802.16j standard (Soldani & Dixit, 2008); although, duplexing can also be achieved with FDMA scheme. However, Yin and Alamouti (2006) noted that OFDMA is a superior access technology compared to TDMA and CDMA, because the fine granularity of resources for OFDMA increases the flexibility and efficiency of OFDMA compared to the other schemes.

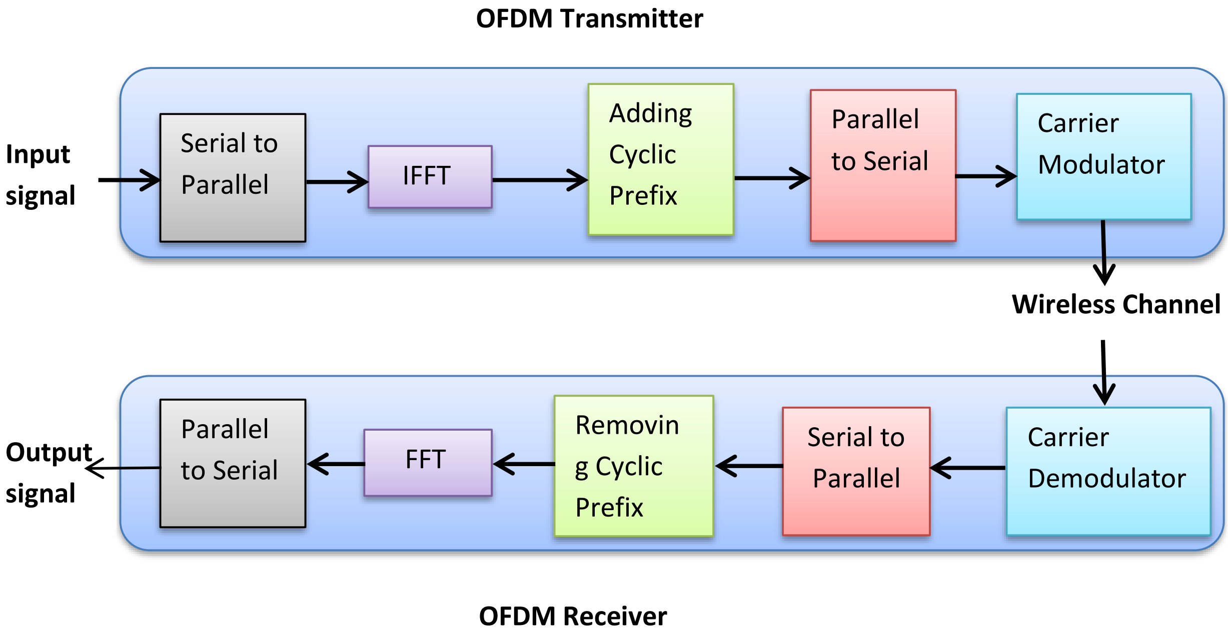


Figure 2.1: OFDM transceiver system model (Patel et al., 2006)

2.4 Physical and Media Access Control System (MAC) Layers

OFDM and OFDMA are usually combined and used as a multi-user access technique where each subcarrier can be allocated to different users at different times. Therefore, they provide both the physical (PHY) and media access control (MAC) layers for radio resource allocation (RRA). To guarantee QoS for all the users in the system, a scheduler at the MAC level is required in the system to manage subcarrier allocation among the users in the system. Figure 2.2 illustrates the PHY/MAC layer issues of an OFDM system. The effective transmission rate of the user can be calculated by knowing the coding rate and modulation order.

Subcarrier and power allocations are crucial to resource management; because allocating the subcarriers to the users with better channel conditions improves the total rate of the system. The subcarrier gains of the users are retrieved at the receiver by inserting the known pilot signals intermittently in the original OFDM signal at the transmitter. After subcarriers are all assigned to the users according to the subcarrier and power allocation algorithm, the adaptive coding and modulation scheme is chosen according to the QoS requirements of the user.

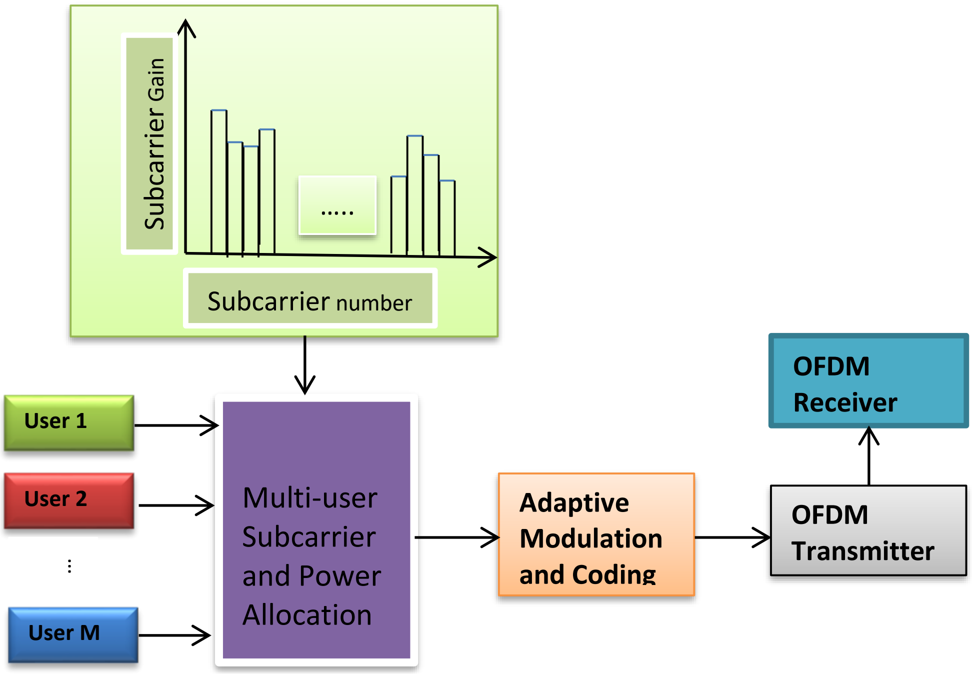


Figure 2.2: OFDM PHY/MAC layer system model (Rasouli, 2012)

2.5 Scheduling and Resource Allocation for OFDMA Systems

Scheduling is a general term used to describe the sharing of limited resources among the contending users; it refers to a method of allocating resources to users over time by following some predetermined procedures (Lee et al., 2009). Since resources for wireless communications are scarce and constrained, scheduling algorithms try hard to allocate resources in an optimal way based on some criteria.

There are different types of scheduling algorithms, with each achieving what it considers as optimal. A simple form of scheduling policy/algorithm proposed in for wired network is the round-robin policy, which selects users according to a fixed cyclic order, thereby yielding a fair allocation of transmission opportunities. However, when there are urgent packets to deliver within certain time duration, the round-robin policy may fail them as they have to wait until their turns come (Lee et al., 2009). Moreover, scheduling algorithms proposed for wired networks are usually not applicable to wireless systems which are confronted with time-varying multi-path fading and location-dependent path loss.

2.6 Subcarrier and Power Allocation

The success of emerging BWA networks will depend on their abilities to manage the shared wireless resources such as bandwidth and power in the most efficient way (Beres & Adve, 2008). The bandwidth is a key wireless resource which refers to the range of frequencies (denoted by Hz) occupied by a transmitted signal. Since bandwidth determines the maximum symbol transmission rate, it puts a fundamental limit on the channel access rate (Lee et al., 2009). Power is also a wireless resource which contributes to the system capacity in a nonlinear manner. Since the achievable data rate is a function of the power allocation (i.e. the data rate increases nonlinearly with the transmit power and vice versa), it is expected that adaptive power allocation (APA) can further improve the system data rate (Ali et al., 2007).

Dynamic subcarrier allocation (DSA) to multiple users can be employed to improve the system data rate. DSA algorithms are categorized as either “margin adaptive” (Wong et al., 1999; Zhang, 2004) or “rate adaptive” (Shen et al., 2005; Song & Li, 2003a; Song & Li, 2003b; Song & Li, 2005b; Song & Li, 2005c). Margin adaptive algorithms aim at minimizing the total transmits power in the system while maintaining each user with its required quality of service requirements; such as data rate and bit-error-rate (BER). In rate adaptive algorithms, the optimization objective is to maximize the total system throughput with the constraint on the total transmit power in the system (Hoo et al., 2004). Ideally, to achieve the optimal solution for the optimization objectives, subcarriers and power should be jointly allocated.

The resource allocation problem for a single user is to maximize the total achievable throughput over allocated subcarriers subject to a total power constraint, which can be optimally solved by means of the water-filling method (Song & Li, 2003b). Water-filling approach pours more power into high-gain subcarriers and allocates no power to the low-gain subcarriers (Tse & Viswanath, 2005). In OFDM/OFDMA system, both the subcarrier and power are jointly managed (allocated) across multiple users in order to maximize the resource allocation efficiencies.

However, joint optimization of subcarrier and power poses a prohibitive computational burden at the base station, especially, in the presence of channel variation. Hence, low-complexity suboptimal algorithms are preferred for cost-effective and delay-sensitive implementations. Low complexity can be achieved by separating the subcarriers and power allocation, because the number of variables in the objective function is almost reduced by half. Jang and Lee (2003) proved that allocating the subcarriers to the users with the highest gain, and then allocating the total power to the subcarriers according to water-filling (Goldsmith, 1997), maximizes the total throughput of the system. They also showed that the throughput of the system degrades negligibly by flat power allocation compared to water-filling approach when high-gain subcarriers are allocated to the users in the system. However, the works (Shen et al., 2003; Tung & Yao, 2002) noted that subcarrier allocation provides more gain than power allocation and that power allocation does not offer substantial gains at high SNRs. Therefore, Rhee and Cioffi (2009) suggested a flat power allocation among the subcarriers in order to reduce the complexity of resource allocation problem.

Different rate allocation strategies such as continuous-rate adaptation (CRA) and discrete-rate adaptation (DRA) have been proposed to take advantages of fading on the wireless channel. The CRA is based on realistic Shannon capacity using the SNR gap. The DRA is a process of using discrete adaptive modulation and coding (AMC) levels, where SNR thresholds for switching between different AMC modes were determined for a target BER. In order to take advantage of fluctuations in the wireless channel AMC have been analyzed for systems relying on adaptive transmissions (Alouini & Goldsmith, 2000; Chung & Goldsmith, 2001; Goldsmith & Chua, 1997; Goldsmith & Chua, 1998). AMC allows resources, at the physical layer, to be adjusted in accordance with the channel quality so that higher (lower) rate and power is allocated as the channel quality increases (respectively decreases) (Wang et al., 2007). AMC is an effective way to increase the spectral efficiency in a time varying wireless channel. By dynamically adapting the modulation and coding scheme that can be supported by the signal-to-noise ratio (SNR), the user data rates and hence the capacity can be maximized.

The quality of the received signal depends on a number of factors such as the distance between the transmitter and receiver, path loss exponent, log-normal shadowing, short-term fading and noise. This implies that the signal to interference plus noise ratio (SINR) varies over time, frequency and/or space. For example, a Rayleigh fading channel causes intermittent reduction in the power level of the received signal, which is known as *deep fades*. During *deep fades* burst errors occur and poor bit error rate (BER) is obtained, while in between fades the received signal level is good and a better BER is obtained. Therefore, transmitter needs a good estimate of the channel to be able to implement the adaptive modulation and coding.

The data rate reduction caused by using adaptive modulation and coding in the system is modeled by the SNR gap, which is defined as the difference between the SNR needed to achieve a data rate for a practical system and the theoretical SNR (Jang & Lee, 2003). The advantages of AMC-based systems over non-adaptive alternatives have been documented (Yanikomeroglu & Zhang, 2008). AMCS have been adopted in current and future standards of most BWAS**.** For example, the 802.16/WiMAX network can use either low efficiency modulations (binary phase shift keying (BPSK) with coding rate 1/2) or very high efficiency ones (64-QAM with coding rate 3/4), depending on the SNR. If the SNR decreases, change is made to a more robust modulation and coding such as BPSK with coding rate 1/2, to improve the performance (data throughput), otherwise a less robust profile can be switched to. Thus, users with better SNR (users closer to the BS) get higher order modulation; those farther from the BS get lower order modulation, ensuring the best performance for each user within the BS coverage.

An OFDMA scheme with adaptive power, in which subcarrier allocation itself plays a very significant role in maximizing the total throughput, by using multiuser diversity (Dañobeitia & Femenias, 2010), is an important consideration in BWASs. The works (Dañobeitia & Femenias, 2011; Femenias et al., 2012), observed that the APA-based strategies improve the performance of uniform power allocation (UPA)-based ones. However, they also found that the performance improvement, although noticeable for discrete rate-based systems, becomes almost negligible when using continuous rate-based schemes; therefore suggesting that using AMC schemes with a large set of modulation formats combined with powerful channel codes with adaptive coding rates can make unnecessary the use of power allocation strategies.

2.7 Cross-Layer Resource Allocation

Scheduling algorithms which take into account only the variability of channel characteristics are classified as channel-aware. In the past, resource allocation schemes were designed using only physical-layer power-bit loading processes (Kullarni et al., 2005; Leke & Cioffi, 1998; Miao et al., 2010; Shakkottai & Srikant, 2002; Yao & Giannakis, 2005). In such designs, data bits and transmitting powers are adjusted across subcarriers to efficiently utilize the network resources. However, although such approaches are simple to implement, without knowing the upper-level packet arrival characteristics or the queuing/buffer conditions, the single layer resource allocation cannot guarantee each user's specific QoS requirements (Javidi & Kittipiyakul, 2004).

Girici et al. (2010) developed an accurate finite buffer queuing model for a channel-based scheduling scheme to obtain optimal buffer partitioning based on that model. Efficient buffer management, essential to ensure good system performance, can be achieved when scheduling algorithms take into account the users' queue backlogs in addition to the channel conditions. More recent studies have proposed cross-layer resource allocation designs to account for both PHY-layer channel conditions and queue dynamics from upper layers (Mokari et al., 2010). Juyeop et al. (2005) presented a cross-layer adaptation framework for interlayer operation between the MAC and PHY layers, and a design example of primitives to exchange PHY information for cross-layer protocol operation. They show that average cell throughput can be significantly improved by applying carefully cross-layer adaptation schemes.

2.8 Utility Function in Resource Allocation

Utility theory was originally conceived for applications in economics (Ganz & Wongthavarawat, 2003), but found attention in communication, especially in scheduling design. In communication networks, utility theory have been used to evaluate the degree to which service requirements of users' applications are met (Kelly, 1997). Utility theory has also been applied to formulate quantitatively the relations between user experience and various network performance metrics (Keeney & Raiffa, 1994). Therefore, a utility function provides a measure of how satisfied a user is in terms of resources allocated to it, and quantifies the benefit of usage of certain resources (Andrews, 2004). A utility function is defined as a curve mapping the amount of network resources received by the application to the performance as perceived by the end-user.

The scheduling rules (Andrews et al., 2001; Kelly, 1997; Knopp & Humblet, 1995; Shenker, 1995) have been considered gradient-based utility functions, i.e. they represent marginal utility (derivative of utility) functions in terms of some QoS measures. For example, the utility function in Knopp and Humblet (1995) is utility function with respect to average throughput $r_i[n]$, i.e. $U_i(r_i[n])$ and which are concave; meaning that the utility of the average throughput for user i during time slot n increases as the average throughput itself increases. The work (Andrews et al, 2001) is a utility function with respect to delay $D_i[n]$, i.e. $U_i(D_i[n])$ and it is a non-concave one, because as the delay of user $i's$ packet in time slot n increases his utility decreases. However, if the marginal utility is of greater interest, then for a linear utility function, $U_i(r_i[n]) = r_i[n]$ the marginal utility function is $U_i'(r_i[n]) = 1$. In this case, subcarrier assignment is made independent for different subcarriers, which means that the assignment of a subcarrier does not affect assigning other subcarriers (Song & Li, 2005a). Shenker (1995) showed that if utility functions are concave with respect to the instantaneous data rates then, for a given number of users, the total utility in the system is maximized when resources are evenly distributed among all users. Utility function can be linear, concave or convex and can be produced by exponential, logarithmic and power methods.

The choice of individual type of utility function depends on the desired rate of change of the utility with respect to the performance metric being considered. However, for a more general scenario in which the utility functions are nonlinear, assigning different subcarriers is not independent anymore, and DSA becomes very complicated (Song & Li, 2005a). Recently nonlinear utility functions have been widely applied in resource allocation schemes for QoS provisioning in multimedia wireless networks (Choi et al., 2009; Jiang et al., 2005; Lee et al., 2005; Lei et al., 2007; Miao & Himayat, 2008; Ryu et al., 2005). Lee et al. (2005) showed that a utility-based approach can improve system capacity over conventional proportional-fair resource allocation schemes.

Miao et al. (2008) noted that a utility based framework may be considered as a QoS-aware framework for resource allocation, where differing QoS requirements across a mix of service types may be directly included in the optimization objective. The key advantage of utility function is that it can inherently reflect the QoS requirements such as delay and throughput of the end user and quantify the adaptability of the application (Ermini, 2011). However, Ermini (2011) noticed that solution of utility-based resource management in wireless networks depends not only on the utility curves or shapes, but also on wireless channel quality, which can differ significantly from user to user, and over time. However, the formulation of utility functions for multimedia traffic remains a problem. Ali and Zeeshan (2012) adopted the sigmoid utility function; Navaie et al. (2007) applied bell-shaped sigmoidal marginal utility functions to schedule both RT and NRT traffics which require different QoS requirements; Ganz and Wongthavarawat (2003) used linear and convex utility functions for adaptive bandwidth allocation; Kwon et al. (2002) formulated the utility function for each class of traffic to reflect its nature of adaptability and Cho and Chong (2005) constructed utility functions using subjective values from the authors' experiments. The idea (Cho & Chong, 2005) is based on the assumptions that all traffics are bandwidth (throughput) sensitive and therefore their QoS provisioning should be based on bandwidth metrics and bandwidth requirements in accordance to their adaptability. However, none of these schemes provide the method to capture the nature of each application in the utility function and map their allocated bandwidth to their QoS requirements.

2.9 Quality-of-Service (QoS) in BWAS

Broadband wireless access systems (BWASs) such as high-speed downlink packet access (HSDPA) (Forkel et al., 2005), WiMAX and LTE are promising and attractive; however, they present more challenging issues: one, the wireless channel is characterized by fast-fading due to user mobility, two it must support a wide range of multimedia applications with diverse Quality of Service (QoS) requirements. The goal of QoS in a BWAS is to guarantee the ability of a network to provide predictable services including dedicated bandwidth, controlled latency and jitter, improved error rate, and several others. QoS involves resource reservation control mechanisms to ensure the provision of the target service in a certain quality level. Therefore, QoS guarantee is not possible unless it is supported by a call admission control in order to block users when there is not enough capacity to support such guarantees (Al-Manthari et al., 2009).

However, time-varying nature of wireless channel makes deterministic QoS guarantee infeasible, hence statistical guarantee for delay requirement have been considered, that is, most packets experience a specified delay (Lee et al., 2009). Depending on the traffic type, three classes of service such as BE, NRT and RT and the associated QoS requirements and priorities must be accounted for in wireless communications (Femenias et al., 2012). Since 802.16/WiMAX is the network that mostly contributed to the success of OFDMA technique, many works on traffic scheduling in OFDMA-based systems refer to this network, or consider it as case study.

The WiMAX standards specify five types of QoS services for both the uplink and downlink scheduling policies: unsolicited grant service (UGS), real-time polling service (rtPS), extended rtPS (ertPS), non-real-time polling service (nrtPS) and BE. Table 2.1 shows the five different service flows and the related applications as well as their QoS specifications and parameters. Each QoS class has its own specification to describe the traffic property and the QoS requirements. Table 2.2 contains QoS parameters for LTE standards, which is the closest rival to WiMAX. The two tables provide a means to compare them in terms of their different QoS mechanisms.

In WiMAX systems, the QoS service types can be generally divided into two categories, one is for delay-sensitive flows such as UGS, ertPS, rtPS, and the other is the delay-tolerant flows such as nrtPS and BE sessions (Goldsmith & Chua, 1998). The UGS and ertPS belong to *conversational class*. Conversational class is the most sensitive class which delivers two-way interactive real-time traffic flow. In the conversational class, the transfer time should be low enough and the time variation (jitters) between information entities should be preserved. The maximum transfer delay is given by human perception of video and audio conversation, and therefore cannot be delayed more than 100*ms* in end-to-end delivery.

The QoS requirement for a UGS flow such as Voice over Internet Protocol (VoIP) without silence suppression, T1/E1 is a guaranteed amount of bandwidth allocation (with constant-bit rate (CBR)) without delay or contention. ErtPS is similar to UGS except that ertPS is designed for VoIP with silence suppression and allocation is dynamic while that of UGS is fixed.

The NrtPS belong to *interactive class*. NrtPS is designed to provide a variable sized data rate for non-real-time traffic transmission such as File Transfer Protocol (FTP), HyperText Transfer Protocol (HTTP) or web browsing. The QoS requirement for nrtPS is minimal bandwidth for transmission. The applications in this class require the remote equipment to send a requesting message and the corresponding data is sent back in response. Therefore, the traffic pattern is bursty in nature; a relatively long time elapses between transmissions of requests, and the data amount of response is generally much larger than that of the request message.

The RtPS belong to *streaming class*. RtPS is designed to provide a variable bit rate (VBR) for real-time traffic transmission such as video and audio streaming. A distinctive difference of streaming class from conversational class is that its service is one-way transport and has no interactivity. This class also requires no preservation of time relation (jitters) between information entities. The QoS requirement for an rtPS is a specific amount of bandwidth allocation within its maximal-tolerant delay requirement.

The BE traffic falls into *background class* and they include applications that are not time sensitive. BE provides a variable bit rate for message-based services such as E-mail, the short message service, and downloading of database. The BE has no QoS guarantee on bandwidth or delay for such services. In practical systems real-time services such as UGS, ertPS and rtPS require the scheduling algorithm to be able to transmit its data before deadline expiration, otherwise such data are useless for the receiver, and then dropped, causing a degradation of the service (Ermini, 2011).

The interval of time within a packet must be transmitted/received, is called “*deadline*". In other words, deadline is the time before expiration of maximum latency. Probably, one of the first and more significant studies on scheduling real-time services in wireless systems with the objective of meeting *deadline* requirements have been carried out by Shakkottai and Srikant (2002). The authors observed that, in wireless systems, there exists a connection between allocating channels in “good" state and meeting deadline requirements. They noticed that some polices like EDD (Earliest Due Date), also known as EDF (Earliest Deadline First), or STE (Shortest Time to Extinction), recognized to be optimal in wire-line systems are not always optimal policies for wireless systems.

Table 2.1: WiMAX QoS and Traffic Parameters (Andrews et al., 2007; Stratogiannis et al., 2010)

QoS Category	**QoS Specifications**	**Services**	**Min. reserved traffic rate (bps)**	**Max. sustained traffic rate (bps)**	**Max. Latency (ms)**	**Max. Jitter (ms)**
Unsolicited Grant Services (UGS)	- Maximum Sustained Rate - Maximum Latency Tolerance - Jitter Tolerance	VoIP	N/A	64000	<20	150
Real-time Polling Services (rtPS)	- Maximum Reserved Rate - Maximum Sustained Rate - Maximum Latency Tolerance - Traffic Priority	Streaming Audio / Video	64000	500000	30	160
Extended real-time Polling Services (erPS)	- Maximum Reserved Rate - Maximum Sustained Rate - Maximum Latency Tolerance - Jitter Tolerance - Traffic Priority	VoIP, voice with activity detection	25000	64000	20	150
Non- real-time Polling Services (nrtPS)	- Maximum Reserved Rate - Maximum Sustained Rate - Traffic Priority	File Transfer Protocol (FTP)	45000	500000	100	300
Best-effort (BE)	- Maximum Sustained Rate - Traffic Priority	Data transfer, web browsing (HTTP), etc.	1000	6400	N/A	N/A

Table 2.2: LTE QoS class identifier (QCI) and traffic parameters (Alasti et al., 2010)

QCI	Resource Type	Priority	Packet Delay (ms)	Packet error loss rate (ms)	Services
1	Guaranteed bit rate (GBR)	2	100	10^{-2}	Conversational voice
2		4	150	10^{-3}	Conversational video (live streaming)
3		3	50	10^{-3}	Real time gaming
4		5	300	10^{-6}	Non-Conversational video (buffered streaming)
5	Non-Guaranteed bit rate (Non-GBR)	1	100	10^{-6}	IMS signaling
6		6	300	10^{-6}	Video (buffered streaming), TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
7		7	100	10^{-3}	Voice, video (live streaming), interactive gaming
8		8	300	10^{-6}	Video (buffered streaming), TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
9		9		10^{-6}	

2.10 Utility-Based Cross-Layer Scheduling and Resource Allocation

When the algorithm focuses only on maximizing throughput without concern for fairness, it can be classified as throughput-oriented. But when objective is to maximize fairness, it is classified as fairness- oriented scheduling algorithm. When the objective is to maximize QoS without sacrificing both throughput and fairness, the scheduling algorithm is classified as

QoS- oriented. Every scheduling algorithm invariably employs a utility function to determine resource usage or distribution of resources among the various users.

Throughput- and fairness-based scheduling algorithms are only suited for BE traffic scheduling because their utility functions contain only QoS metric. QoS-based scheduling algorithms can be designed to use a single utility function for different class of services or use different (multiple) utility functions for different class of services. The categorization in terms of service class to which each type of algorithm belongs is illustrated by Figure 2.3. The symbol U(BE) represents a single utility function designed to allocate resources for only a BE traffic. U(BE, nrtPS, rtPS, ertPS) denotes a single-utility function which can allocate resources to any of the four traffic classes listed therein and (U_{BE}, U_{nrtPS}, U_{rtPS}, U_{ertPS}) is a multiple-utility function, in which each utility function is dedicated to a traffic class.

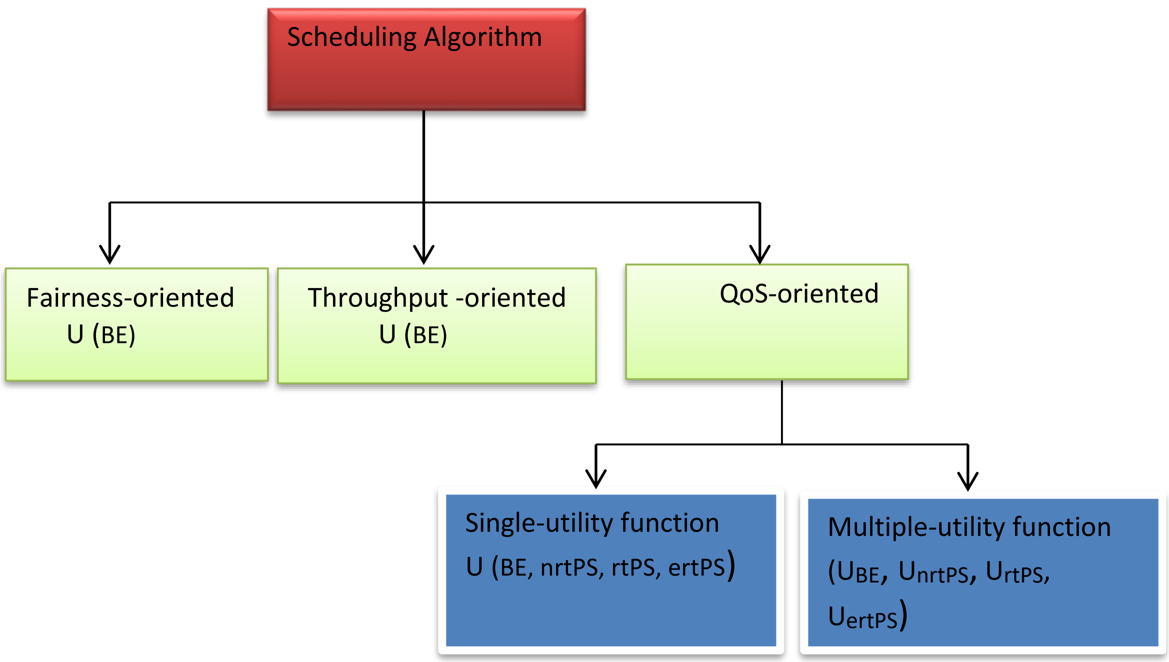


Figure 2.3: Categorization for BWAS scheduling algorithms

Utility-based scheduling decisions can be made using only the users' instantaneous channel conditions, but this can lead to inefficient utilization of available resources. However, combining users' instantaneous channel, queuing and QoS conditions can provide cross-layer benefits such as efficient trade-offs between throughput and fairness on one hand, and between throughput and QoS.

2.10.1 Throughput-Oriented Scheduling Algorithms

Throughput-oriented scheduling algorithms opportunistically allocate network resources. The term "opportunistic" means that the resources will be dynamically allocated based on users' instantaneous CSI to achieve maximum capacity. Knopp and Humblet (1995) proposed a maximal-sum-rate (MSR) scheduling to allocate the time slot to the strongest channel user and thus maximize the total system capacity. To maximize the capacity, the MSR takes advantage of the independent channel variations across users to substantially improve the network capacity through *multiuser diversity*, whose gain increases with the number of users (Liu et al., 2001). The MSR was initially designed for single-carrier network to maximize the total system throughput, $\sum_m r_m [k,t]$ where $r_m[k,t]$ is the achievable transmission rate per Hz (determined by Shannon capacity formula) for m^{th} user at k^{th} subcarrier at time slot t. However it can be extended to multicarrier networks such as OFDMA to exploit the multiuser diversity gain.

The core idea of opportunistic scheduling is to schedule a user with good channel conditions to transmit packets. In this case, the users in more favorable channel conditions may get more chances to be served than the other users in poor channel conditions, which results in unfairness in resource allocation among users whose channel conditions, are not identically distributed. In other words, those users experiencing bad channel quality conditions may suffer from starvation (Femenias et al., 2012). Therefore, MSR sacrifices maximal system throughput for an unfair resource allocation for users with poor channel conditions.

2.10.2 Fairness-Oriented Scheduling Algorithms

Improving system efficiency without consideration for resource allocation fairness has been considered insufficient, since it could churn out the subscribers who are unfairly treated. As a result, attention has been focused on scheduling algorithms that can effectively provide trade-off between spectral efficiency and fairness. Throughput-oriented schedulers 'opportunistically' serving only the users with favorable channel quality conditions raise the issue of fairness. Issues of fairness arise whenever something is shared. Fairness can be defined in many ways, including equality of utility derived from the network, such as user average throughput or equality of opportunity to use network resources, such as amount of time during which a station is permitted to transmit. Therefore, if all channels are independent and identically distributed, each user gets the same number of transmission opportunities as long as it can wait for enough time (Lee et al., 2009).

The objective of resource allocation schemes with rate-adaptive optimization often is to improve the overall system performance, such as throughput and fairness with constraints on power consumption. Two main fairness definitions have been identified: resource- or QoS-based. In the former, fairness is related to the equality of opportunity to use network resources, for example the number of frequency resources a user is allocated or the amount of time during which a user is permitted to transmit. In the latter, fairness is associated with the equality of utility derived from the network, e.g. flow throughput, delay distribution. However, modern broadband wireless access networks (BWANs) are expected to deal with multimedia services with diverse quality-of-service (QoS) requirements. But satisfying QoS and at the same time providing fairness are a conflicting requirement; because one cannot be fair if the strict QoS requirement of each user must be satisfied, especially when heterogeneous mixed traffics have to be simultaneously scheduled.

The round-robin scheduler (Shen et al., 2003) achieves absolute fairness by allocating the same number of time slots to all the users in a round-robin fashion. Since this scheduling algorithm does not consider the channel condition, it is not implementable in wireless system. Three representative types of fairness that have been proposed are PF (Jalali et al., 2000; Kelly et al., 1998; Kim & Han, 2005), temporal fairness (Liu et al., 2001) and max–min fairness (Rhee & Cioffi, 2000) to provide a trade-off between throughput and resource allocation fairness. The max-min fairness algorithm (Rhee & Cioffi, 2000) provides maximum fairness by allocating equal amount of throughput to each user, but it achieves this by maximizing the worst users' capacity. Because this results in all users achieving a similar data rate, min-max cannot accommodate different levels of service because different users require different data rates (Shen et al., 2005).

Kim et al. (2008) modified the proportional fair (PF) scheduling algorithm for systems with multiple carriers such as OFDMA systems whereby the scheduler assigns users each subcarrier to maximize the sum of logarithmic transmission rate. The scheme achieves noticeably better performance in delay distributions, although gives the similar performance as PF scheduling in long-term throughput. Even though PF scheduling provides multiuser diversity with fair resource allocation, it still schedules more time slots to the users in favorable channel conditions than to other users with poor channel conditions.

Absolute fairness implies equal transmission opportunities or equal allocation of resources to users, so it will be more desirable to allocate an equal number of time slots to each user while maximizing the total sum-rate throughput. To achieve such fairness, Liu et al. (2008) proposed the optimal temporal fairness scheduling algorithm (OTFSA) which determines to which user the current time slot should be assigned to achieve optimal temporal fairness. Although, the temporal fairness algorithm guarantees pre-assigned time fraction to each user, weak users may not enjoy the multiuser gains as the scheduler prefers strong users. Hence, Shen et al (2006) achieved fairness by maximizing the total capacity subject to user rate proportionality constraints, instead of maximizing the minimum user's capacity.

In Han et al. (2005) a new fairness criterion was proposed, which is a generalized proportional fairness based on Nash bargaining solutions and coalitions for maximizing the overall system rate, under each user's maximal power and minimal rate constraints, while considering the fairness among users. Among the various fairness criteria, the PF scheduling is widely considered as a good solution because it provides an attractive tradeoff between the maximum average throughput and the user fairness. This tradeoff is provided by exploiting the temporal diversity and game-theoretic equilibrium in a multiuser environment (Kelly et al., 1998) and seems more attractive than the max-min fairness in wireless networks. However, it has been shown that the PF scheme, which is implemented for 1xEVDO systems, is fair only in ideal cases when users experience similar channel conditions. In Eryilmaz and Srikant (2007), for instance, a congestion-control mechanism is proposed with such policy employed to introduce user fairness through traffic policing if the arrival rates are elastic, i.e., the traffic sources can adapt their rates (Dañobeitia & Femenias, 2011). But Andrews (2004) shows that proportionally-fair scheduling can lead to unstable queues, even for low arrival rates since it does not utilize the buffer state.

Parag et al. (2005) observed that the capacity achieved by the maximal-sum-rate and proportional fairness scheduling algorithms that are solely based on channel conditions does not necessarily translate into throughput when the input traffic is bursty. Because these algorithms work with two implicit assumptions that there is no constraint on the maximum size of the transmission buffer (or queuing delay) and the scheduled users always have data to transmit. This is only true for the application that is delay-tolerant (such as NRT and BE) and the buffer size is relatively large. However, real networks implement finite buffers and real-time applications require delay constraints. Therefore, throughput-oriented and fairness-oriented scheduling algorithms cannot support delay sensitive services such as voice and RT video, because they cannot guarantee the QoS of RT services (Lei et al. 2007) as widely required in BWASs.

2.10.3 QoS-Oriented Scheduling Algorithms

Fairness is an important consideration when designing a resource allocation algorithm for a wireless network; however, when diverse traffic flows are present in a network and each requires a different QoS, fairness is no longer sufficient as it cannot account for how a user is satisfied in terms of his QoS requirement; a user call satisfaction measure does. Therefore, reasonable objective to be pursued in wireless networks is the maximization of the number of satisfied users (Dhrona et al., 2009; Duval et al., 2010). Many scheduling algorithms have been proposed in the literature for broadband wireless access (BWA) systems and their performances have been well-studied for individual traffic types and QoS classes.

A unified scheduling framework should be able to accommodate all three classes of service such as real-time (RT), non-real-time (NRT) and BE (Liu et al., 2008). However, when common resources are shared by RT and NRT traffics which deal with different QoS constraints, respectively, it is difficult to compare the urgency of each class and to decide which traffic to serve with how much resource; the performance metric is different for the two classes (Lee et al., 2009). Because of the inherent difficulty of providing different scheduling policies for different traffic flows, Wang and Jia (2010) implemented all the QoS services for WiMAX standards using only a strict traffic priority.

Ganz and Wongthavarawat (2003) proposed Uplink Packet Scheduling (UPS) to support all types of service flows and applied a combination of a strict priority service discipline, Earliest Deadline First (EDF), and a Weight Fair Queuing (WFQ) one. However, user is scheduled based on both static bandwidth request and strict priority. Traffic prioritization raises the issue of allocation fairness as it already determines the order of access. Furthermore, QoS provided by traffic prioritization is only relative; a higher priority does not even ensure any absolute performance but only provides relatively better performance (Kelly et al., 1998).

Shin and Byung-Han (2004) proposed a Packet Loss Fair Scheduling (PLFS) algorithm, in which the packet loss of each user from different real-time traffic streams is fairly distributed according to the tolerable packet loss requirements of all the users. However, it is better to estimate the performance of RT traffic in terms of packet delay. Wang et al. (2008) presented a Joint Urgency and Efficiency Scheduling (JUES) algorithm which takes into account spatial characteristics of the multi-cell OFDMA-based system to reduce user's packet drop ratio. The NRT has minimum data rate constraint and BE is satisfied as long as the payload content is preserved. Liu et al. (2007) proposed a joint spatial-frequency scheduler based on Modified Proportional Fairness (MPF) and Quality Guaranteed (QG) priority function for mixed RT and NRT services for MIMO OFDMA system. However, this scheduler is only effective for traffic types which have no explicit constraints on the minimum achievable average data rate and/or the maximum allowable absolute delay.

Ukil (2009) formulated a cross-layer framework for WiMAX networks, based on channel-Queue- and QoS awareness, to optimize the system performance as well as maintaining the end-to-end QoS of individual users. Their algorithm shows better performance for non-real-time traffic in QoS diversified heterogeneous traffic condition. The algorithms are also said to be optimal in terms of providing maximal possible throughput. However, throughput-optimal policies are not fairness oriented in principle (Salem et al., 2010), as they aim to stabilize all user queues under any heterogeneous traffic flows within the system's capacity region. Lei et al. (2007) designed an algorithm which aims at maximizing system throughput while satisfying QoS requirements of the RT and BE services. However, for QoS requirements, the algorithm considers arrival rate as the required average transmission rate for RT and BE and average absolute deviation of transmission rate for RT. No specific delay requirement was provided for RT.

Andrews et al. (2000) proposed the modified largest weighted delay first (M-LWDF) algorithm, based on utility function of waiting time in the queue, instantaneous potential transmission rate and maximum tolerable delay for resource allocation. Shakkottai and Stolyar (2001) designed the exponential (EXP) scheduler also based on exponential utility function of waiting time in the queues, the instantaneous potential transmission rates and the maximum tolerable delay requirements. The EXP uses exponential-based marginal utility function in its scheduling rule aimed at minimizing the total delay in the system by equalizing the delay for all the users.

Maximum Delay Utility (MDU) (Song & Li, 2005a; Song, 2005) scheduling policy, based on joint channel-and-queue-aware scheduling approach, which maximizes the total utility with respect to average delays or average waiting times in the queues, was proposed. The MDU can provide better throughput and delay Jain's fairness indexes for RT and NRT traffic classes, but such a fair behavior is obtained at the cost of QoS satisfaction (Femenias et al., 2012). The MDU scheduling was proved to have better throughput-delay performance than the other scheduling schemes mentioned above. It was also found capable of reducing queuing delay substantially with a moderate or heavy traffic load. The algorithm meets the QoS requirements and outperforms the modified largest weighted delay first (M-LWDF) algorithm in terms of throughput for BE users and packet drop rate for RT users. However, for QoS requirements, the algorithm considers arrival rate as the required average transmission rate for RT and BE and average absolute deviation of transmission rate for RT. No specific delay requirement was provided for RT and besides, infinite buffer was considered for BE users only.

Ali and Zeeshan (2012) developed a U-delay scheduling rule comprising of a utility function based on Sigmoid utility to allocate bandwidth among various service classes and then another utility function based on deadline to select users within each class. The scheduling decision takes two scheduling epochs, hence increases scheduling time. Lin et al. (2008) proposed a Greedy-Latency scheduler using packet latency ratio and packet dropping policy. The algorithm is designed for only variable bit rate (VBR)-type real-time applications with maximum delay requirements.

Mohanram and Bhashyam (2007) designed a sub-optimal joint subcarrier and power allocation algorithm (JSPA) for Channel-Aware Queue-Aware (CAQA-JSPA) scheduling on a multiuser OFDM downlink by optimizing a user's power allocation immediately after each subcarrier is allocated to the user. The CAQA-JSPA was found to achieve a greater delay performance over CAQA scheduling with fixed power allocation (CAQA-FPA) and M-LWDF. This scheduler, however, is only applicable for traffic types without any constraint on delays (Kim & Han, 2005). Ryu et al. (2005) proposed an *urgency* and *efficiency* based packet scheduling (UEPS) algorithm, which is able to schedule RT and NRT traffic, at the same time. Its design goal is to maximize throughput of NRT traffics while satisfying QoS requirements of RT traffics. The objective of the UEPS algorithm is to maximize throughput for NRT traffic, while satisfying QoS requirements for RT traffic, such as packet delay and packets loss ratio.

In order to satisfy QoS demands, these algorithms attempt to provide optimal balance of utilization and fairness using delay (deadline) as the scheduling weight. This approach has the advantage of low complexity of implementation, as it only maps only the delay requirements for different traffics into the scheduling weight. All the foregoing scheduling policies are based on a single type of QoS such as traffic priority or packet delay to schedule heterogeneous traffics. Therefore, even if they are able to provide some QoS differentiation, they cannot guarantee that each traffic flow is satisfied in terms of his QoS requirement. Besides, any performances obtained are only relative since the QoS metrics upon which they are allocated common resources are unrelated to the specific requirement of each application; although the algorithms can provide some trade-off between system capacity and fairness. However, user QoS satisfaction is a more appropriate performance metric in when diverse traffic flows are present in a network and each requires a different QoS.

In heterogeneous traffic scenario, multiple utility functions are required, in which each different utility function is dedicated to different class of services, to guarantee that each user's QoS requirement is satisfied. However, Wang and Jia (2010), using strict priority implemented all the QoS services for WiMAX standards in the proposed algorithm. The algorithm uses strict priority disciple which allows the higher priority connections to starve the lower priority connection of bandwidth. However, a higher priority does not even ensure any absolute performance but only provides relatively better performance (Lee e al., 2009). Wang et al. (2007) considers support for three major classes of service, i.e. BE traffic with no QoS requirement, NRT traffic with data rate requirement, and real-time traffic with delay requirements. However, the algorithms do not aim at scheduling the different traffics at the same time for common resources.

In Balakrishna and Canbark (2014), a traffic-aware QoS provisioning scheduling algorithm was proposed for constant-bit rate, video streaming and BE. However, the utility functions are based on average waiting time and traffic priority; no application-specific QoS requirements are included. Hence, efficient QoS provisioning will be difficult to achieve. However, Lee et al. (2009) noted that traffic prioritization raises the issue of fairness as it already determines the order of access. The utility function designed to allocate resources to heterogeneous traffics must consider the QoS parameters appropriate for each traffic flow. Utility function can measure the amount of utility (satisfaction) a user derives in terms of resources allocated to him. Therefore,

Rodrigues and Casadevall (2009) proposed adaptive delay-based fairness (ADF) and adaptive throughput-based fairness (ATF) to schedule RT and NRT traffics, respectively. Although, the utility functions contain the appropriate QoS metrics, they did not specify any QoS requirements hence scheduling priority cannot be dynamically adjusted when the QoS metric exceeds the minimum or maximum requirement. Al-Manthari et al. (2009) proposed a fair class-based (FCB) algorithm that incorporates three utility functions, one for delay-based traffic, one for minimum rate-based traffic and the other for maximum-rate-based traffic was proposed. The algorithm specifies three different types of constant parameters for each type of traffic for the purpose of differentiating their QoS requirements. The main problem with this approach is the complexity of the algorithm and the difficulty of choosing an optimal value for each set of the constant parameters.

Lima et al. (2014) developed two utility-based radio resource allocation (RRA) policies, the TSM policy and the DSM policy based on sigmoid utility function and both aimed at maximizing the number of satisfied users in the system. However, the combined DSM/TSM scheduling policies are unable to guarantee efficient QoS differentiation when heterogeneous mixed traffics have to be scheduled at the same. Besides, the DSM/TSM allocates most of the networks resources to higher priority traffics, thus starving lower priority traffics. As a result, the algorithm cannot guarantee both fairness and efficient user satisfaction performances. The reason for this is that DSM/TSM uses a similar bell-shaped utility functions for both RT and NRT traffic, which makes it difficult to achieve a fair distribution of network resources on the basis of the differing QoS parameters.

In resource allocation problems, it is widely accepted that the higher the data arrival rate of a traffic flow, invariably the higher is its average throughput. But in a mixture of diverse traffics, the amount of sharable resources that the users get depends not only on their data arrival rates but also on their QoS constraints. However, when arrival rates are same or slightly different the traffic with higher priority QoS requirement must be satisfied more. Supporting QoS implies an adaptation to various applications (Lee et al., 2009), and to achieve efficient QoS differentiation an optimal scheduler needs to consider implementing different utility functions for different traffic classes.

Therefore, to provide satisfactory performance for the system, a utility-based resource allocation framework which takes into account the specific QoS requirement of each application is required at the media access control (MAC) layer of the base station (BS). Musabe and Larijani (2014) developed a cross-layer scheduling scheme which improves only one real-time application namely VOIP in terms of throughput and delay but it causes a huge starvation for best effort Traffic (FTP). Mushtaq et al. (2014) proposed a novel scheduling algorithm which improved the performance for VOIP traffic and compared it with well-known algorithms. The proposed scheme shows better performance in terms of delay sensitivity and Packet Loss Ratio (PLR) whereas it didn’t study the non-real time applications.

Vulpe et al. (2014) proposed an approach which prioritizes the QoS users but it doesn’t concern about QoS users with high delay sensitivity which significantly increases the Packet Loss Ratio (PLR). The Frame Level Scheduler (FLS) was proposed to give attention to the users with tightest delay (Piro et al., 2011). FLS scheme is a two level scheduler which shows an acceptable performance for real time applications which are sensitive to delay but highly starves the non-real time users. A Scheduling algorithm for QoS and non-QoS (NQoS) users (Hajjawi & Ismail, 2015) was proposed. The model is separated into two levels where the resources are initially distributed among QoS and NQoS users in the first level and in the second level the users with tightest delay requirements are prioritized. However, two-level increases the scheduling time.

2.11 Resource Allocation in Relaying Systems

Theoretical studies have shown that MIMO antenna system can provide spatial diversity in the system (Tse & Viswanath, 2005). This technique is attractive for its significant improvement to information rate and transmission reliability (Goldsmith et al., 2003; Zhang et al., 2007) However, to achieve acceptable and meaningful diversity the multiple antennas at each terminal are expected to be reasonably separated from one another by at least a quarter of a wavelength of the operating frequency. Therefore, the high cost and the complexity of implementation arising from this is a major challenge to MIMO systems.

However, as the radio spectrum increasingly becomes competitive new strategies must be developed to increase the spectral efficiency of wireless networks, especially where MIMO antennas are not deployable at the terminal. One of such fast and economic methods is to employ cooperative relay terminals, where each relay re-transmits the signal received from a remote source to the destination (Boyer et al., 2004). This technique has been found to be an effective way to combat wireless fading by providing spatial diversity without the need of multi-antenna configurations (Kramer et al., 2005; Laneman et al., 2003; Laneman et al., 2004).

Cooperative relaying for wireless networks has received considerable interest, as it provides coverage extension and reduced power consumption (Lin et al., 2010; Peter & Heath, 2009) without incurring the high costs of additional BS deployment. It is also considered a promising solution to improve spatial diversity gains through the cooperation between the source and the relay nodes where multiple antennas are not deployable at the terminal. The works (Lin et al., 2010; Peter & Heath, 2009) showed that the cooperative system achieves higher throughput compared to the non-cooperative one. Because of its potential application, cooperative relaying has been incorporated into many wireless standards, such as 3GPP long-term evolution (LTE) (Dahlman et al., 2013), IEEE 802.16j (wireless multi-hop relay) (Peter & Heath, 2009) and IEEE 802.16m (WiMAX2) (IEEE, 2011).

Cooperative communications have emerged as promising techniques to improve spatial diversity gains through the cooperation between the source and the relay nodes. Cooperative diversity is a set of techniques that exploit the potential of spatial dispersed terminal/node antennas to improve communication reliability. Various cooperative diversities where each user or node is equipped with single antenna were proposed and analyzed in the works (Kramer et al., 2005; Laneman et al., 2003) to mimic the performance advantages of multi-antenna systems.

Relaying strategies such as DF and AF have been investigated for cooperative communications (Janani et al., 2004; Kramer et al., 2005; Laneman et al., 2004; Lin et al., 2010; Nabar et al., 2004; Peter & Heath, 2009). In DF protocol, also known as digital relay or regenerative relay (Hasna & Alouini, 2002) the relay node decodes the received message and then forwards the decoded message to the destination node. In AF relaying protocols (Patel et al., 2006), also called analog relay or non-regenerative relay, the relay simply scales the received signal by a certain gain factor, G, and then retransmits (forwards) it to the destination node. Thus AF relays have relatively less complex circuitry and cost. However, as simple repeaters, AF relays are well known for transmitting an amplified version of noise signals over the communication channel since it amplifies the noise parts as well as the signal during the scaling process.

In relaying, duplexing can be achieved with FDMA and TDMA schemes where the available bandwidth and time frame, respectively, are shared. Due to the path loss and fading effects, the channel condition and capacity of these orthogonal channels are considerably different. TDMA-based relaying protocol is the most common and simplest form of protocol among the relaying protocols in the literature. This protocol is the most-employed protocol in practical systems such as IEEE 802.16j standard (Soldani & Dixit, 2008).

The relay operation can be classified into the full-duplex (FD) and the half-duplex (HD) relaying modes (Cheng et al., 2012). In FD mode, the relay node can transmit and receive simultaneously at the same frequency band in one frame duration, but requires self-interference cancellation due to the signal leakage between the relay output and input. In HD mode, the relay is restricted to transmit and receive over alternate sub-frames (requiring twice the time slot to transmit a symbol) thereby resulting in a loss of spectral efficiency but without incurring self-interference penalty (Patel et al., 2006). HD mode enjoys much lower implementation complexity than the FD mode but wastes resources. However, the full duplex is a more challenging and attractive mode which can provide better performance than the HD mode if the self-interference can be reduced efficiently (Patel et al., 2006). Recent research has shown that FD relaying is feasible by using interference cancellation techniques and transmit/receive antenna isolation (Fan et al., 2008; Riihonen et al., 2009).

The capacity enhancement of relay networks employing MIMO technologies has recently been investigated (Li & Jafarkhani, 2009; Nabar et al., 2003). For multi-antenna relay networks, opportunistic relay algorithm will lead to more consumed energy to obtain multiple CSIs and complicated relay selecting criteria to evaluate the whole state of a multi-antenna terminal. Bletsas et al. (2006) proposed opportunistic relaying to select the relay with best instantaneous channel condition, based on an 802.11b-like MAC protocol.

Selective relaying is a technique through which the source adaptively chooses to relay on certain subcarriers depending on the potential gains (Duval et al., 2010). Various studies on the relay selection have been carried out to increase the diversity order, reduce adverse channel effects, and to overcome the half-duplex loss of the use of relay (Bletsas et al., 2006; Duval et al., 2010; Emmanouil et al., 2009; Fareed & Uysal, 2009; Hasan et al., 2011; Kim & Kim, 2009). Bletsas et al. (2006) show that the diversity gains can be achieved in the order of the number of relays assisting the communication. This also compensates for the half-duplex loss. Fareed and Uysal (2009) proposed a relay selection scheme to maximize the SNR at the destination.

Depending on the network configuration and available resources, three main relaying schemes have been proposed in cooperative networks: amplify-and-forward (AF), decode-and-forward (DF), compress and forward (CF) (Laneman et al., 2004; Simoens et al., 2007). However, amplify-and-forward and decode-and-forward are the most widely used protocols in practice. Therefore, most of these relay selection works are based on relaying protocols such as AF and DF. Previous works (Abrar et al., 2012; Ng et al., 2012; Rasouli, 2012; Saleh et al., 2009) have all compared AF and DF. Their results, respectively, show that spectral efficiency for DF usually outperforms that of AF. This is due to the fact that the existing AF relays amplify the thermal noise power in the case of HD and FD relaying and loop interference in the case of FD relaying (Ng et al., 2012). Both the AF and DF systems were implemented (Emmanouil et al., 2009), and compared in terms of implementation loss and the complexity. They showed that the AF protocol is less complex and has lower implementation loss; the performance is very similar to the theoretical studies. However, the main disadvantage of AF relaying is noise amplification.

In the literature, different types of AF relaying protocols have been studied. These include the fixed-gain (FG) relay (Hasna & Alouini, 2004), variable-gain (VG) relay (Emamian et al., 2002) and unlimited-gain (UG) relay (Hasna & Alouini, 2002). The FG gain factor only requires the knowledge of the average power received by the relay; therefore, it is called a fixed-gain relay. The VG gain factors require the relay to have knowledge of instantaneous state information (CSI) of the source-to-relay (S-R) channel and are, therefore, called variable gain relays; because they continuously adjust their gain depending on the instantaneous channel response (Anghel & Kaveh, 2004). The performance of VG relay systems is also compared to that of FG relay systems (Hasna & Alouini, 2004). The outage performance of a two-hop system with semi-blind FG relay is compared to that of a system with VG relay, the exact performance of which was given (Emamian et al., 2002). Emamian et al. (2002) also derived the outage probability expression for a two-hop communication through one VG relay. When CSI is not available at the source, the second-order statistics (variance) of the backward channel has been used in (Cheng et al., 2012) to perform blind/semi-blind relaying for FG and VG protocols. In blind FG relaying, the scaling to the received signal is performed using a fixed constant, thus the relay does not need to measure the instantaneous CSI, whereas, in semi-blind it is assumed that the AF relay has some average knowledge of the S-R channel.

The performance of the systems employing VG relays over a Rayleigh fading channel was first studied in (Laneman & Wornell, 2000). Hasna and Alouini (2002) present the UG protocol as a benchmark protocol that gives tight bounds on the performances of the VG protocol for Rayleigh and Nakagami channels, respectively. In both works, the performance of a two-hop communication system with UG relay is compared to that of a regenerative system. UG relay is a hypothetical relay able to invert the channel regardless of its amplitude; this relay protocol is not applicable in practice because when the fading coefficient is too small, the relay gain is supposed to become infinitely large to compensate the fading effect.

Lee (2009) noted that applying the upper bound on relay gain of UG, when the transmits power at the relay forced to unity, can drive the relay gain to infinity as the S-R channel coefficient goes a little towards zero. The UG relay gain only sets the upper bound for the performance of AF, it is not a practical relay and more its performance is still poor compared to DF. Therefore, they proposed a clipped-gain (CG) relay with a gain normalizing factor whose value is set by comparing the SNR at the S-R channel with the SNR threshold at the receiver. When the measured SNR is less than the SNR threshold, the signal is clipped; whereas, as the channel coefficient tends to infinity, the signal is not transferred. However, the CG protocol yields a slightly better performance in terms of probability of outage than other practical AF protocols including the benchmark protocol UG only at low SNRs. However, in resource allocation, the CG protocol would offer almost similar performances as VG and UG.

Relaying systems should manage their spectral resources in order to maximize their performance metrics such as throughput and fairness. Therefore, Ng et al. (2012) proposed a distributed resource allocation algorithm, which enables the exploitation of the benefits of different relaying protocols and duplexing schemes, fulfills heterogeneous QoS requirements. However, the algorithm only enables dynamic selection between AF relaying and DF relaying with FD and HD relays. Previous work (Riihonen et al., 2009) proposed resource allocation in FD mode, but they did not consider the delay constraints. Such an information-theoretic approach maximizes the system throughput but cannot satisfy users' requirements in terms of delay constraints that may be requested for both real and non-real time applications.

Unlike traditional wireless repeaters, advanced wireless relays should be both channel- and queue-aware, supporting a wide range of delay constraints, and not only optimizing at the physical layer but also providing diverse QoS guarantees at upper-protocol layer (Cheng et al., 2012). Bi and Zang (2010) proposed relay-assisted resource allocation which achieves same diversity gain as opportunistic scheduling in TDMA without incurring fairness penalty. However, the algorithm did not consider a system with heterogeneous QoS requirements. Furthermore, no solution has been proposed so far to achieve a specific level of fairness when absolute fairness is not required in the system. Rasouli (2012) proposed a scheme in which cooperating users are divided into groups and different cooperation (scaling) coefficients are statically assigned to each group for resource allocation. The scheme maintains the same level of fairness but obtains higher data rates compared with a similar algorithm which does not consider cooperation.

Recently, few distributed relay-assignment algorithms are proposed for single antenna relay networks. The scheduling schemes exploit the transmission opportunities at the various nodes to effectively utilize the capacity of the network (Kim & Sichitiu, 2010). However, duplicating scheduling algorithms at relay nodes can lead to complexity because these functions must also be centrally coordinated. In the emerging OFDMA-based standards such as 3GPP LTE and IEEE 802.16j, the multi-hop relay concept has been introduced to satisfy the QoS of users near the cell edge (Wang et al., 2010).

Relaying techniques are integrated with OFDM-based wireless systems to improve the system performance by taking advantage of both techniques. But to fully exploit the benefits of relaying in an OFDM system, efficient management of resources such as subcarriers and power is required. Rasouli (2012) also noted that relay station parameters play a major role in the resource allocation to the active users in the OFDM relaying system. In (Berger & Wittneben, 2005), coherent gain allocation schemes that achieve a distributed spatial multiplexing gain are discussed. The authors noted that one approach to allow multiple users to access the channel simultaneously is to compute the relay gain factors such that the source/destination streams are completely orthogonalized in space (multiuser zero-forcing (MUZF) relaying). However, most existing scheduling algorithms do not take advantage of this.

2.12 Chapter Summary

The main characteristics of wireless channel were reviewed in this chapter. Medium access techniques were also reviewed and their main advantages and disadvantages were mentioned. OFDM was shown as an appropriate physical layer technique to combat frequency selectivity of the wireless channel which provides an appropriate framework for dynamic resource allocation in the OFDMA system. Also in this chapter, a review of important scheduling and subcarrier allocation algorithms proposed in the literature for OFDM systems were carried out. The subcarrier allocation algorithms were classified as throughput-oriented, fairness-oriented and QoS-oriented. While throughput-oriented subcarrier allocation maximizes the throughput, fairness-oriented subcarrier allocation prioritizes the fairness and QoS-oriented subcarrier allocation assures QoS guarantee in the system. Relaying in OFDMA systems offers opportunities to extend coverage and also increase data rate as an economic alternative to installing MIMO antennas. However, attentions now mostly focus on maximizing the system efficiency without consideration for resource allocation fairness or QoS. Therefore, a review of different resource allocation techniques in relaying systems presented in literature were carried out. All these lead us to Chapter 3, where the methodology to be used in developing the algorithms proposed in this thesis is described.

Chapter 3

Research Methodology

The main requirement for the next generation of wireless network is that it should cost-effectively provide guaranteed quality-of-service (QoS), especially in terms of delay and average throughput requirement, with ubiquitous high data rate coverage, when and where required (Yanikomeroglu & Zhang, 2008). Therefore, utility functions designed for scheduling and resource allocation algorithms should necessary incorporate either one or the two QoS parameters.

The average throughput, $\bar{R}_m[t]$, of user m can be obtained by using a simple exponential smoothing filter as follows:

$$\bar{R}_m[t+1] = (1-\beta)\bar{R}_m[t] + \beta R_m[t] \tag{3.1}$$

where $0 < \beta < 1$ is the forgetting factor, $\bar{R}_m[t]$ and $\bar{R}_m[t+1]$ are the average throughput or average achievable data rate during the current time slot and at the beginning of the next time slot, respectively. $\bar{R}_m[t]$ is a not good measure of the actual amount of resources allocated, so $\bar{R}_m[t]$ for which the instantaneous data rate $R_m[t]$ for user m is summed over all subcarriers $k \in (1, K)$ and time slots $t \in (1, T)$ is better (Wang et al., 2007). The recursive HOL packet delay is approximately computed by:

$$D_m^{hol}[t+1] = D_m^{hol}[t] + \frac{\bar{\omega}_m T_s - R_m[t] T_s}{\bar{\omega}_m} \tag{3.2}$$

where T_s is time slot duration and $\bar{\omega}_m$ is the mean bit arrival rate for user m. The term $\frac{\bar{\omega}_m T_s - R_m[t] T_s}{\bar{\omega}_m}$ represents the instantaneous HOL packet delay. When $R_m[t]$ is zero, the HOL packet delay is incremented by T_s. When the arrival rate for user m equals his achievable data rate, i.e., $\bar{\omega}_m = R_m[t]$, the instantaneous packet delay is zero.

3.1 Related Utility-Based Scheduling and Resource Allocation Rules

In cross-layer resource management architecture, utility function offers a tangible metric for pricing the benefit of taking use of certain radio resource to guarantee the practical transmission services (Li et al., 2012). Therefore, the scheduling rules below have been formulated as utility-based ones to support traffics in wireless networks. Each of the rules is designed with different maximization objectives; some aim at maximizing the total throughput while some others focus on providing resource allocation fairness and/or QoS satisfaction.

3.1.1 Maximum Sum Rate (MSR)

The utility function for MSR (Knopp & Humblet, 1995) is $U_m(R_m[k,t]) = R_m[k,t]$ and the gradient-based utility function $U'_m(R_m[k,t]) = w_m[k,t]$, where the utility-based weight factor $w_m[k,t] = 1$ indicating that the utility is independent for different subcarriers and thus maximizes users with the highest transmission data rate (Song & Li, 2003b). In MSR, the user with index m^* is chosen if the following condition is satisfied:

$$m^* = \arg\max_{\Re_m}\{1.R_m[k,t]\} \tag{3.3}$$

3.1.2 Proportional Fairness (PF)

The PF (Kelly et al., 1998) scheduling provides each connection a priority inversely proportional to its data rate. The PF scheduler is based on channel-aware scheduling rule aiming at maximizing the logarithmic-sum-throughput, $\sum_m \ln(\bar{R}_m[t])$. However, the gradient-based utility function, which is equal to the scheduling weight, is given as $U'_m(t) = w_m[k,t]$, where $w_m[k,t] = \frac{1}{\bar{R}_m[t]}$.

Therefore, the user with index m^* is chosen if the following condition is satisfied:

$$m^* = \arg\max_{\Re_m} \left\{\frac{R_m[k,t]}{\bar{R}_m[t]}\right\} \tag{3.4}$$

3.1.3 Urgency-and Efficiency-Based Packet Scheduling (UEPS)

The utility function used by the UEPS (Ryu et al., 2005) algorithm for RT traffics is a relaxed z-shaped sigmoid function represented by:

$$U_m(D_m^{hol}[t]) = 1 - \frac{1}{[1+exp\{-\mu*(D_m^{hol}[t]-D_m^{max})\}]} \tag{3.5}$$

$$= \frac{exp\{-\mu*(D_m^{hol}[t]-D_m^{max})\}}{[1+exp\{-\mu*(D_m^{hol}[t]-D_m^{max})\}]}$$

Where $\frac{1}{[1+exp\{-\mu*(D_m^{hol}[t]-D_m^{max})\}]}$ corresponds to original s-shaped sigmoidal function for $D_m^{hol}[t]$, $D_m^{hol}[t]$ is the current HOL packet delay for user m in time slot t at subcarrier k and D_m^{max} is the maximum tolerable HOL packet delay. A decreasing sigmoid is achieved when the slope-shaping parameter μ is set to -1. The marginal (derivative of) utility function is obtained as:

$$U'_m(D_m^{hol}[k,t]) = \left|\frac{-\mu*\, exp\{-\mu*(D_m^{hol}[t]-D_m^{max})\}}{[1+exp\{-\mu*(D_m^{hol}[t]-D_m^{max})\}]^2}\right| \tag{3.6}$$

where μ still set to -1, $U'_m(D_m^{hol}[k,t])$ is a bell-shaped utility function.

For NRT traffics the utility function is represented by:

$$U_m(D_m^{hol}[t]) = 1 - \frac{\exp(\mu * D_m^{max})}{\exp(D_m^{hol}\,[t])^2} \tag{3.7}$$

Differentiating (3.7) with respect to $D_m^{hol}\,[t]$ yields the marginal utility function represented by:

$$U'_m(D_m^{hol}[t]) = \frac{\mu * \exp(\mu * D_m^{max})}{\exp(D_m^{hol}\,[t])} \tag{3.8}$$

The constant μ is set to 1 to produce a positively increasing utility function. At each scheduling time, the user with index m^* is chosen based on the following condition:

$$m^* = \arg\max_{\Re_m} \left\{ |U'_m(t)| \frac{R_m[k,t]}{\bar{R}_m[t]} \right\} \tag{3.9}$$

3.1.4 Delay-Based Satisfaction Maximization (DSM)/Throughput-Based Satisfaction Maximization (TSM)

This work extended the idea in (3.5) by including the parameter δ in the utility function.

The bell-shaped Delay-based Satisfaction Maximization (DSM) (Lima et al., 2014) can be expressed by:

$$U'_m(D_m^{hol}[t]) = \frac{\delta * \mu * \exp\{\mu * \delta * (D_m^{hol}[t] - D_m^{max})\}}{[1 + \exp\{\mu * \delta * (D_m^{hol}[t] - D_m^{max})\}]^2} \tag{3.10}$$

where μ and δ are the constants that determine the direction and the shape of the sigmoid function, respectively. The parameter μ is set to 1 and the shape-controlling parameter δ is defined as $\delta = \frac{log\left(\frac{1-0.01}{0.01}\right)}{0.5 * D_m^{max}}$. To provide separately for a rate-based resource allocation, the utility function in (3.10) is modified, by replacing delay with data rate parameters, to obtain a Throughput-based Satisfaction Maximization (TSM) (Lima et al., 2014), expressed by:

$$U'_m(\bar{R}_m[t]) = \frac{\mu * \delta * \exp\{\mu * \delta * (\bar{R}_m[t] - R_m^{min})\}}{[1 + \exp\{\mu * \delta * (\bar{R}_m[t] - R_m^{min})\}]^2} \tag{3.11}$$

In this case μ is set to -1 and $\delta = \frac{log\left(\frac{1-0.01}{0.01}\right)}{0.5*R_m^{min}}$. Therefore, the user with index m^* is chosen according to the following:

$$m^* = \arg\max_{\Re_m}\{U'_m(t)R_m[k,t]\} \tag{3.12}$$

The bell-shaped marginal utility functions described by (3.10) (DSM) and (3.11) (TSM), respectively, are as shown in Figure 3.1.

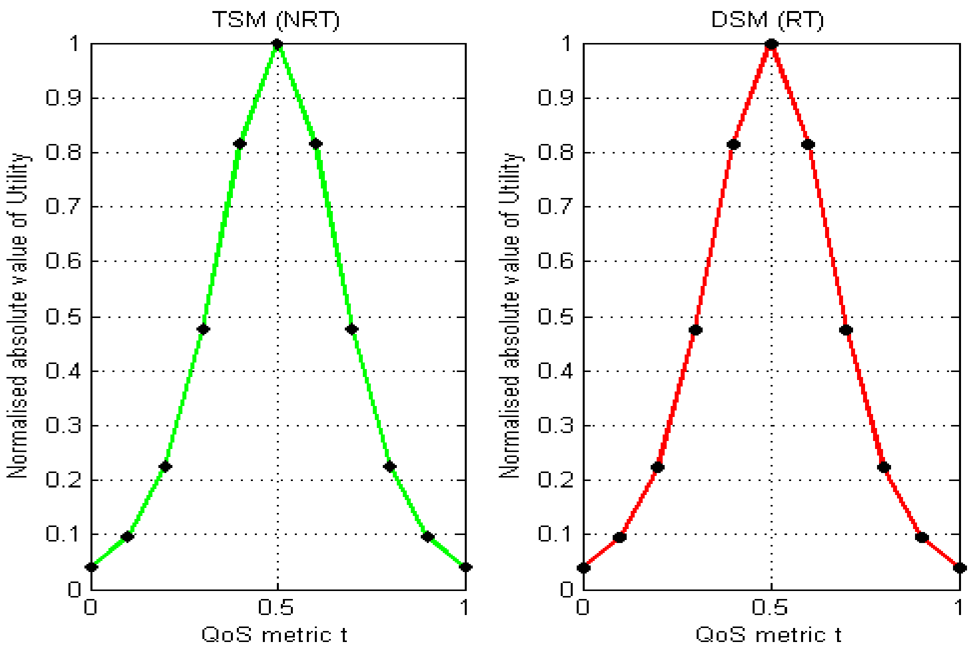


Figure 3.1: Utility functions for NRT (TSM) and RT (DSM) traffics

3.2 Proposed Utility-Based maximum QoS satisfaction (MQS) Rule

The MQS scheduling algorithm has the main objective to maximize the total utility of a wireless network in terms of user satisfaction. Therefore, in constructing a utility function, the nature and characteristics of each traffic is considered. The shape of the utility function should vary according to the adaptive characteristics of the application (Ermini, 2011). This implies that utility function must capture the specific nature of a heterogeneous traffic. The s-shaped sigmoidal utility function can be represented as (Ryu et al., 2005):

$$U(t) = \frac{1}{1+\ exp\{-\mu p(t-c)\}} \tag{3.13}$$

Where p and c are the parameters that determine the slope and location of inflection point, μ is a constant (-1 or 1) that determines if the sigmoid is a decreasing or increasing function. Suppose the scaled version of (3.13) can be written as:

$$U(t) = \frac{1}{1+\ exp\{-2\mu p(t-c)\}} \tag{3.14}$$

Multiply (3.14) throughout by $exp\{\mu p(t-c)\}$ to obtain:

$$U(t) = \frac{exp\{\mu p(t-c)\}}{exp\{\mu p(t-c)\}+\ exp\{-\mu p(t-c)\}} \tag{3.15}$$

The modified sigmoidal utility function in (3.15) has a property that the decreasing and increasing exponential functions in the denominator counter-balance each other; increase in one causes exponential decrease in the other and thus produces utility values that are closely distributed for the same values of μ and p. Therefore, utility function formulated in (3.15) provides proportional utility fairness. Figure 3.2 shows different utility curves for $p = \{1,2,3\}$ for NRT and RT traffics respectively. It can be seen that the higher the value of a the more nonlinear is the utility curve. As t increases the utility function remains convex, which implies that the user is less satisfied in terms of the utility derived so far. At the point of inflection$(t = c)$, the user has reached his maximum satisfaction. As t increases above c, the utility becomes concave and marginal utility is derived.

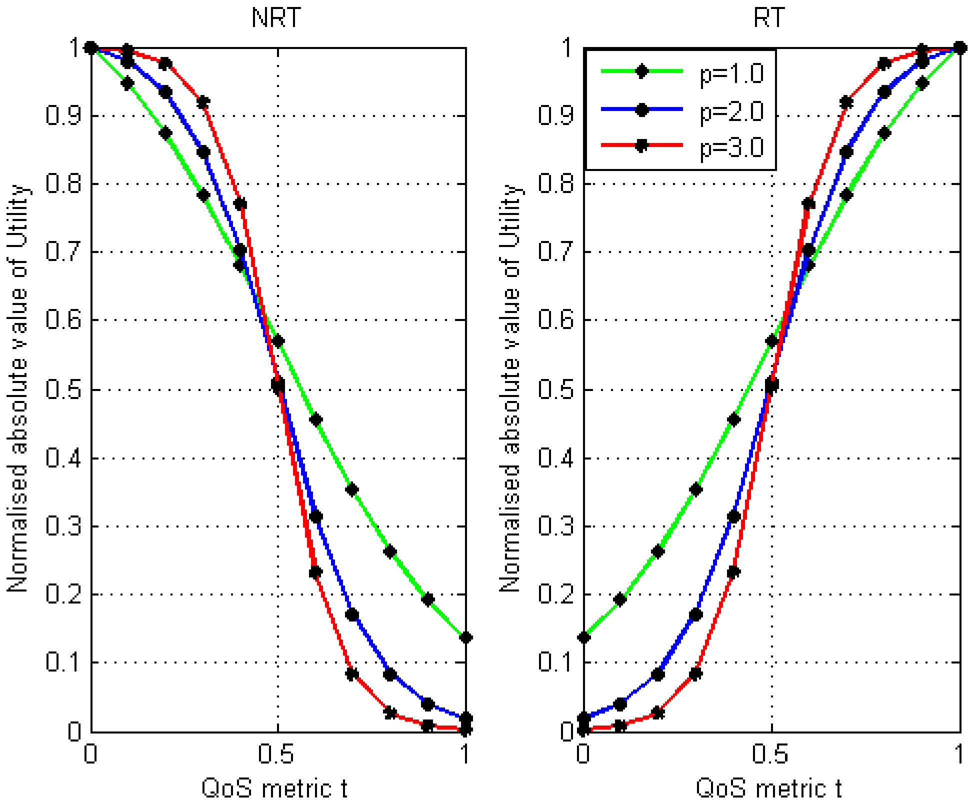


Figure 3.2: Utility functions for NRT and RT traffics for different values of scaling parameters

In practical application, t and c have to be substituted for q and q^{Req} representing the QoS metric (average throughput or average HOL packet delay) and QoS requirement (minimum required throughput or maximum tolerable delay), respectively.

Therefore (3.15) can be re-written as:

$$U(c) = \frac{exp\{\mu p(q-q^{Req})\}}{exp\{\mu p(q-q^{Req})\}+ exp\{- \mu p(q-q^{Req})\}} \tag{3.16}$$

The delay-based traffic is formulated as a sigmoidal-type utility function which is concave in terms of packet delay, D_m^{hol}, and which can be expressed by:

$$U(D_m^{hol}[t]) = \frac{exp\{p(D_m^{hol}[t]-D_m^{max})\}}{exp\{p(D_m^{hol}[t]-D_m^{max})\}+exp\{-p(D_m^{hol}[t]-D_m^{max})\}} \tag{3.17}$$

where D_m^{max} is the maximum delay requirement of user m and the parameter, $p = \frac{a}{D_m^{max}}$ is the normalizing parameter and $a \subset \{1,2,3\}$ is the constant that determines the shape of the utility curve. The RT users' utility derived from the network increases as the HOL packet delay, D_m^{hol}, increases; that is, the user's chances of being allocated resources increases as his HOL packet delay increases with respect to his maximum delay requirement, D_m^{max}.

However, in a heterogeneous mixed traffic involving NRT and RT services, the average throughput for NRT must be allowed to gradually decrease after it has achieved its required minimum rate so that the RT traffics will be able to satisfy their delay requirements. Therefore, a positive and decreasing utility function obeying the law of diminishing marginal utility will better capture this objective for NRT traffic, and can be mathematically modeled by:

$$U(\bar{R}_m[t]) = \frac{exp\{-p(\bar{R}_m[t] - R_m^{min})\}}{exp\{p(\bar{R}_m[t] - R_m^{min})\} + exp\{-p(\bar{R}_m[t] - R_m^{min})\}} \tag{3.18}$$

Similarly, the parameter $p = \frac{a}{R_m^{min}}$ is the normalizing parameter and $\bar{R}_m$ is the data rate of user m averaged over all time slots. BE traffic is generally considered an NRT service, but which requires no minimum rate guarantee. Therefore, the minimum rate requirement, R_m^{min}, in (3.18) can be substituted with zero to obtain a utility function for a BE traffic as:

$$U(\bar{R}_m[t]) = \frac{exp\{-p(\bar{R}_m[t])\}}{exp\{p(\bar{R}_m[t])\} + exp\{-p(\bar{R}_m[t])\}} \tag{3.19}$$

However, the normalizing parameter, $p = \frac{a}{L}$. The packet length, L (in bits), is used for BE traffic to prevent its data rate from increasing to infinity. The value of $a = 1$ is adopted in the three utility functions, because it provides a more linear curve. Therefore, the user with index m^* is chosen according to the following:

$$m^* = \arg\max_{\Re_m}\{U_m(t)R_m[k,t]\} \tag{3.20}$$

3.3 Amplify-and-Forward (AF) Relaying Technique

In this section, we establish a relaying model and then propose a relay station parameter called equalized scaling coefficient, to be used to control a utility function of some scheduling algorithms for the purpose of improving resource allocation. Performance metrics such as average system throughput, user call satisfaction and call fairness are analyzed and compared with some pre-existing scheduling algorithms.

3.3.1 System Model and Transmission of Dual-Hop Single Relaying

The symmetric connection system model as illustrated in Figure 3.3 in which the source (S) is assumed fixed but both the relay (R) and destination (D) can move is considered. It is assumed that the source can communicate directly with the destination as well as with it through the relay. In this model, an independent Rayleigh fading for the channel gains $h_{SR} \sim \mathcal{CN}(0, \sigma_{SR}^2)$,$h_{RD} \sim \mathcal{CN}(0, \sigma_{RD}^2)$ and $h_{SD} \sim \mathcal{CN}(0, \sigma_{SD}^2)$ is also assumed. The receiver noise are $z_{SR} \sim \mathcal{CN}(0, \sigma_n^2)$, $z_{RD} \sim \mathcal{CN}(0, \sigma_n^2)$ and $z_{SD} \sim \mathcal{CN}(0, \sigma_n^2)$. But the noise components$z_{SR} = z_{RD} = z_{SD} = z$, since all the node inputs experience identical additive white Gaussians noise (AWGN). Therefore, the noise power at each node is σ_n^2. The term $\mathcal{CN}(.)$ represents the complex Gaussian random variable with the first and the second parameter denoting the mean and the variance, respectively.

The source and relay transmits powers are denoted by P_S and P_R, respectively. Therefore, the total transmits power in the cell, $P_T = P_S + P_R$. The instantaneous SNR of S-D, S-R and R-D channel gains are defined, respectively, as:

$$Y_{SR} = \frac{P_S|h_{SR}|^2}{\sigma_n^2}, Y_{SD} = \frac{P_S|h_{SD}|^2}{\sigma_n^2}, Y_{RD} = \frac{P_R|h_{RD}|^2}{\sigma_n^2} \tag{3.21}$$

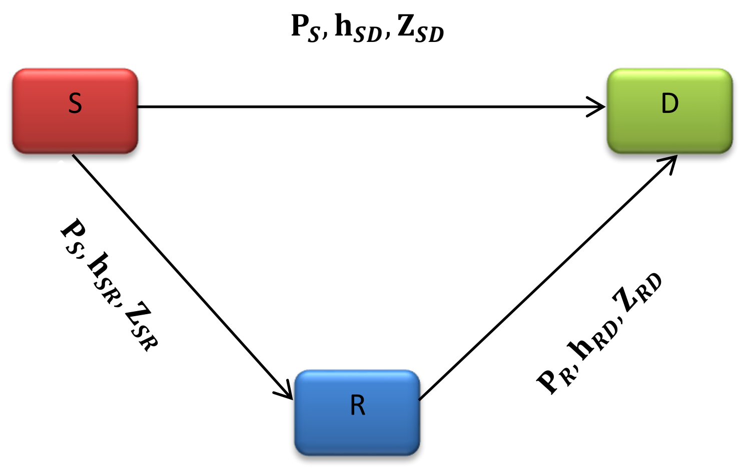


Figure 3.3: The relay transmission model

Table 3.1 describes the relay transmission in half-duplex (HD) mode. In this system, the relay node simply amplifies the received signal and then forwards it to the destination.

Table 3.1: TDMA transmission protocol.

	Source (S)	**Relay (R)**	**Destination (D)**
Timeslot 1	Transmits	Listens	Listens
Timeslot 2	-	Transmits	Listens

In half-duplex (HD) mode, the AF relay takes two time slots to transmit a packet from source to destination. In the first timeslot, the source broadcasts its unit-energy signal $x(t)$, and the signal received by destination is:

$$y_{SD} = \sqrt{P_S} h_{SD} x(t) + z \tag{3.22}$$

The signal received by relay is:

$$y_{SR} = \sqrt{P_S} h_{SR} x(t) + z \tag{3.23}$$

In the second timeslot, the relay multiplies the received signal by a relay gain, G_{HD}, and then forwards the amplified signal to destination. Therefore, at the destination, the received signal can be written as:

$$y_{HD} = G_{HD}\sqrt{P_s}\, h_{SR} x(t)\sqrt{P_R}\, h_{RD} + G_{HD}\sqrt{P_R}\, h_{RD} z + z \tag{3.24}$$

The SNR at the destination, through the relay link, can be expressed by:

$$\Upsilon_{HD} = \frac{P_S |h_{SR}|^2 P_R |h_{RD}|^2 G_{HD}^2}{P_R |h_{RD}|^2 G_{HD}^2 + 1} \tag{3.25}$$

where the variable-gain (VG) relay factor in HD mode can be expressed (Laneman & Wornell, 2003) by:

$$\mathrm{G}_{HD} = \sqrt{\frac{1}{P_S |h_{SR}|^2 + \sigma_n^2}} \tag{3.26}$$

Feasibility of relaying in FD mode has been investigated. Therefore, as in HD, the received signal, in the first time slot, at the relay is expressed (Ng et al., 2012) by:

$$y_{SR} = \sqrt{P_s}\, h_{SR} x(t) + \sqrt{P_R}\, h_{LI} q_{LI} + z \tag{3.27}$$

According to Ng et al. (2012), the relay subtracts the loop interference cancellation signal $C_{SR} = \sqrt{P_R}\, \hat{h}_{LI} q_{LI}$ from y_{SR} for loop interference cancellation to yield:

$$\hat{y}_{SR} = y_{SR} - C_{SR} \tag{3.28}$$

$$= \sqrt{P_s}\, h_{SR} x(t) + \sqrt{P_R}\, \Delta h_{LI} q_{LI} + z$$

where $\hat{h}_{LI}$ is the estimated loop interference channel and $\Delta h_{LI} \sim \mathcal{CN}(0, \sigma_e^2)$ is the residual loop interference channel due to imperfect channel estimation, σ_e^2 is the variance of the residual loop interference and q_{LI} is the accumulated loop interference signal at relay caused by FD relaying. In the second timeslot, the relay multiplies the signal received from source by

G_{FD} for FD and forwards it to the destination. Then the received signal at destination is given as:

$$y_{RD} = G_{FD}\sqrt{P_s}\, h_{SR}x(t)\sqrt{P_R}\, h_{RD} + G_{FD}\sqrt{P_R}\, h_{RD}z + G_{FD}\sqrt{P_R}\, h_{RD}\sqrt{P_R}\, \Delta h_{LI}q_{LI} + z \qquad (3.29)$$

The SNR at the destination, through the relay link, in FD can be expressed by:

$$\Upsilon_{FD} = \frac{P_S|h_{SR}|^2 P_R|h_{RD}|^2 G_{FD}^2}{P_R|h_{RD}|^2 G_{FD}^2 + 1} \qquad (3.30)$$

where G_{FD} is written as:

$$G_{FD} = \sqrt{\frac{1}{P_S|h_{SR}|^2 + P_R\Upsilon_{LI}\sigma_n^2 + \sigma_n^2}} \qquad (3.31)$$

and $\Upsilon_{LI} = \frac{\Delta h_{LI}}{\sigma_n^2}$ denotes the residual loop interference SNR.

3.4 Computing Equalized Scaling Coefficient

The SNR at destination can also be written as:

$$\Upsilon = \frac{P_S|h_{SR}|^2}{\sigma_n^2} . a \qquad (3.32)$$

Where the scaling coefficient, $a \in (0,1)$, can be derived as:

$$a = \frac{P_R|h_{RD}|^2 G^2}{P_R|h_{RD}|^2 G^2 + 1} \qquad (3.33)$$

where G can be substituted for the relay gain in (3.26) and (3.31), respectively, for HD and FD operating modes. However, the scaling coefficient in (3.33), which determines the proportion of the SNR on the S-R (backward) channel that can be transferred to destination, can be highly uneven across subcarriers because their values depend on the backward channel variation. Applying these directly to control resource allocation will be inefficient, because they will lead to more unfairness; and, therefore, they need to be equalized. This unevenness is caused by the relay gain, G. Therefore a generic gain, called the equalizing gain, is computed as:

$$\hat{G} = G + \frac{1}{P_R|h_{RD}|^2G} \tag{3.34}$$

$$= \frac{P_R|h_{RD}|^2G^2+1}{P_R|h_{RD}|^2G}$$

The equalized scaling coefficient can be written as in (3.33) as:

$$\hat{a} = \frac{P_R|h_{RD}|^2\hat{G}^2}{P_R|h_{RD}|^2\hat{G}^2+1} \tag{3.35}$$

$$= \frac{\left(P_R|h_{RD}|^2G^2+1\right)^2}{(P_R|h_{RD}|^2G^2+1)^2+P_R|h_{RD}|^2G}$$

Note that in the second time slot, the output signal of the relay can be expressed as:

$$t(t) = G\left(\sqrt{P_s}h_{SR}x(t) + z\right) \tag{3.36}$$

The relay gain, G, in (3.26) and (3.31) are, therefore, formulated to satisfy the constraint that the average transmits power at the relay is unity:

$$\varepsilon[|t|^2] = \varepsilon[P_s|h_{SR}|^2|G|^2] + \varepsilon\left[|G|^2\sigma_n^2\right] = 1 \tag{3.37}$$

where the operator, $[.]$, denotes statistical expectation. Thus, the scaling coefficient, a, also satisfies this constraint as it is restricted between zero and one. Interestingly, the equalized scaling coefficient, $\hat{a}$, is also bounded between 0 and 1. In the work (Rasouli, 2012), arbitrary values of scaling (cooperation) coefficients were assigned to different cooperating group of users to compute their achievable data rate on each subcarrier. Instead of arbitrary assignment, the equalized scaling coefficient, which is adaptively generated from relay gain, will be used to compute users' achievable data rate. The comparison of $\hat{a}$ and a in HD mode is as shown in Table 3.2.

Table 3.2: Comparison of scaling coefficients

Subcarrier	1	2	3	4	5	6	7	8
$Y = \frac{P_S \lvert h_{SR} \rvert^2}{\sigma_n^2}\ (dB)$	16.310	19.444	22.868	24.748	24.991	26.059	26.843	31.352
a	0.5381	0.4262	0.7471	0.8549	0.6961	0.6343	0.5417	0.2441
$\hat{a}$	0.8009	0.8035	0.8411	0.8896	0.8254	0.8117	0.8011	0.8442

3.5 Resource Allocation Problem Formulation

The idea of utility pricing system is to map the channel frequency resource and QoS requirements into the corresponding evaluation values, and then solve the established utility-based optimization problem by taking appropriate DSA scheduling (Song & Li, 2005a). On this basis, the utility function for the cross-layer DSA scheduling can be formulated as (x_m) , which is relative to some generic variable x_m that can represent a resource usage or QoS metric of user m. The objective function for the utility-based cross-layer optimization can be formulated as (Song, 2005):

$$max_{\Re_m} \sum_{m=1}^{M} U(x_m) R_m[t] \tag{3.38}$$

Subject to

C1: $\bigcup_{m=1}^{M} \Re_m \subseteq \Re$

C2: $\Re_m \cap \Re_n = \emptyset,\ \ n \neq m, \forall n, m \in \{1,2, \dots \dots M\}$

where M is the total number of users in a cell, $\Re$ is the set of all subcarriers in the system, $\Re_m$ is the subset of subcarriers assigned to user m and $R_m[t]$ is the instantaneous data rate of user m in time slot t. The optimization problem in (3.38) is the maximization of the utility weighted sum rate. Constraints (C1) and (C2) state that the union of all subsets of subcarriers assigned to different users must be contained in the total set of subcarriers available in the system, and that these subsets must be disjoint, i.e., the same subcarrier cannot be shared by two or more users in the same time slot.

Power constraints are not included because joint optimization of subcarrier and power is nonlinear and so it is complex to solve. Besides, optimal solutions are often difficult to be found. However, sub-optimal solutions that have been proposed in literature considered segregating the problem into two steps: first, dynamic resource assignment with fixed power allocation, second, adaptive power allocation with fixed resource assignment (Jang & Lee, 2003).

The works (Shen et al., 2003; Tung & Yao, 2002) have found that adaptive power allocation does not offer substantial gains over equal power allocation at high SNRs. Furthermore, Rhee and Cioffi (2000) noted that equal power allocation offers a low complexity. When equal power allocation is applied, the problem in (3.38) has a closed form solution as its objective function is now linear with respect to $R_m[t]$; thus reducing the optimization problem to a dynamic subcarrier allocation problem. Therefore, the optimization objective function with equal power allocation among subcarriers is to maximize the total utility weighted sum rate (Liu et al., 2007; Song, 2005).

The optimization problem can be formulated, respectively, for non-relay and relay networks as:

$$max_{\Re_m} \sum_{m=1}^{M} U(x_m).min\left(\frac{B}{K} log2\left(1 + \frac{P_m[k,t]|h_m[k,t]|^2}{\sigma_n^2\Gamma}\right), \frac{Q_m[k,t]}{T_s}\right) \quad (3.39)$$

$$max_{\Re_m} \sum_{m=1}^{M} U(x_m).min\left(\frac{B}{K} log2\left(1 + \frac{P_{S_m}[k,t]\left|h_{SD_m}[k,t]\right|^2}{\sigma_n^2\Gamma} + \frac{P_{S_m}\left|h_{SR_m}[k,t]\right|^2}{\sigma_n^2\Gamma}\hat{a}_m[k,t]\right), \frac{Q_m[k,t]}{T_s}\right)$$

(3.40)

where utility function $U(x_m)$ corresponds to the scheduling weight for user m on subcarrier k in time slot t, $Q_m[k,t]$ is the queue for user m. The $P_m[k,t]$ and $P_{S_m}[k,t]$ are the equal powers on each subcarrier; $|h_m[k,t]|^2$, $\left|h_{SD_m}[k,t]\right|^2$ and $\left|h_{SR_m}[k,t]\right|^2$ are channel gains; σ_n^2 is the noise power, $\hat{a}_m[k,t]$ is the equalized scaling coefficient and $Q_m[k,t]$ is the queue for user m on subcarrier k in time slot t. The B and K are the total system bandwidth and the total available subcarriers. The slot duration and the SNR gap are given by T_s and Γ, respectively.

At the beginning of time slot t, user m is assumed to have $Q_m[t]$ bits in the queue. If there are $A_m(t)$ bits arriving during time slot t, the queue length at the end of this time slot, assuming queues of infinite capacity, can then be expressed as

$$Q_m[t+1] = Q_m[t] - R_m[t]T_s + \bar{\omega}_m \quad (3.41)$$

where $R_m[t] = min\left(r_m[k,t], \frac{Q_m[t]}{T_s}\right)$, $\bar{\omega}_m = \mathrm{E}(A_m[t])$ is the mean arrival rate, $A_m[t]$ is the arrival bits, T_s is the slot duration and $r_m[k,t]$ is user m data rate in subcarrier k, time slot t. An active user is defined here as the user who has data to transmit during time slot t, i.e. $N_m = \{m: Q_m(t) > 0\}$ is a set in which each queue associated with a user m is not empty at time slot t.

In order to avoid wasting resources and thus fulfill the frugality constraint (FC) (Song, 2005), the cross-layer resource allocation strategy selects a transmission rate:

$$r_m[k,t] \leq \frac{Q_m[t]}{T_s} \tag{3.42}$$

This is achieved with the function $min(\cdot)$. Therefore, the objective function of each of the algorithms selects the user with index m^* to transmit on the subcarrier k at the time slot t if the condition below is satisfied:

$$m^* = \arg max_{\Re_m} \left\{ U(x_m).min\left(r_m[k,t], \frac{Q_m[k,t]}{T_s}\right)\right\} \tag{3.43}$$

where the instantaneous achievable transmission rate of the subcarrier k with respect to user m during time slot t, $r_m[k,t]$ corresponds to the first argument of the $min(.,)$ in (3.39) and (3.40). The AMC for WiMAX standards is as shown in Table 3.3. However, the AMCS can also be computed (Rodrigues & Casadevall, 2009) on the data rate as:

$$r'_m[k,t] = 2.round\left(\frac{r_m[k,t]}{2}\right) \tag{3.44}$$

Therefore, the objective function when AMC is applied selects the user with index m^* to transmit on the subcarrier k at the time slot t if the condition below is satisfied:

$$m^* = \arg max_{\Re_m} \left\{ U(x_m).min\left(r'_m[k,t], \frac{Q_m[k,t]}{T_s}\right)\right\} \tag{3.45}$$

The optimization objective function can be regarded as simply a dynamic resource allocation, whose weights are adaptively controlled by the utility function. .

Table 3.3: Adaptive Modulation and Coding Schemes (AMCS) for IEEE 802.16 OFDMA PHY (Zhu et al., 2008)

Mode	Modulation	Coding rate	Information (bits/symbol)	Receiver SNR (dB)
1	QPSK	1/2	1.0	5
2	QPSK	3/4	1.5	8
3	16-QAM	1/2	2.0	10.5
4	16-QAM	3/4	3.0	14
5	64-QAM	1/2	3.5	16
6	64-QAM	2/3	4.0	18
7	64-QAM	3/4	4.5	20

The pseudo codes for the implementation of the subcarrier assignment for RT and NRT traffics are as depicted, respectively, in Algorithms 3.1 and 3.2 in Appendix B which consider equal power over subcarrier with frugality constraint (FC). In order to ensure that the system's delay budget is satisfied, packet dropping policy is implemented for RT traffic, such that HOL packets are dropped when their maximum tolerable delays are exceeded. Generally, packets are also dropped when the buffer is full.

3.6 System Model and Assumptions

An OFDMA-based wireless network as shown in Figure 3.4 is considered. The total bandwidth of B is divided into K independent subcarriers and shared by M users who are randomly located at various distances and angles from their serving base station (BS) within a cell; therefore each user experiences a path loss. The base station (BS) transmits a total power of P which is uniformly allocated (uniform power allocation) among the total number subcarriers K, and it is assumed to be equipped with single transmit antenna to provide service to M active users, each equipped, without loss of generality, with a single receive antenna. Transmission between the BS and active users or mobile stations (MSs) takes place in time slots of a fixed duration, T_s, which is assumed to be less than the channel coherence time τ_c . Thus, the channel gain, h_m, is constant during each time slot and is independent of the channel for other time slots; a quasi-static fading channel is assumed.

In the BS, the incoming packets of each user arrive from some upper layers and then buffered in its first-come-first-out (FIFO) queue with a finite space of F bits waiting to be scheduled. It is assumed that each user (or subscriber) only has one traffic flow which can be chosen from Voice over Internet Protocol (VoIP), video/audio streaming, File Transfer Protocol (FTP) and Hypertext Transfer Protocol (HTTP) as depicted in Figure 3.4; and each traffic flow m is assigned a queue Q_m.

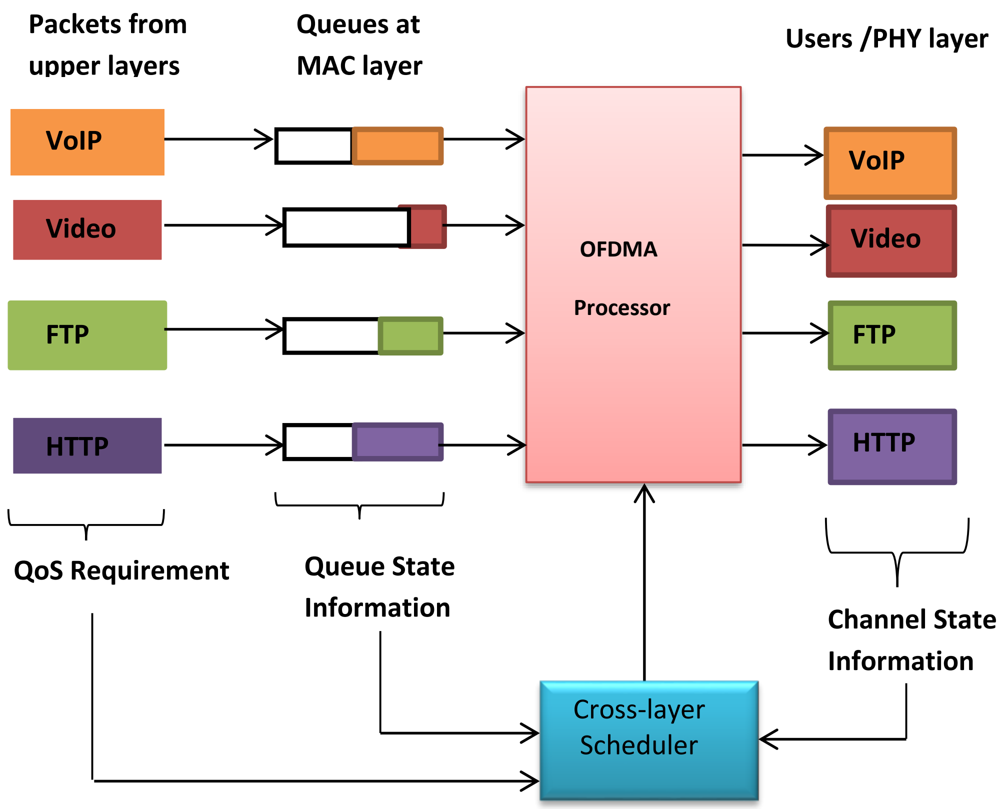


Figure 3.4: OFDMA-based wireless network model

In the wireless network model in Figure 3.5, the mobile stations (MS) or users are randomly distributed in the cell and equally divided among the different classes service when they are multiplexed for resource allocation. The channel between BS and MS, are modeled as non-light-of-sight (NLOS).

Figure 3.6 illustrates the cooperative relaying model. As shown, there are three relay stations (RSs) in a single cell. The green circle is BS Region, the users located in this BS region is served directly by the BS. The blue area is RS regions. The users in the RS region are divided into 3 sub-regions: RS1, RS2, RS3 regions. The users in each of the 3 sub-regions can be served by the BS directly or indirectly through the corresponding RS. Hereafter, the BS shall be referred to as the source (S), the RS as the Relay (R) and MS as the Destination (D) nodes, respectively.

It is assumed that S-R and R-D links use the same frequency band, and S-R and R-D transmissions follow TDMA protocol. We further assume that transmission within the sector is synchronized so that there is no intra-sector interference. Multiuser interference from other sectors and cells is ignored to simplify the implementation of the scenario. The channel between S and D; and between R and D are modeled as light-of-sight (LOS), whereas the channel between S and R are considered NLOS. However, this work considers that users are only located in the RS region and none in the BS region to enable performance analysis to capture the effect of relaying. Furthermore, only one relay region will be analyzed. The two network models assume that the links are error-free and this can be justified, e.g., by saying that the link uses directional antennas in both ends.

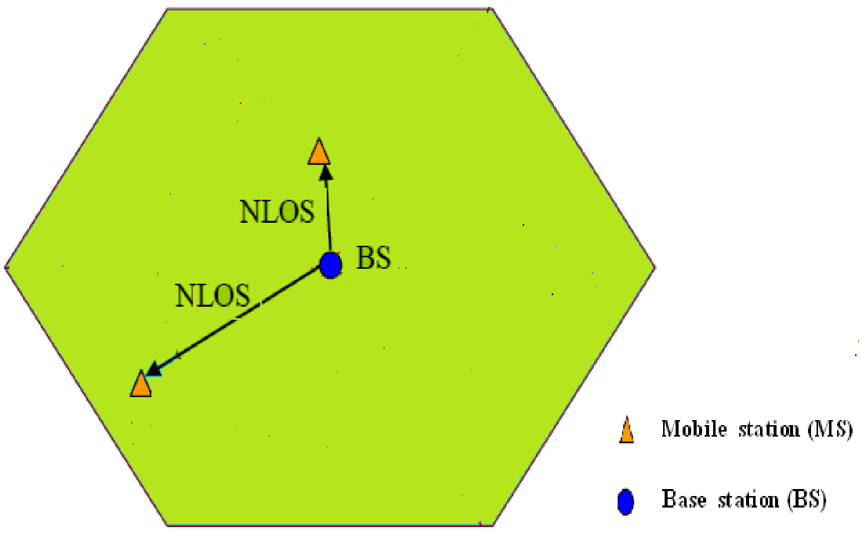


Figure 3.5: Multi-user wireless network model

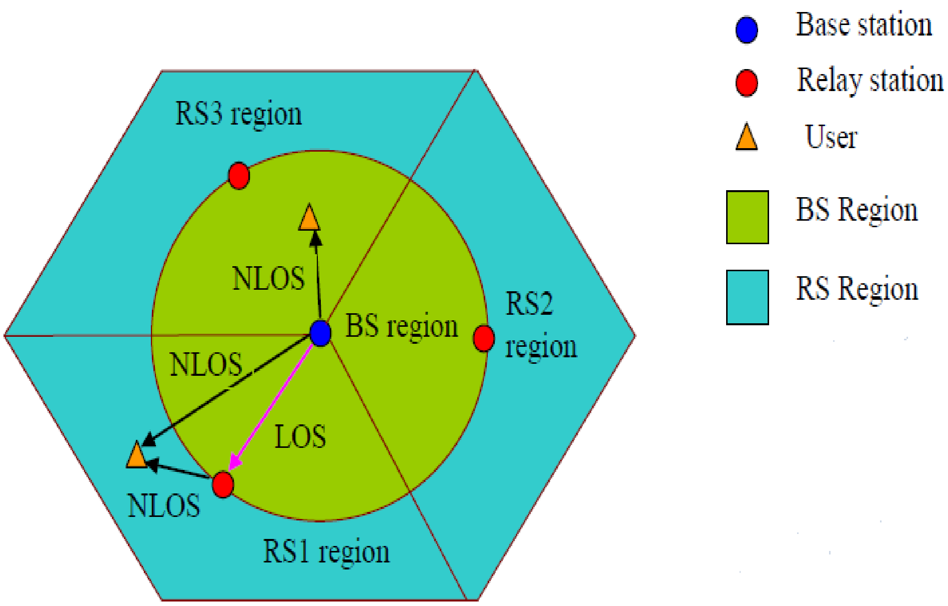


Figure 3.6: Multi-user Multi-relay wireless network model

3.7 Performance Metrics

The following performance metrics such as average system throughput, throughput fairness and user satisfactions are compared for different scheduling and resource allocation algorithms in the downlink of an OFDMA system:

3.7.1 Average System Throughput

Average system throughput is the average number of successfully delivered bits over the lifetime of the user's connection. Mathematically, average system throughput is the data rate, $R_m[k,t]$, of user m averaged over subcarriers $k \in (1,K)$ and time slots $t \in (1,T)$ and expressed by:

$$\overline{T}_m = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} R_m[k,t] \tag{3.46}$$

3.7.2 User Call Satisfaction

Generally, packets in excess of base station buffer size are dropped and so also are HOL packets of RT traffic that exceed their budget delays. Therefore, scheduling algorithm for RT traffic is designed to ensure that most of the network resources are allocated within the packets' deadlines. For NRT traffic packet, the requirement of the scheduling algorithm is to ensure that it is allocated network resources equal to its minimum data rate requirement during the entire session duration. Session duration corresponds to the number of scheduling epochs and in simulation it is taken as equal to the number of time slots. Therefore, user call satisfaction allows us to compare scheduling algorithms with respect to their abilities to guarantee satisfaction of QoS requirement for each application. Users are said to be satisfied when resources are allocated within the delay budget for RT user and minimum data rate target is met within the NRT user session duration.

The number of satisfied users in NRT and RT service classes can be expressed (Femenias et al., 2012), respectively, by:

$$J_{RT}^{sat} = \begin{cases} 1, & D_m^{hol}[t] \leq D_m^{max} \\ 0, & Otherwise\ (\ HOL\ packet\ is\ also\ dropped) \end{cases} \tag{3.47}$$

$$J_{NRT}^{sat} = \begin{cases} 1, & \overline{R}_m[t] \geq \bar{R}_m^{min}\ over\ the\ session\ duration \\ 0, & Otherwise \end{cases} \tag{3.48}$$

The user call satisfaction is, therefore, the ratio between the numbers of satisfied users J_i^{sat} to the total number of users J_i in service class i. Mathematically, this can be expressed in percent by:

$$SI_i = \frac{J_i^{sat}}{J_i} * 100 \tag{3.49}$$

where SI_i ranges between 0 and 100.

3.7.3 User Call Fairness

Several computations of the system-centric fairness index have been proposed. The most popular one is the Jain's fairness index (Jain et al., 1984) which is modeled by:

$$JFI_{\mathcal{C}} = \frac{(\sum_{m\in\mathcal{C}} \psi_m)^2}{|\mathcal{C}| \sum_{m\in\mathcal{C}} (\psi_m)^2} \quad , \psi_m \geq 0 \ \ \forall m \tag{3.50}$$

where ψ_m is the QoS-based performance metric (i.e. average throughput or delay) for user m in the set of users $\mathcal{C}$ who belong to the same QoS class and $|\mathcal{C}|$ denotes the number of users in that set (Chia et al., 2008). Note that if all users in $\mathcal{C}$ get the same ψ_m, then $JFI_{\mathcal{C}} = 1$ indicating a totally fair allocation. If all the resources are allocated to only one user, then $JFI_{\mathcal{C}} = 1/|\mathcal{C}|$. However, as $|\mathcal{C}|$ tends to ∞, $JFI_{\mathcal{C}}$ gets close to 0. Therefore, $JFI_{\mathcal{C}}$ is bounded between 0 and 1.

3.8 Chapter Summary

The methodology that was used in developing the scheduling algorithms proposed in this thesis is described. Firstly, related utility-based scheduling rules are presented, and then a new one based on novel sigmoidal-type utility functions is proposed. An AF scaling coefficients are derived for HD and FD operating modes and analyzed. Because, they are unequally distributed across subcarriers, modified versions called equalized scaling coefficients are developed for subcarrier allocation in relaying systems. Then a simple optimization method of dynamic subcarrier assignment is developed to study the effects of a mixed service scenario on system capacity, fairness and user satisfaction in the downlink of an OFDMA wireless system.

Chapter 4

Results and Discussion

4.1 Simulation Parameters

The simulation considers a system bandwidth of 5MHz divided into 128 subcarriers and slot duration of 2.0571ms based on WiMAX standards (Andrews et al., 2007). The total transmit power at source and noise power at the receiver front-ends are set at 33.989dBm and -151dBm, respectively. The Rayleigh flat fading which is based on the Stanford University Interim (SUI) channel model 4, which is the widely adopted channel model for simulating and testing WiMAX systems (Erceg et al., 1999). The transmitted signal also undergoes a distant-dependent path-loss given by:

$$PL(d) = 20 * log\left(\frac{4\pi d_0}{\lambda}\right) + 10 * n * log\left(\frac{d}{d_0}\right) + \chi_0 \quad [dB] \tag{4.1}$$

where $d(m)$ is the distance between the BS (Source) and MS (Destination) or between RS (Relay) and MS (Destination), path-loss exponent, $n = 3$; reference distance, $d_0 = 100m$; shadowing, $\chi_0 = 8dB$ and wavelength, $\lambda = 120mm$ at a carrier frequency of 2 GHz in a cell with radius, $n = 1000m$. The SNR gap, $\Gamma = -[ln(5 * BER)/1.5]$, where $BER = 10^{-5}$ is assumed. To capture the dynamics of the system in the time domain as well as the traffic pattern of the users, iteration will be performed over 10,000 time slots for slot duration of 2.0571ms, which corresponds to a simulation time of 20s. The AF-FD outperforms AF-HD at average loop interference power of $\Upsilon_{LI} \leq 26dB$ (Ng et al., 2012); and to ensure that for this analysis, it is set at $\Upsilon_{LI} = 10dB$. The Matlab codes used for the simulations are as shown in Appendix C.

4.2 Traffic Model

It is considered that a user only has one traffic flow which can be chosen from Voice over Internet Protocol (VoIP), Video streaming, File Transfer Protocol (FTP) and Hypertext Transfer Protocol (HTTP). The buffer size of each traffic flow is set at $\mathrm{F} = 128$ Kbytes. The parameter setting for the traffic is as shown in Table 4.1 where the settings except packet sizes are adapted from Table 2.3, which are based on WiMAX standards. The minimum reserved data rates are used as the packet arrival rate for each service class.

Table 4.1: Traffic model based on WiMAX standards

Class	Service type	Minimum reserved data rate (kbps)	Maximum sustainable data rate (kbps)	Packet Size (bits) (L)	Maximum required delay (Latency) (D^{max}) (ms)	Minimum required throughput (R^{min}) (kbps)
ertPS	VoIP	25	64	512	20	N/A
rtPS	Video streaming	64	500	640	30	N/A
nrtP	FTP	45	500	1200	N/A	45
BE	HTTP	1	64	512	N/A	N/A

4.3 Utility-based Scheduling and Subcarrier Allocation

4.3.1 Subcarrier Allocation in a Scenario with Real-Time Services

This section compares the average system throughput, user call satisfaction and user call fairness against user arrival rate for MQS, DSM/TSM, PF and MSR scheduling algorithms. User arrival rate is defined as the number of users arriving in the cell per timeslot. In RT simulation set-up, packet dropping policy is implemented; as such HOL packet is dropped from its associated buffer if it exceeds its delay budget (maximum tolerable delay). Figure 4.1 depicts the average system throughput as a function of the number of video streaming users. As it can be seen, the MQS achieves a slightly better average throughput performance than DSM at high user arrival rates.

One would expect the opportunistic schedulers such as PF and MSR to present higher average system throughputs than the delay-aware schedulers such as MQS and DSM/TSM. However, this was not so because PF and MSR do not provide in their utility functions a means of avoiding excessive delays as do MQS and DSM/TSM. Besides, the MSR maximizes the system capacity by always choosing a few users, and by nature of the RT traffic model used, the buffers of these few users do not have so much data to transmit hence the poorest throughput performance. Similarly, PF only slightly provides higher average system throughput than MSR, as it attempts to increase the number of selected users by using the user's relative channel condition to fairly distribute the available system resources.

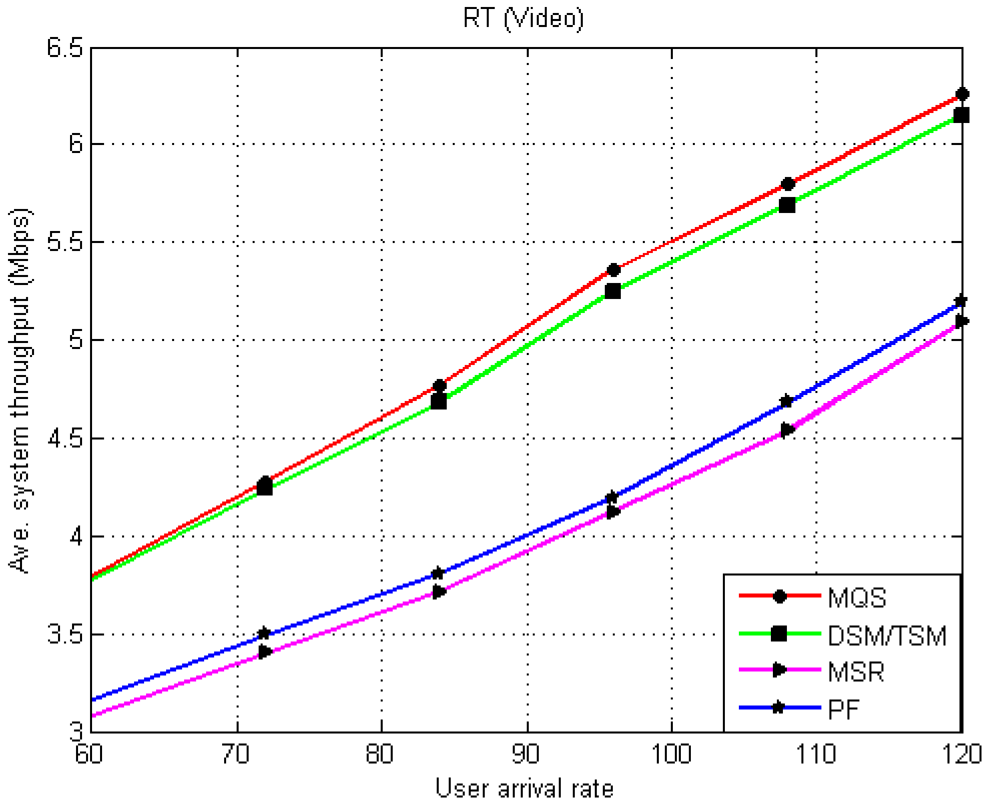


Figure 4.1: Average system throughput for Video services

Figure 4.2 shows the user call satisfaction for different scheduling schemes. It can be seen, that as the traffic load increases users become less satisfied because the available resources have to be shared by the increasing number of users. However, both the MQS and DSM/TSM provide the highest user call satisfaction than the PF and MSR schedulers. Generally, with RT traffic scenario, the delay-aware schedulers tend to perform better than the opportunistic ones, because they are more adapted to avoiding excessive delays which results in satisfying more users in terms of their QoS requirements.

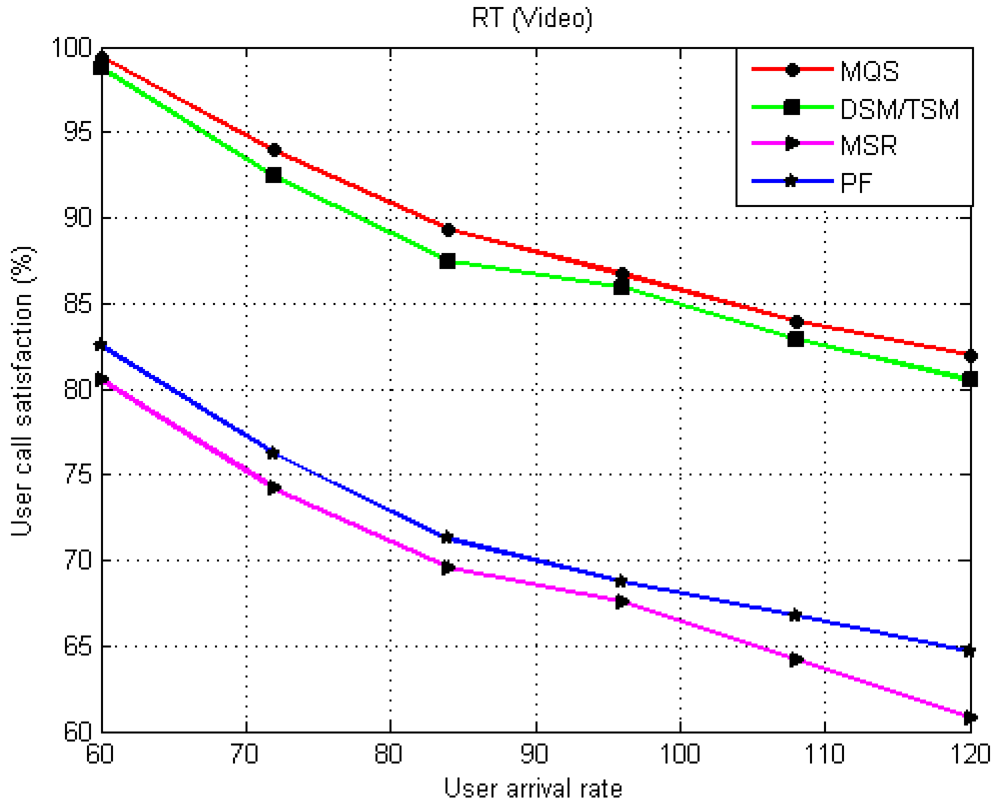


Figure 4.2: User call satisfaction for video streaming services

In Figure 4.3, the curves of the call fairness seem to follow the pattern of the user call satisfaction in Figure 4.2 for all the scheduling algorithms. This is because, in the scheduling framework HOL packets whose instantaneous delays exceed their maximum tolerable delays are dropped from the buffer and the entire traffic is not allocated subcarrier resources during that scheduling epoch. This measure makes the user call fairness show a similar pattern as the user call satisfaction.

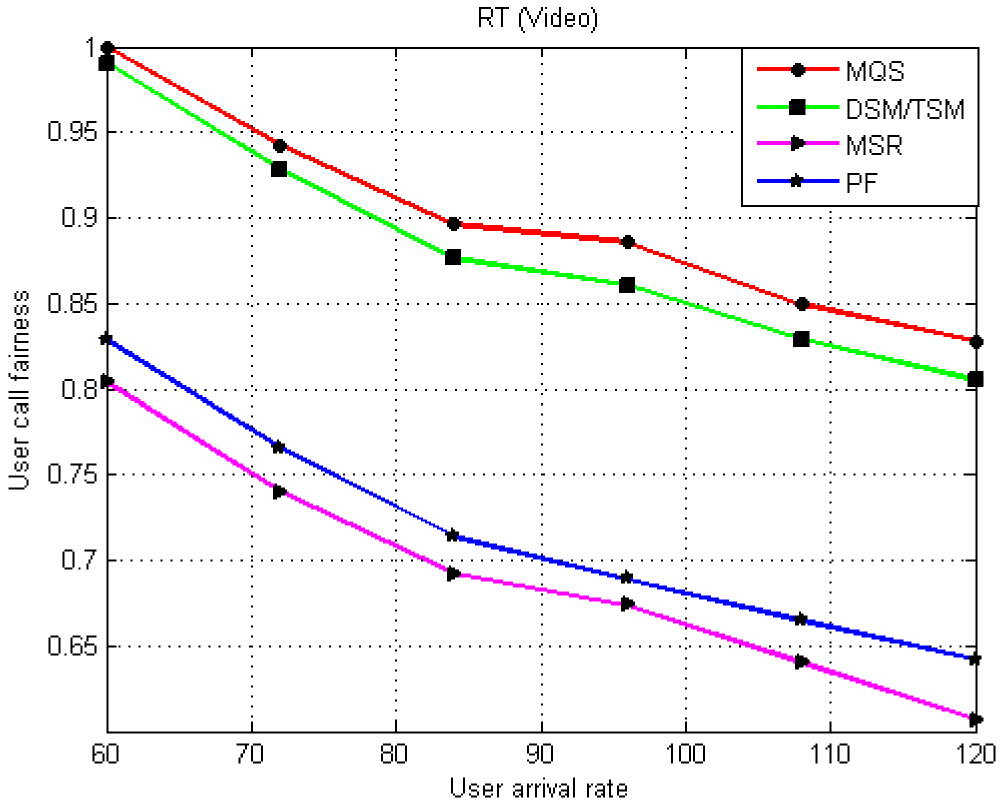


Figure 4.3: User call fairness for Video services

4.3.2 Subcarrier Allocation in a Scenario with Non-Real-Time Services

Here, the performance metrics for the MQS, DSM/TSM, PF and MSR scheduling policies in a scenario with FTP traffic are compared. The average system throughput for various FTP traffic loads is depicted in Figure 4.4. It can be seen that MQS and DSM/TSM provide highest average system throughputs at all user arrival rates, in that order. The figure also shows that MSR provides a superior performance than PF; because are able to exploit the multi-user diversity better.

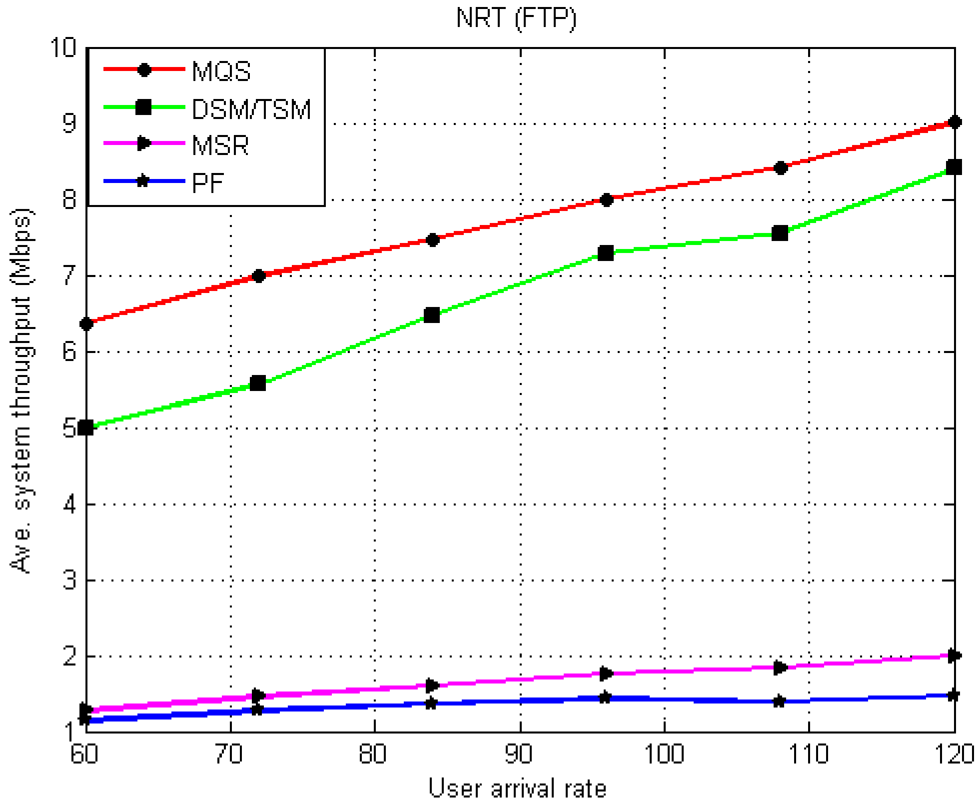


Figure 4.4: Average system throughput for FTP services

As depicted in Figure 4.5, the MQS achieves the highest user call satisfactions than the DSM/TSM, PF and MSR in that order. The superior performances of MQS and DSM/TSM is because they both have integrated in their utility function both the QoS metric of average data rate and minimum required throughput (data rate), which effectively controls the resource usage. The PF uses only the average data rate in its utility function to select much larger number of users than the MSR to satisfy in terms of their required minimum throughput requirement, hence the poorest user call satisfaction performance achieved by the MSR.

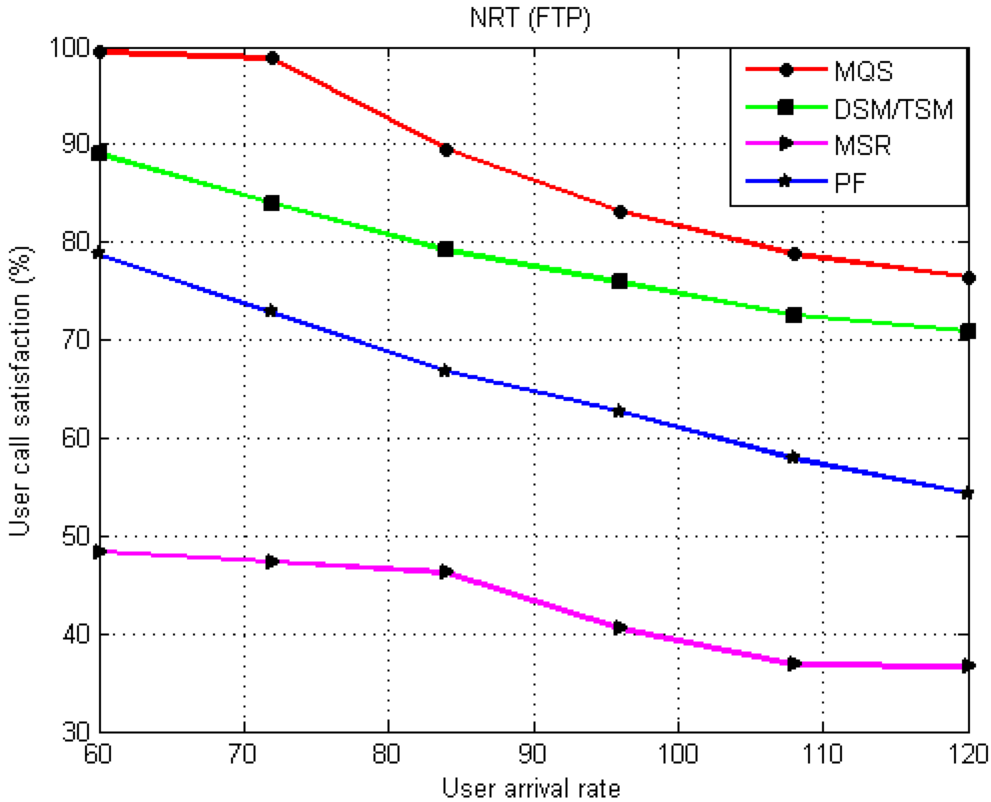


Figure 4.5: User call satisfaction for FTP services

As shown in Figure 4.6, the MQS and DSM/TSM achieve the highest call fairness. This is possible because they both use utility functions that control the usage of the resources and thus share these resources in a more controlled and fair manner. Although, the PF is designed to provide proportional fairness, its fairness performance is poorer compared to both the MQS and DSM/TSM. The reason may be that both MQS and DSM/TSM use sigmoidal type of utility functions which inherently equalize the utilities derived better than does the average user throughput used in PF for the same purpose. The MSR, as expected shows the worst fairness performance, because it opportunistically allocates the resources without concern for fairness.

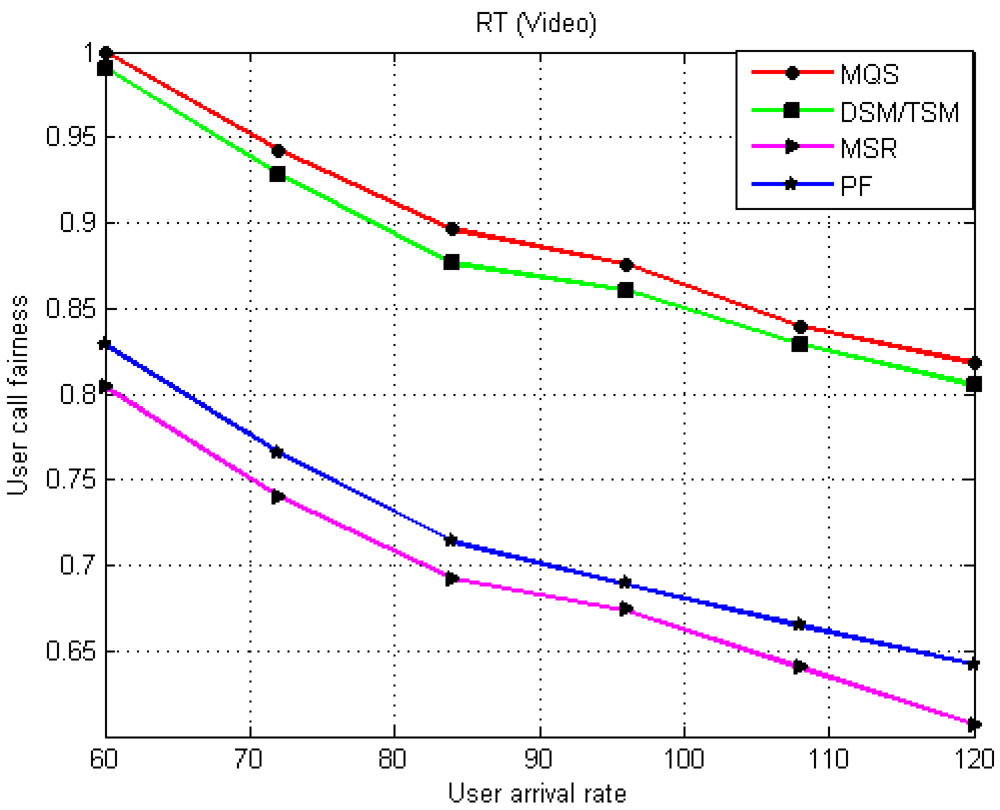


Figure 4.6: User call fairness for FTP services

4.3.3 Subcarrier Allocation in a Scenario with Mixed Real-Time and Non-Real-Time Services

This section compares the average system throughput, user call satisfaction and call fairness when an RT traffic providing a video streaming service is multiplexed with an NRT traffic providing an FTP service for some selected scheduling algorithms. Distribution of throughput resources among different traffic types depends, among others, on the arrival rate of that traffic and the utility function which controls their QoS requirements.

As depicted in Figure 4.7, the MQS, PF and MSR allocate higher throughput to FTP than to video streaming users. This is because the FTP packets arrive at a higher rate than that of video streaming. This can contribute to its higher throughput allocation since queue with greater number of bits but having the same channel quality with other queues, usually gets a higher priority to transmit. However, the DSM/TSM is rather seen to allocate almost all the throughput to video streaming users at the expense of FTP ones who appear completely starved as the user arrival rate increases above 72. The figure also shows that both the MSR and PF allocate much throughput resources to FTP than to video streaming, since they are throughput-oriented scheduling schemes. The MSR compensates for lower average throughput allocated to video streaming compared to the MQS and DSM/TSM by providing the highest average throughput for FTP. However, the PF in allocating more throughputs to FTP completely starves the video streaming users; because the relative channel condition it uses to schedule traffic favors NRT users better.

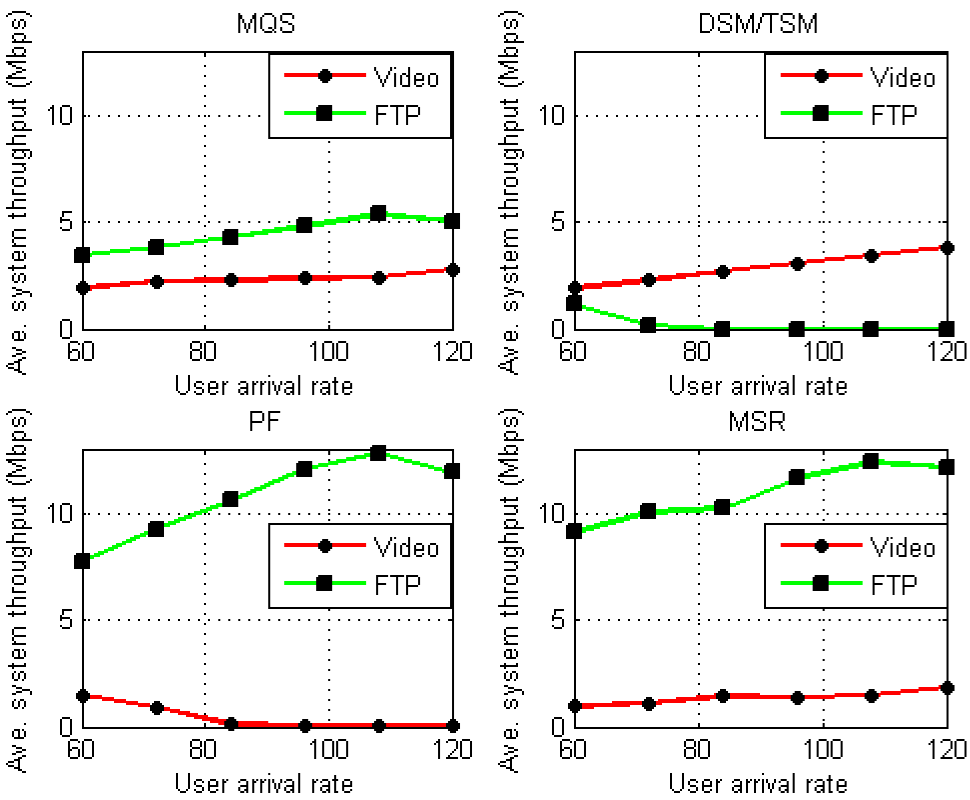


Figure 4.7: Average system throughput for a mixture of Video and FTP services

In Figure 4.8, the MQS and DSM/TSM are shown to differentiate the QoS of the video streaming and FTP by providing higher user call satisfaction for the former; the PF and MSR do the opposite, by providing higher user call satisfaction to FTP than video streaming service. Video streaming is considered a higher priority than FTP. The reason for this is that both the PF and MSR are not naturally designed to account for QoS requirements of applications, hence in mixed traffic scheduling it cannot properly differentiate. But interestingly, as it can be observed the DSM/TSM provides the highest user call satisfaction to the video streaming users at the expense of FTP users who become completely dissatisfied (zero user call satisfaction) as the user arrival increases above 92. The reason for this may be due to the bell-shaped utility function employed by DSM/TSM which is geometrically similar for both RT and NRT traffic. MQS applies different utility functions with different shape; hence it able to differentiate QoS better in mixed traffic scheduling.

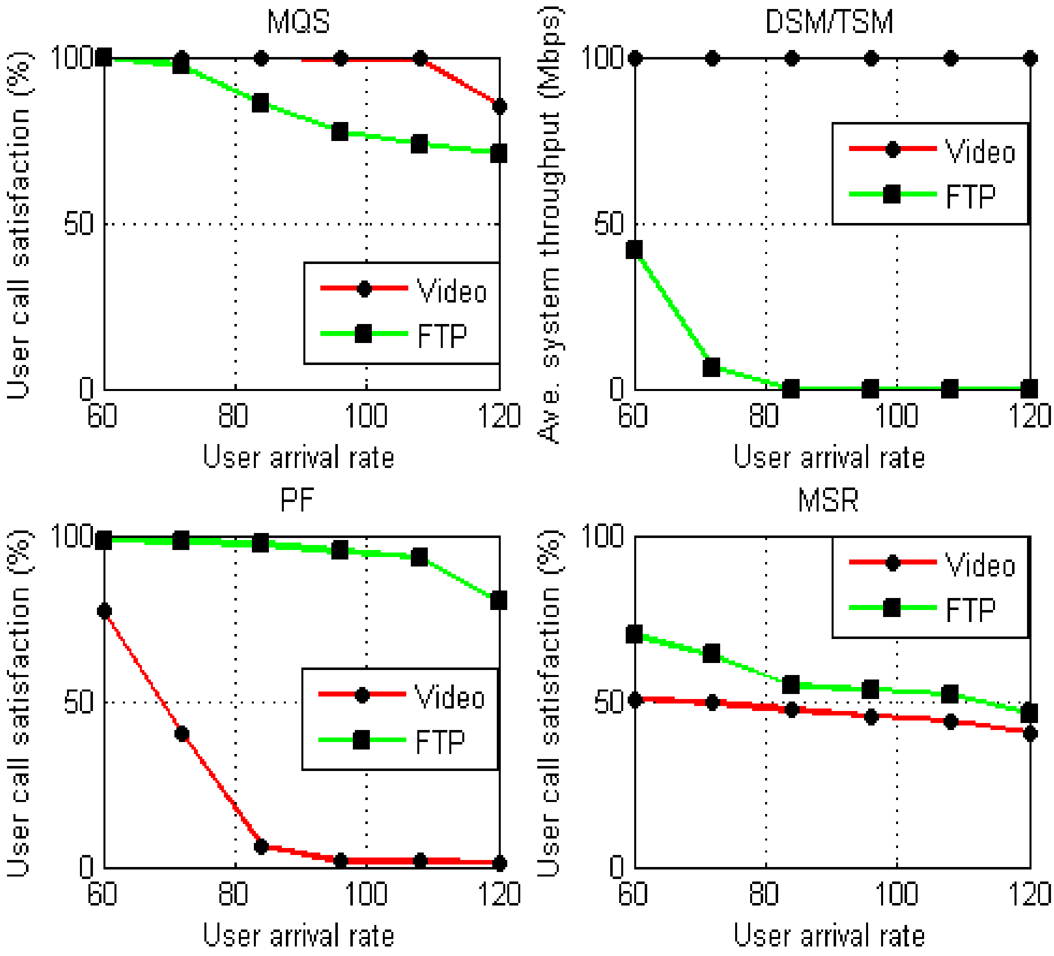


Figure 4.8: User call satisfaction for a mixture of Video and FTP services

In Figure 4.9, the DSM/TSM achieves highest call fairness for video streaming followed by the MQS. But while MQS provides highest user call fairness for FTP, in the DSM/TSM the FTP users are allocated zero call fairness above the user arrival rate of 92. Similarly, like in the case of user call satisfaction in Figure 4.8, both the PF and MSR assign higher resource allocation fairness indexes to FTP than video streaming.

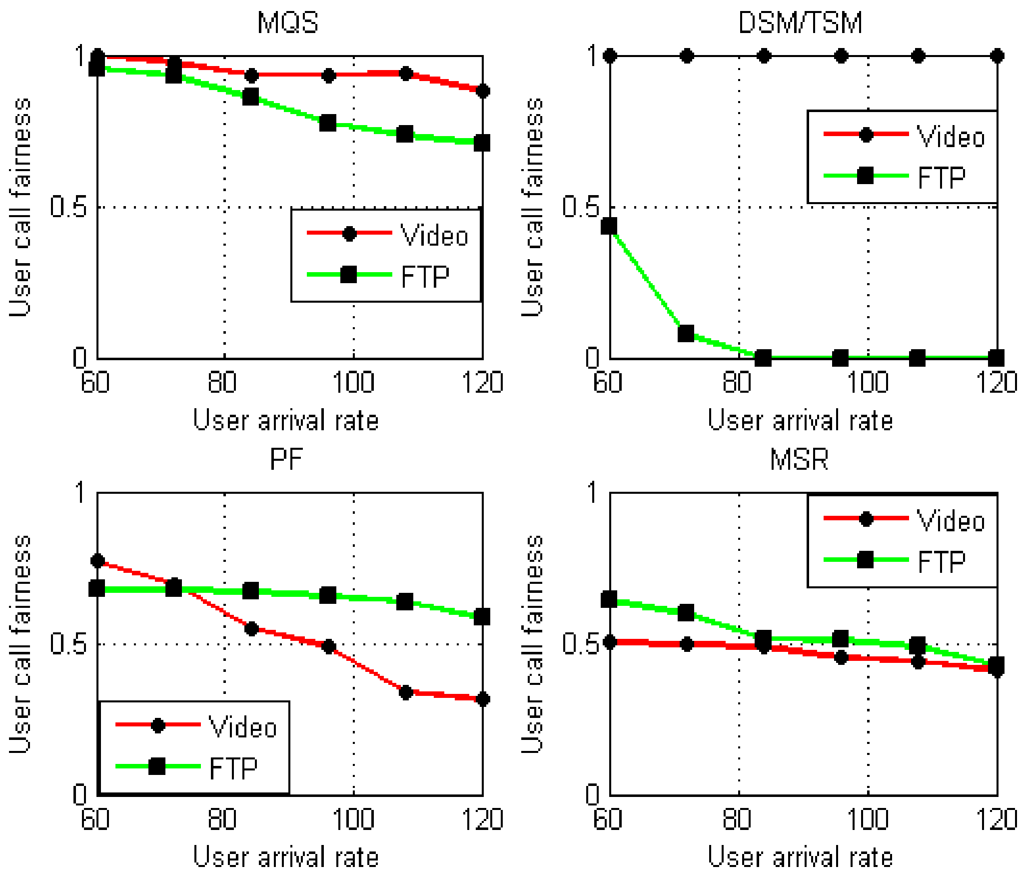


Figure 4.9: User call fairness for a mixture of Video and FTP services

When two different RT services are multiplexed as in Figure 4.10, higher average system throughputs are allocated to VoIP users than video streaming ones by the MQS and DSM/TSM as may be expected. Because frugality constraint is being applied the higher the amount of data that is available in the buffer, the greater the chances of being selected for transmission when their delay budgets are closely related. Therefore, one expects that VoIP will be given the higher priority to transmit more of the time.

. However, it can be seen in the figure also that both the PF and MSR rather allocate more throughput resources to video streaming than VoIP, with VoIP users getting zero allocation above the user arrival rate of 108, in the case of the PF scheduler.

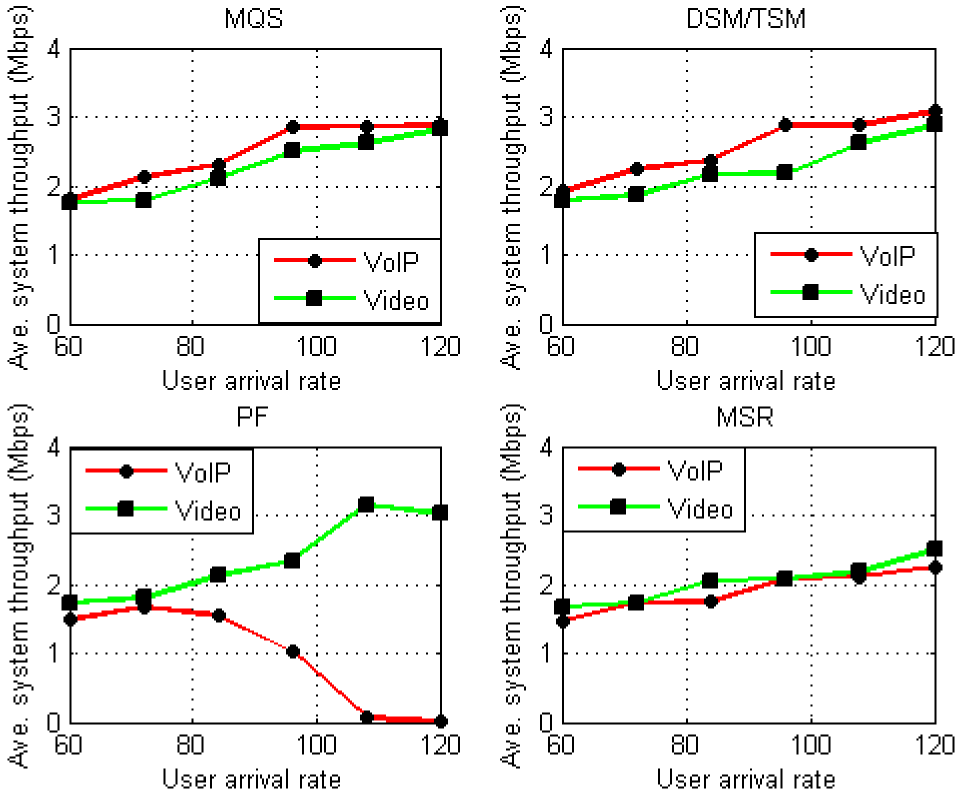


Figure 4.10: Average system throughput for a mixture of VoIP and Video services

Figure 4.11 compares the user call satisfaction assigned by the scheduling algorithms in a mixture of VoIP and Video streaming traffics (homogeneous traffic). Although, the DSM/TSM is shown to provide the best performance for VoIP the MQS achieves the highest user call satisfaction for video streaming users. The figure shows that both the MSR and PF are able to allocate higher user call satisfaction to VoIP than to video streaming; because they are not delay-sensitive and so cannot properly differentiate RT services with delay requirement.

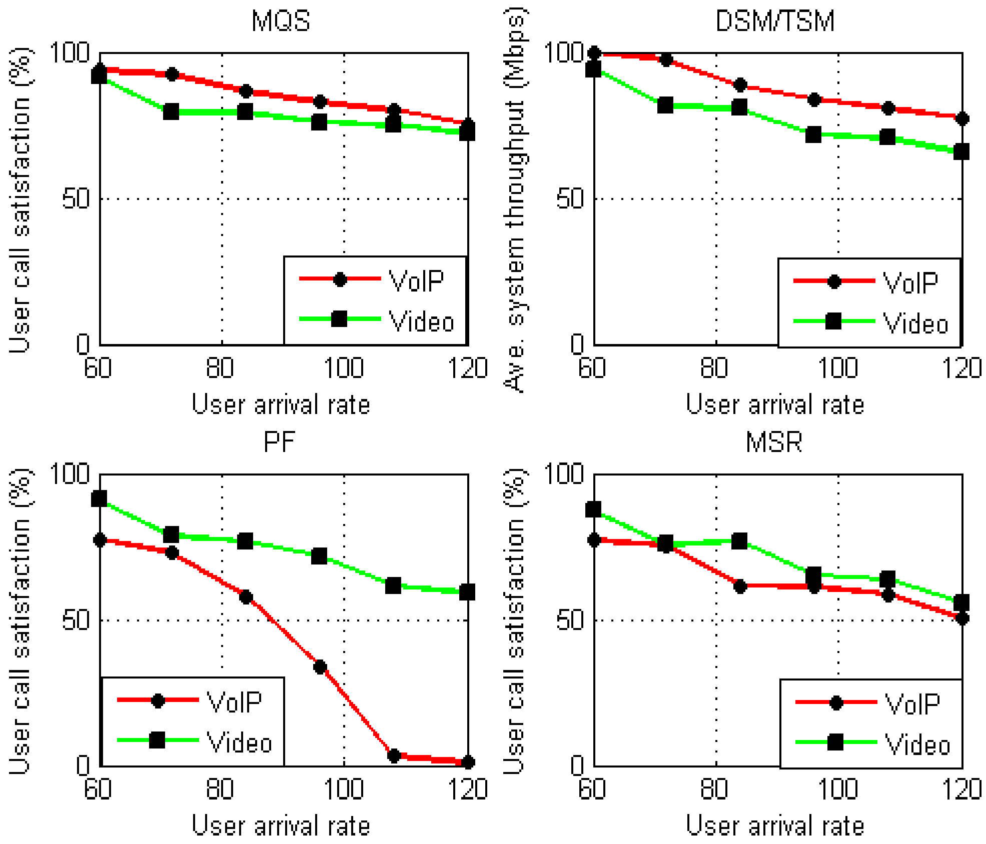


Figure 4.11: User call satisfaction for a mixture of VoIP and Video services

As shown in Figure 4.12, the call fairness for all the scheduling algorithms depicts similar pattern as in the case for user call satisfaction in Figure 4.11. This is because, call fairness and user call satisfaction provide similar measure with RT traffic as delay constraints are often applied to control the distribution of the resources.

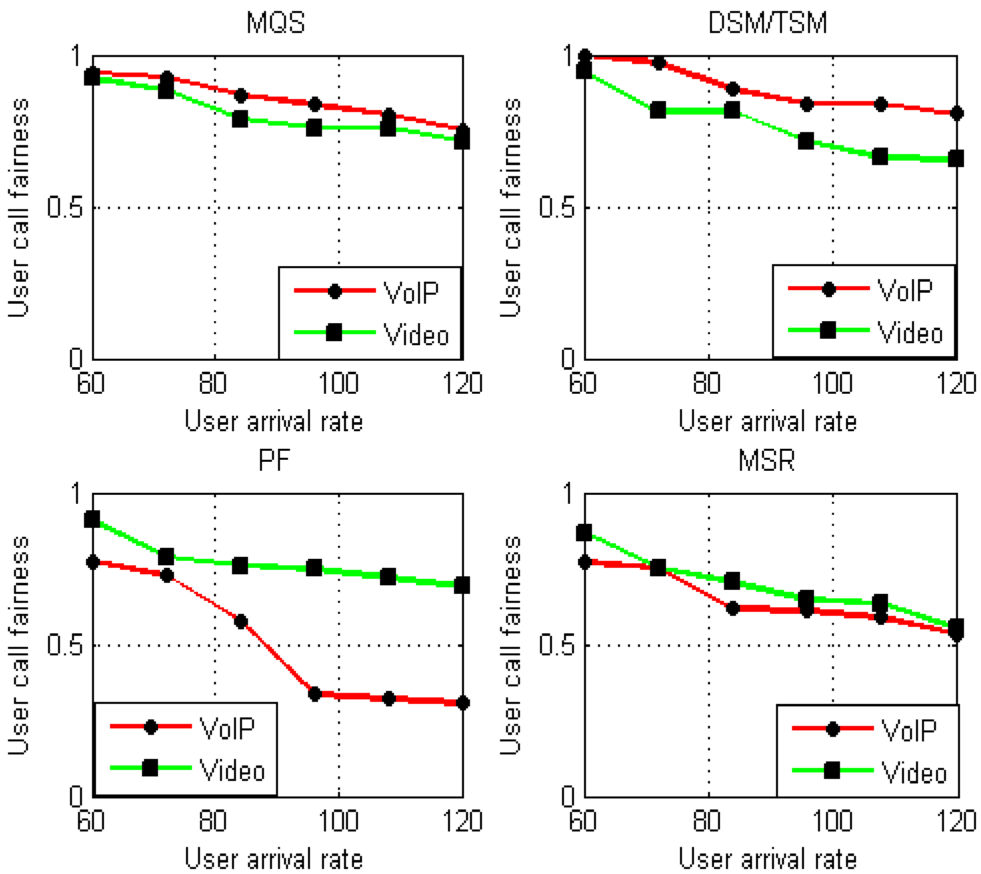


Figure 4.12: User call fairness for a mixture of VoIP and Video services

In Figure 4.13, the average system throughput is compared for FTP and HTTP, another homogeneous traffic. The figure shows that higher average system throughputs are achieved by both the MQS and DSM/TSM compared with PF and MSR for FTP users than HTTP ones. However, it can be noticed that for MQS and DSM/TSM the average throughputs in the case of video streaming initially increases from user arrival rate of 60 until arrival rate of 92 and 102, respectively, and then begin to decrease. These decreases occur for HTTP while those for FTP continue to increase. The may be because HTTP is a BE service with little or no requirement in terms of minimum throughput, while FTP needs to satisfy its higher minimum throughput requirement. Contrarily, both the MSR and PF experience increase in average throughputs for FTP and HTTP at all user arrival rates; because they do not consider their minimum throughput requirements to allocate resources.

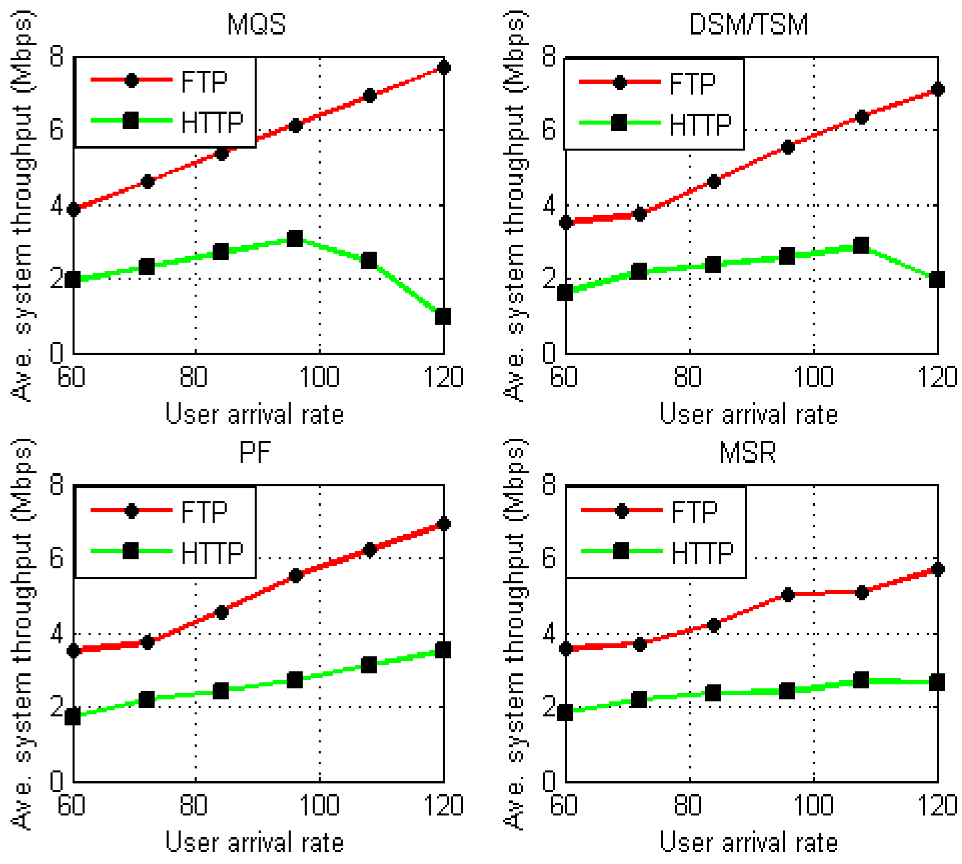


Figure 4.13: Average system throughput for a mixture of FTP and HTTP services

Figure 4.14 shows that only the MQS allocates higher user satisfaction to FTP than HTTP as would be expected; because FTP is usually considered higher priority traffic. Noticeably, the DSM/TSM, MSR and PF all assign higher user call satisfaction to HTTP than FTP making it look as if the former is of higher QoS priority. However, this is an indication that these algorithms are only unable to differentiate properly between the two homogeneous NRT services.

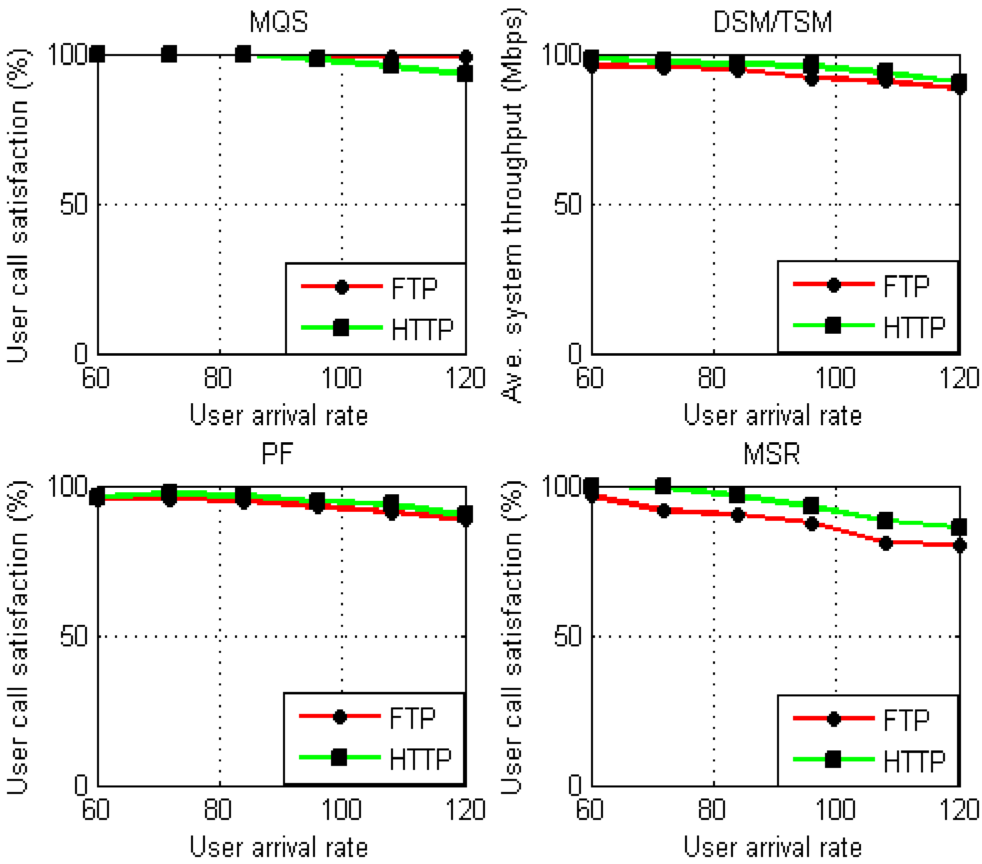


Figure 4.14: User call satisfaction for a mixture of FTP and HTTP services

As depicted in Figure 4.15, the MQS provide highest call fairness to both FTP and HTTP in that order. Surprisingly, the DSM/TSM, MSR and PF still follow the patterns in Figure 4.14 to assign higher call fairness to HTTP than FTP. The MSR achieves the lowest call fairness for the two services. In this case also, the MQS has shown the greatest consistency in considering the QoS constraints of each application to more fairly distribute the network resources.

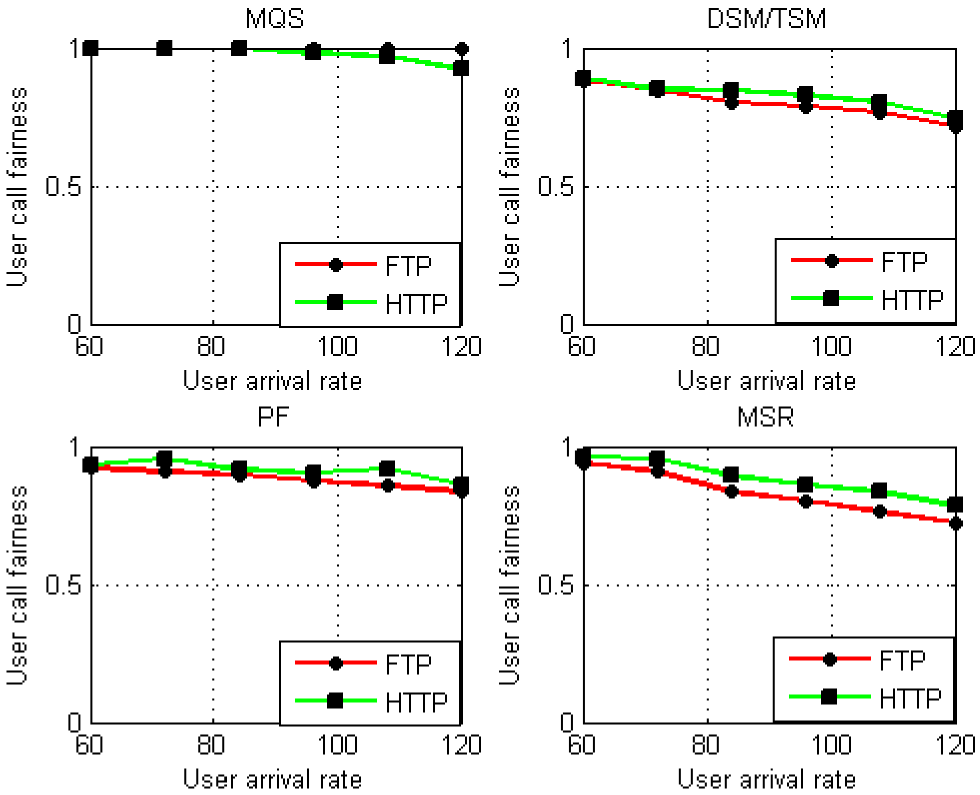


Figure 4.15: User call fairness for a mixture of FTP and HTTP services

In a multiplexed traffic scheduling of VoIP, video streaming, FTP and HTTP as shown in Figure 4.16, the MQS achieves the highest total throughput as the number of user increases, because the achievable average system throughputs for FTP and HTTP by DSM/TSM are zero above the user arrival of 84. This means that DSM/TSM cannot simultaneously accommodate traffics with diverse QoS requirements in the same cell.

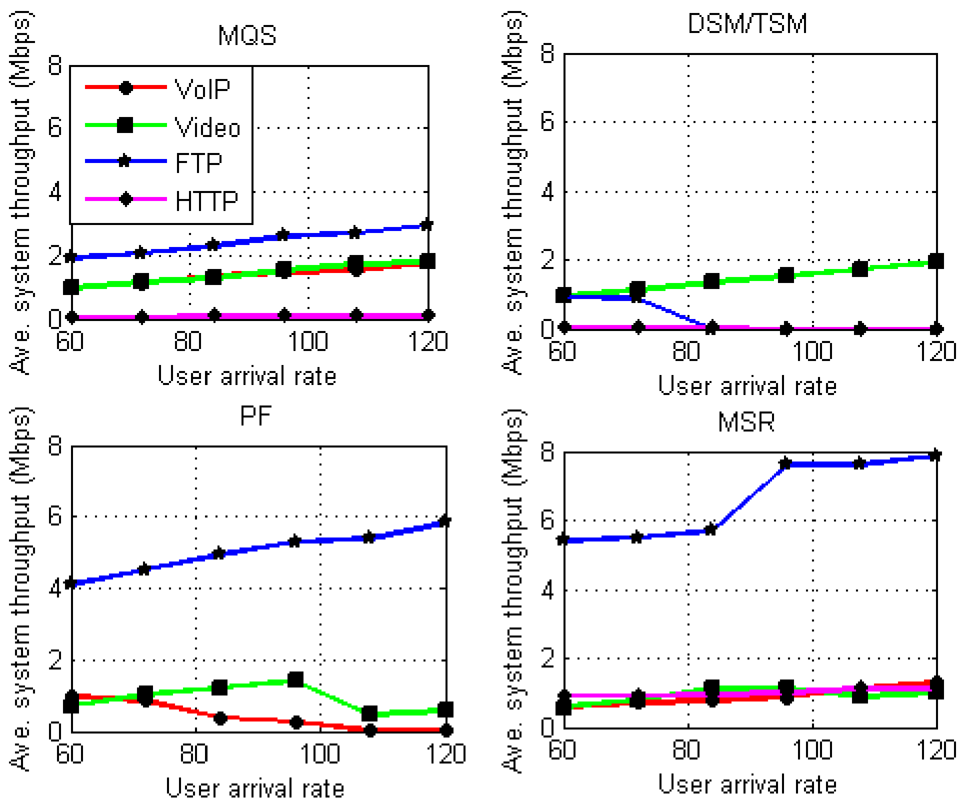


Figure 4.16: Average system throughput for a mixed VoIP, Video, FTP and HTTP services

User call satisfaction for a mixture of traffics involving VoIP, video streaming, FTP and HTTP is shown in Figure 4.17. The results show that both the MQS achieves a more balanced distribution of user call satisfaction; being able to do this in order of service priority. The DSM/TSM achieves almost 100% user satisfaction for VoIP and video streaming users. However, this comes at the expense of dis-satisfying both the FTP and HTTP users (having zero percent user satisfaction) above the traffic load of 84 and 96 users, respectively. This means that apart from DSM/TSM assigning higher user call satisfaction to HTTP than FTP; it can increase the churn rate for FTP and HTTP users who are largely unsatisfied as the user arrival rate increases in the cell. It is very clear from the figure that both the MSR and PF are unable to provide any QoS in heterogeneous mixed traffic.

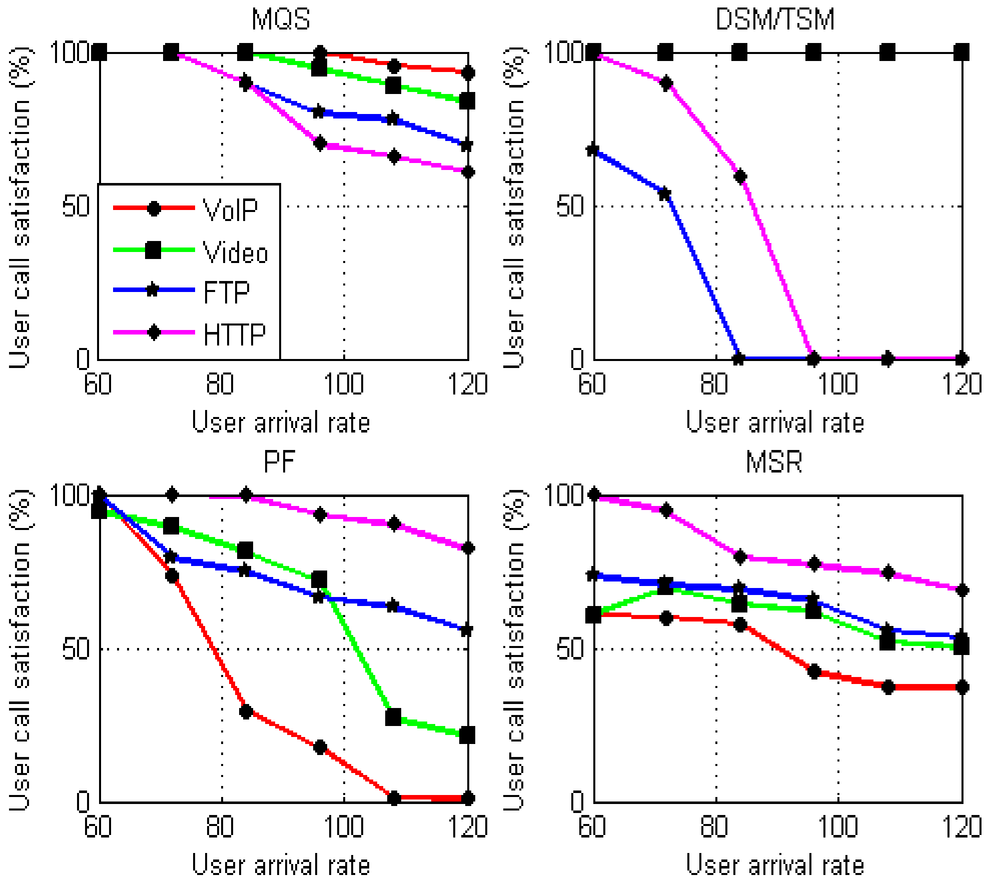


Figure 4.17: User call satisfaction for a mixed VoIP, Video, FTP and HTTP services

In Figure 4.18, the scheduling algorithms are compared for call fairness in scenario with a mixture of traffics involving VoIP, video streaming, FTP and HTTP. The figures show that MQS is the only scheduling algorithm that is able to allocate call fairness according to the traffic priority, i.e. higher call fairness are assigned to VoIP, Video, FTP and HTTP in that order. The DSM/TSM allocates higher priorities to VoIP and Video streaming accordingly. But, it also allocates higher call fairness to HTTP than FTP, both of which are eventually starved of resources above the traffic arrival of 84 and 96 users, respectively. Both the PF and MSR are shown to allocate call fairness without regard for service priority.

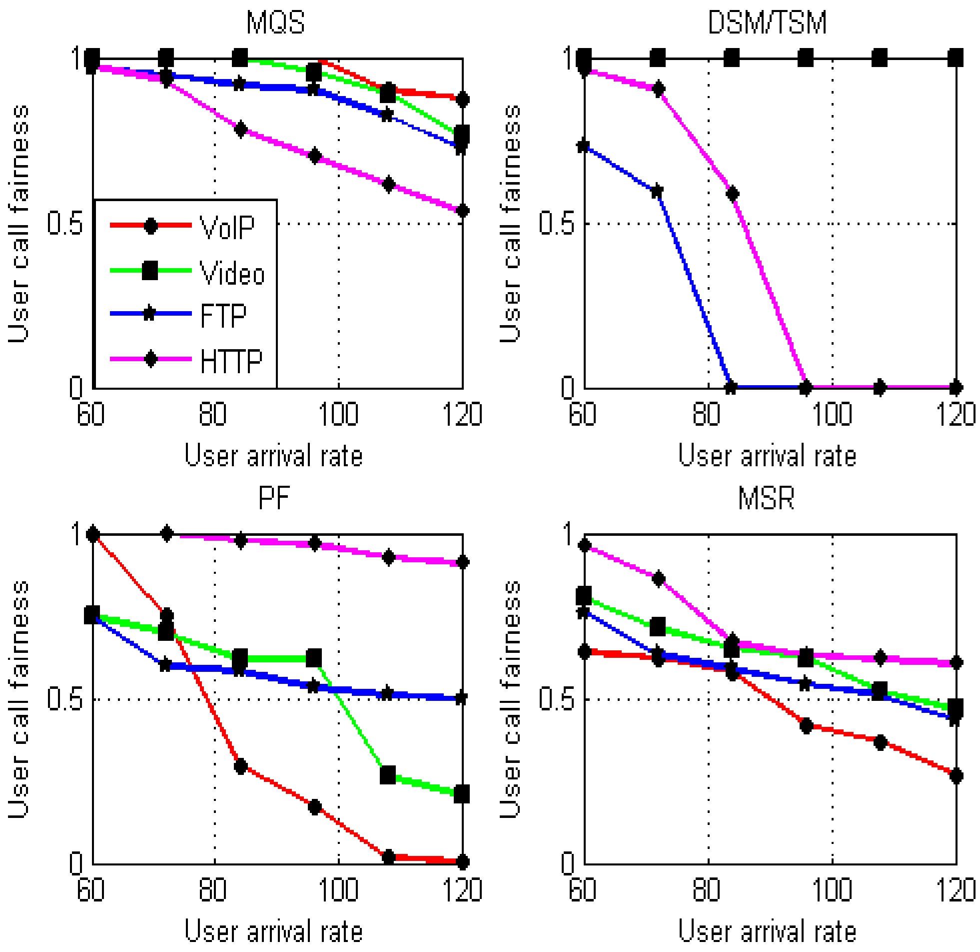


Figure 4.18: User call fairness for a mixed VoIP, video, FTP and HTTP services

4.4 Subcarrier Allocation in Relaying System

In this section and the average system throughput, user call satisfaction and call fairness are compared for two scheduling and resource allocation schemes: MQS-sc and DSM-sc (or DSM/TSM-sc) which applies the proposed scaling coefficient and MQS and DSM (or DSM/TSM) which use the conventional scaling coefficient, respectively, to determine the achievable data rate on each subcarrier. In the simulation, both the RT VoIP and video services are multiplexed using HD transmission.

Figure 4.19 shows that for VoIP, both the DSM-sc and MQS-sc which achieve almost the same average system throughputs are able to significantly increase the throughputs for the DSM/TSM and MQS, respectively, from arrival rates of 84 and 108 users. In video streaming, equal system throughputs are achieved by both the DSM-sc and MQS-sc; however the throughput increases over DSM and MQS are much more compared to the case of VoIP. It is very clear from the figure that the DSM-sc contributes higher gains for both VoIP and video users than in MQS-sc.

In Figure 4.20, the scheduling schemes are compared for user call satisfaction and as it can be seen both the MQS-sc and DSM-sc, achieving about 100% each, perform better than the MQS and DSM, respectively, in scenario with both RT VoIP and video traffic. However, the DSM-sc benefits better from the proposed resource allocation scheme. Figure 4.21 show that both the MQS-sc and DSM-sc, following similar pattern as in Figure 4.20, increases, respectively, the call fairness in RT VoIP and video services when compared to the MQS and DSM.

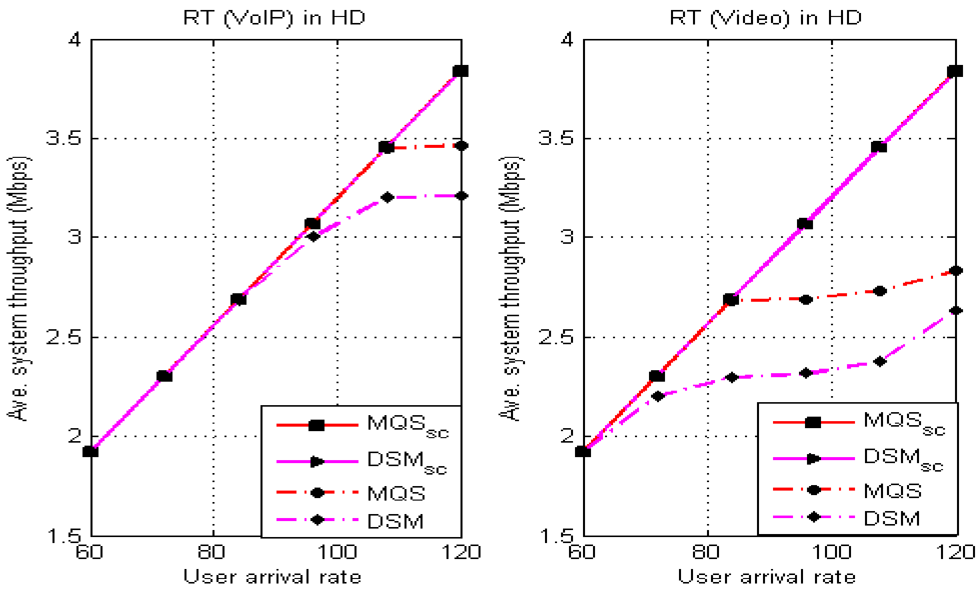


Figure 4.19: Average system throughput for multiplexed VoIP and Video services in HD mode

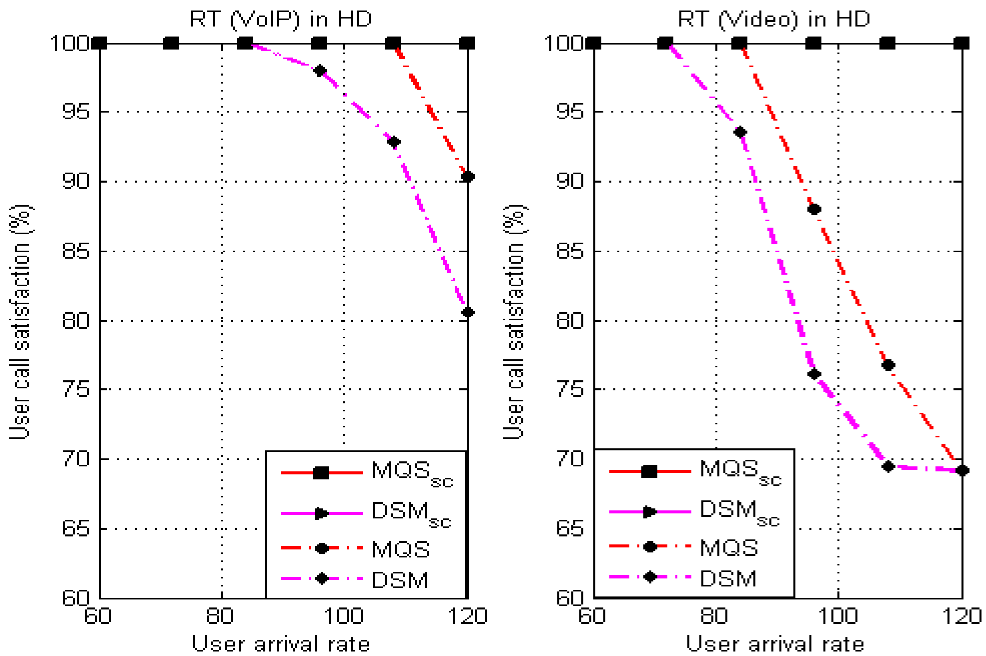


Figure 4.20: User call satisfaction for multiplexed VoIP and Video services in HD mode

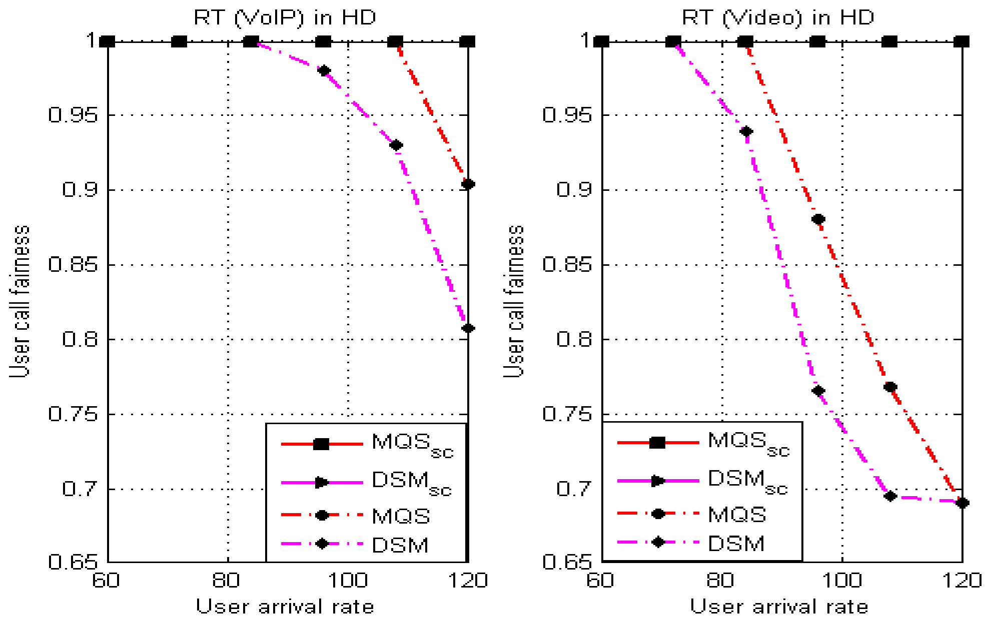


Figure 4.21: User call fairness for multiplexed VoIP and Video services in HD mode

In FD mode with RT VoIP and video services, as shown in Figure 4.22, both the MQS-sc and DSM-sc achieve higher average system throughputs compared to the MQS and DSM/TSM, respectively. In FD mode, the MQS-sc seems to achieve higher throughputs than DSM/TSM-sc, DSM-sc still provides higher throughput gains for VoIP and video. Figure 4.23 show that the MQS-sc and DSM-sc both achieve higher user satisfaction than MQS and DSM, respectively, in VoIP and video. However, in VoIP the user call satisfactions achieved by the different schemes are generally higher than that for video streaming. Almost similar pattern of user call satisfaction in Figure 4.23 is repeated in Figure 4.24 for call fairness. This confirms again that in RT when packets are dropped and traffic is not scheduled as a result of deadline miss, call fairness and user call satisfaction show almost similar pattern of results.

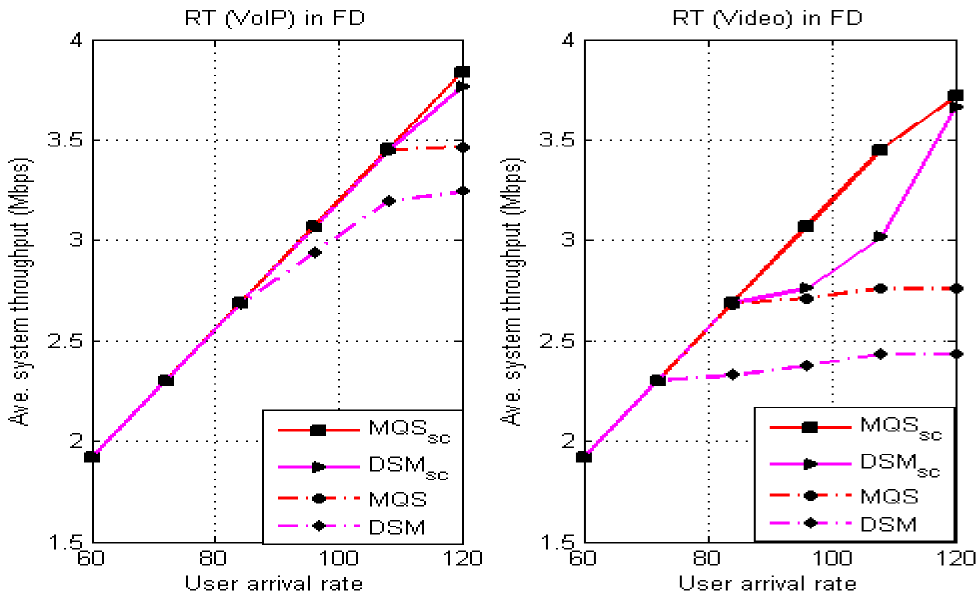


Figure 4.22: Average system throughput for multiplexed VoIP and Video services in FD mode

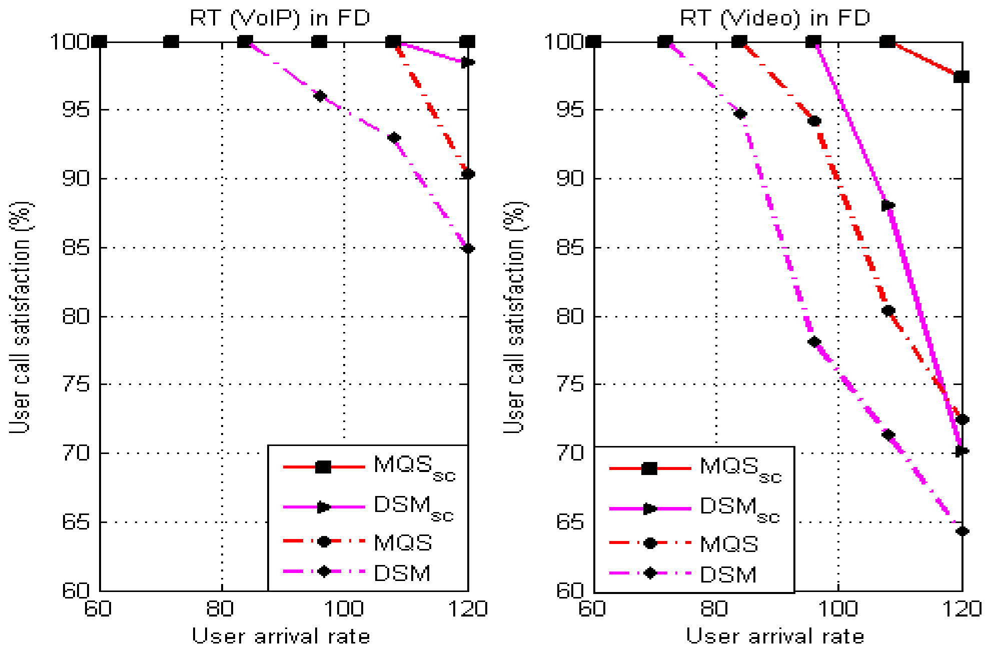


Figure 4.23: User call satisfaction for multiplexed VoIP and Video services in FD mode

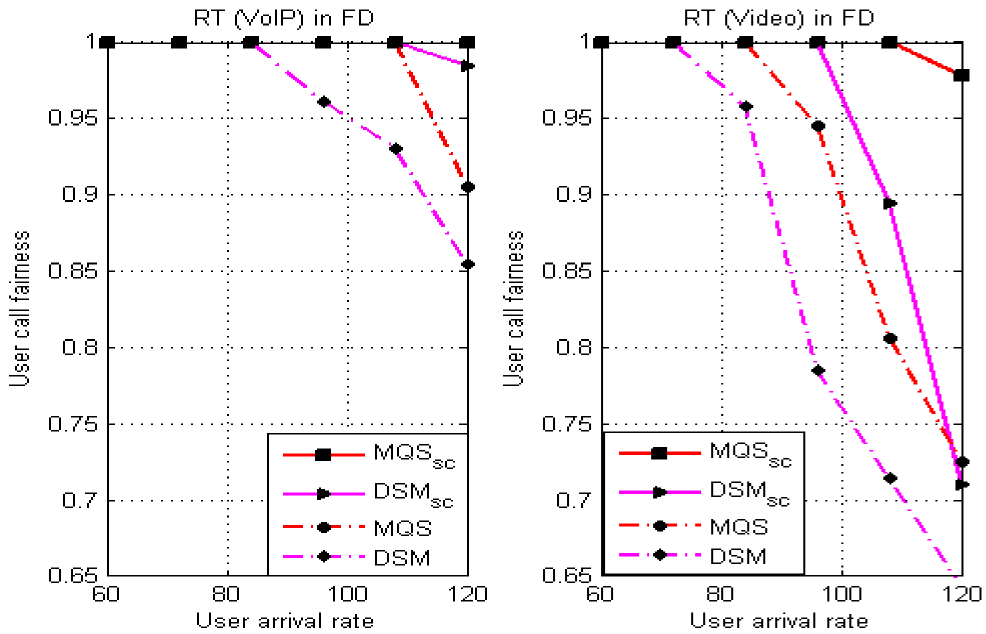


Figure 4.24: User call fairness for multiplexed VoIP and Video services in FD mode

When the system is operated in HD mode in multiplexed RT (video) and NRT (FTP) services and as shown in Figure 4.25, both the MQS-sc and DSM-sc are able to increase the average system throughput for video, although slightly the DSM/TSM-sc. However, in case of FTP, the MQS-sc and DSM-sc still show significant throughput gains; although both DSM-sc and DSM are seen to exhibit decreasing average system throughputs for increasing user arrival rates. This is expected because of the utility shapes of DSM as previously analyzed. Figure 4.26, compares the user call satisfaction and both the MQS-sc and DSM-sc are seen to provide higher user call satisfaction gains, respectively, over the MQS and DSM, respectively.

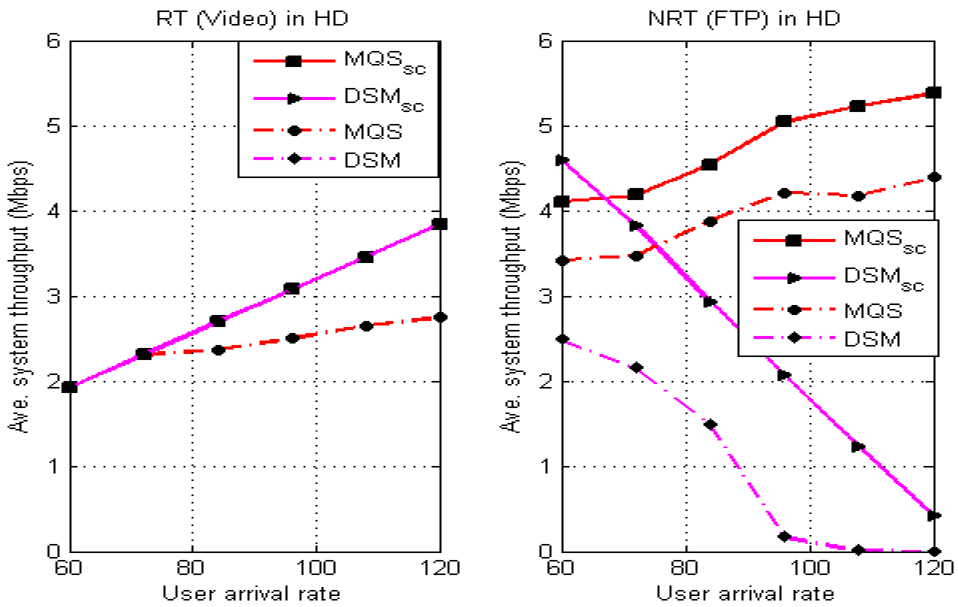


Figure 4.25: Average system throughput for multiplexed Video and FTP services in HD mode

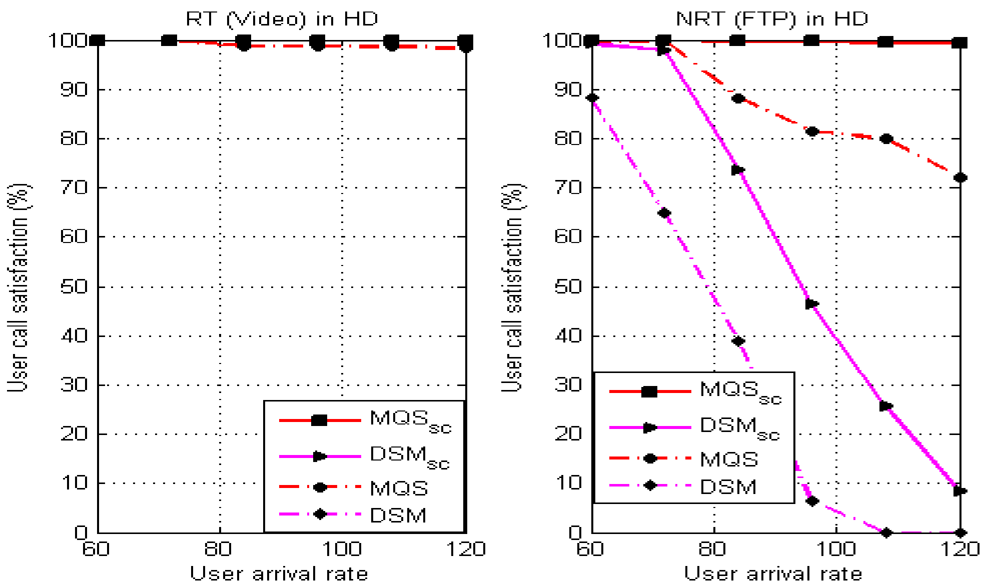


Figure 4.26: User call satisfaction for multiplexed Video and FTP services in HD mode

In Figure 4.27, the user call fairness achieved by both the DSM-sc and DSM are equally high. This is obviously achieved at the expense of FTP which shows much lower user call fairness compared to MQS-sc and MQS, respectively. However, it can be seen the proposed provide high fairness gains.

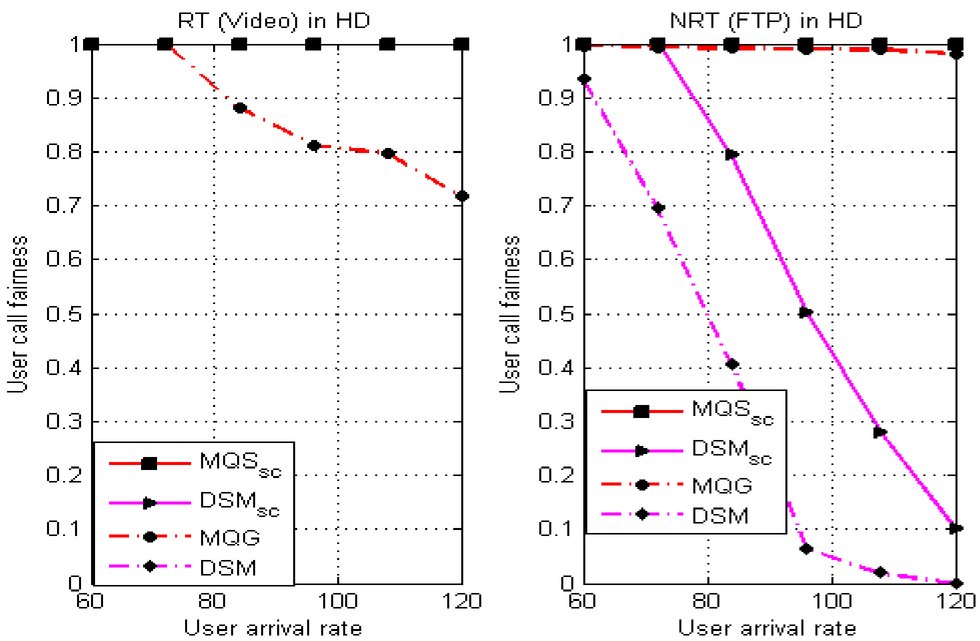


Figure 4.27: User call fairness for multiplexed Video and FTP services in HD mode

Operating in FD mode and as shown in Figure 4.28, both the MQS-sc and DSM-sc achieve higher throughput gains over the MQS and DSM, respectively. Similar patterns as in Figure 4.26 and Figure 4.27 are repeated in Figure 4.29 and Figure 4.30, where both user call satisfaction and call fairness gains seem to be higher in video than FTP for the MQS-sc than DSM-sc. Whereas, DSM-sc provides higher user call satisfaction and call fairness gains than MQS-sc in FTP than in video.

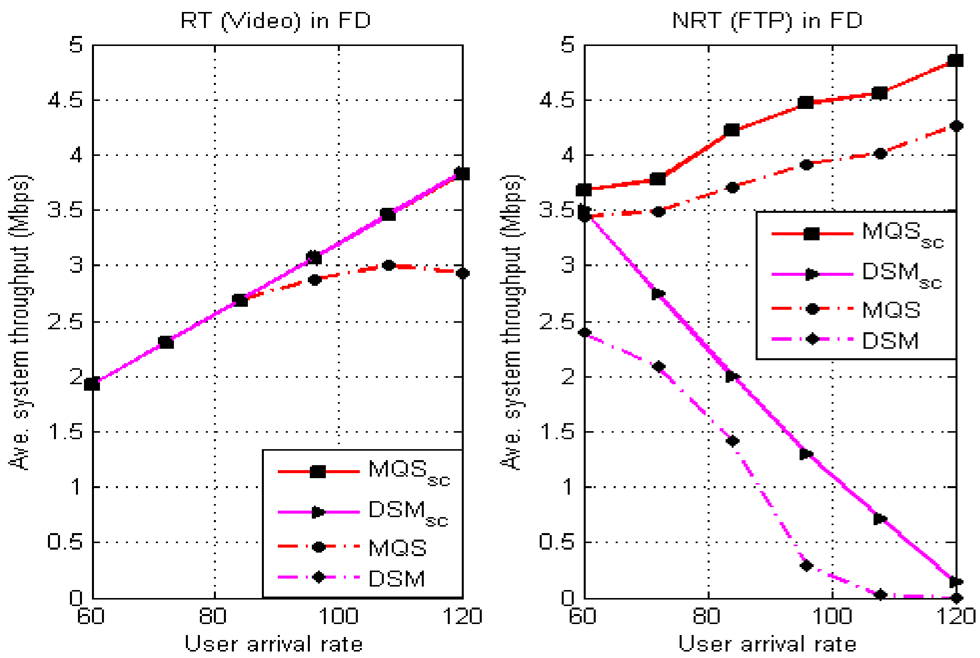


Figure 4.28: Average system throughput for multiplexed Video and FTP services in FD mode

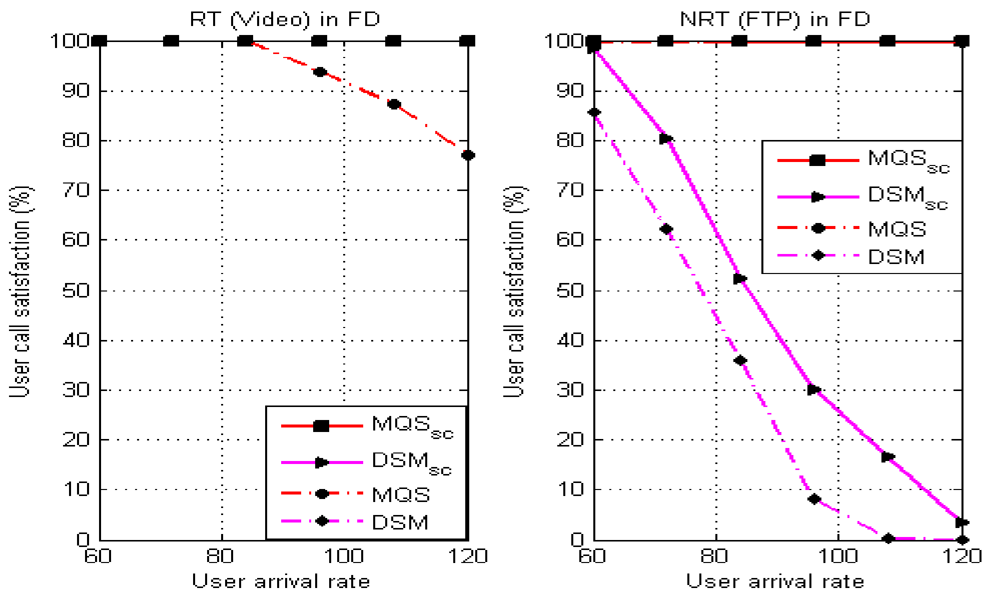


Figure 4.29: User call satisfaction for multiplexed Video and FTP services in FD mode

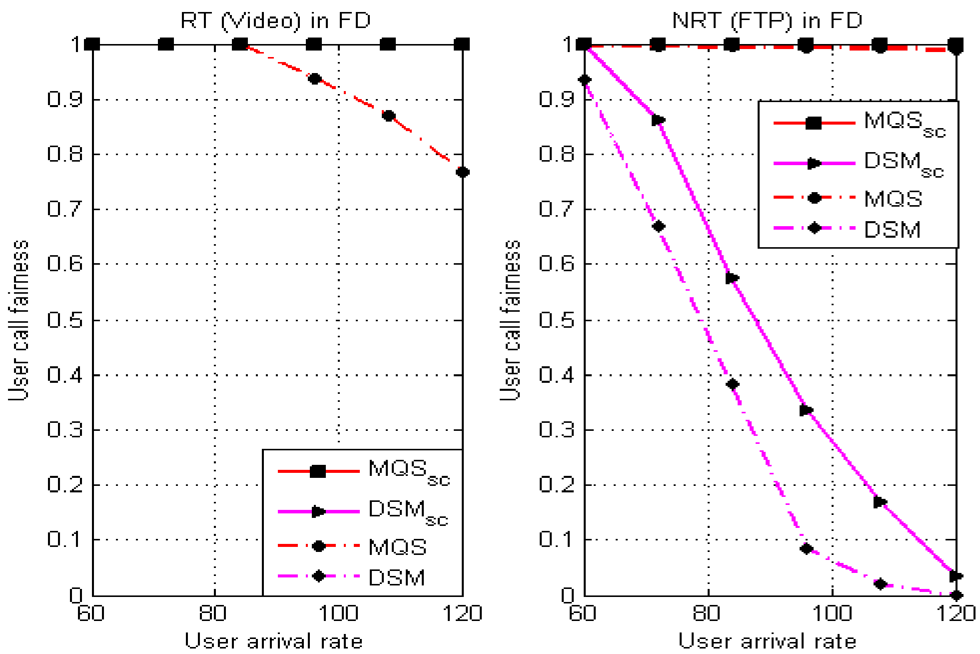


Figure 4.30: User call fairness for multiplexed Video and FTP services in FD mode

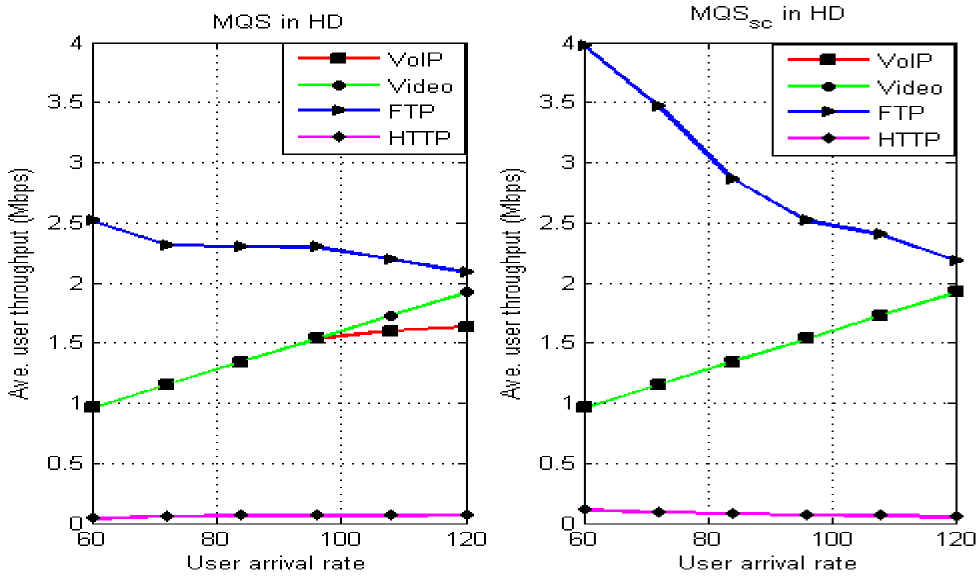
Multiplexing all the four services in HD mode, Figure 4.31a shows that MQS-sc is able to increase the all throughputs. However, the DSM-sc in Figure 4.31b only significantly improves the average system throughputs for FTP and HTTP; preventing HTTP from being starved at high user arrival rates.

In Figure 4.32a, both MQS-sc and MQS preserve the QoS priority in assigning user call satisfaction, while the performances for all the services are improved by MQS-sc. However, the DSM-sc increases the user call satisfaction for all the services and thus rescues both the FTP and HTTP users from being totally dis-satisfied beyond arrival rate of 102 users per timeslot, as shown in Figure 4.32b. However, it is still unable to properly distinguish the QoS classes. Figure 4.33a and Figure 4.33b in user call fairness achievement by both the MQS-sc and DSM-sc, respectively, surpass those of the MQS and DSM.

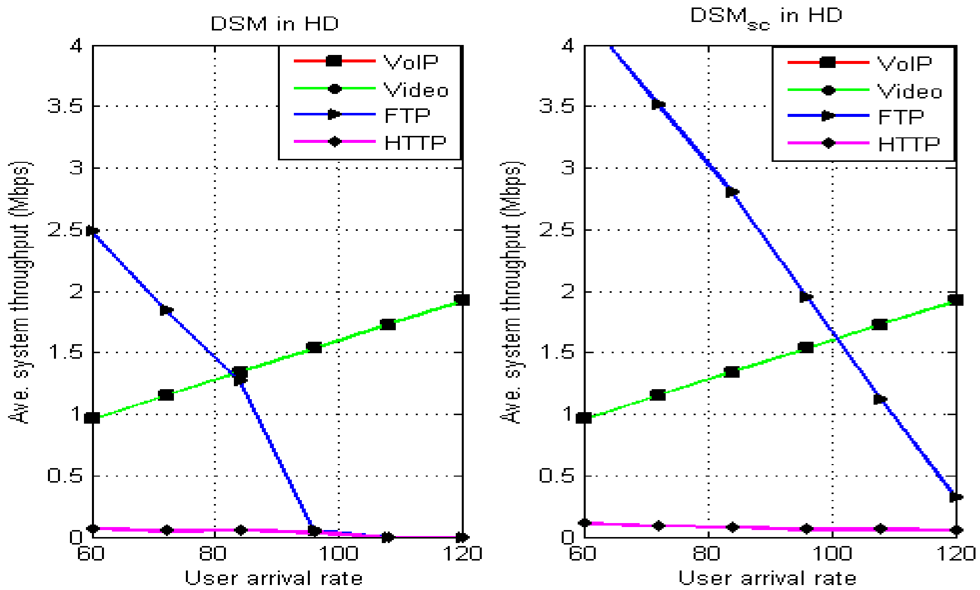
In Figure 4.34a and Figure 4.34b, both the MQS-sc and DSM-sc only improve the average system throughputs in FTP and HTTP services, the ones for VoIP and video are almost un-affected when the system is operated in FD mode.

In Figure 4.35a, the MQS-sc shows improved user call satisfaction while it maintains the QoS priority for all the services. In the case of DSM-sc, Figure 4.35b shows that the user call satisfaction for each of the services is improved; however, making sure that both the FTP and HTTP are made satisfied.

In Figure 4.36a and Figure 4.36b similar pattern of improvement as in Figure 4.34a and Figure 4.34b are shown for the user call fairness. In these figures, DSM-sc and DSM are depicted as being unable to assign utility according to service priority in heterogeneous mixed service scenario.

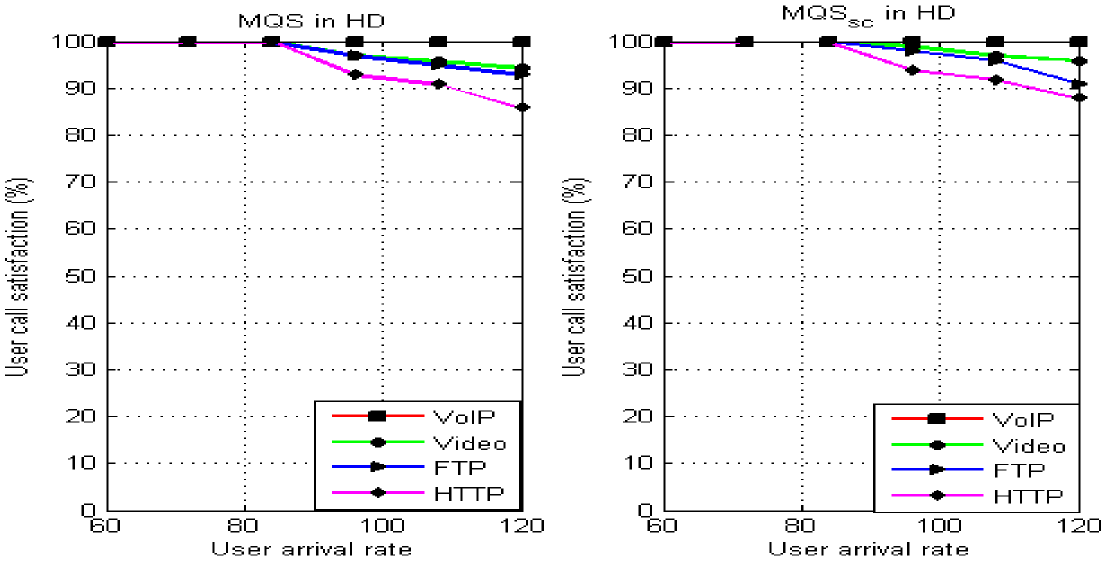


(a)

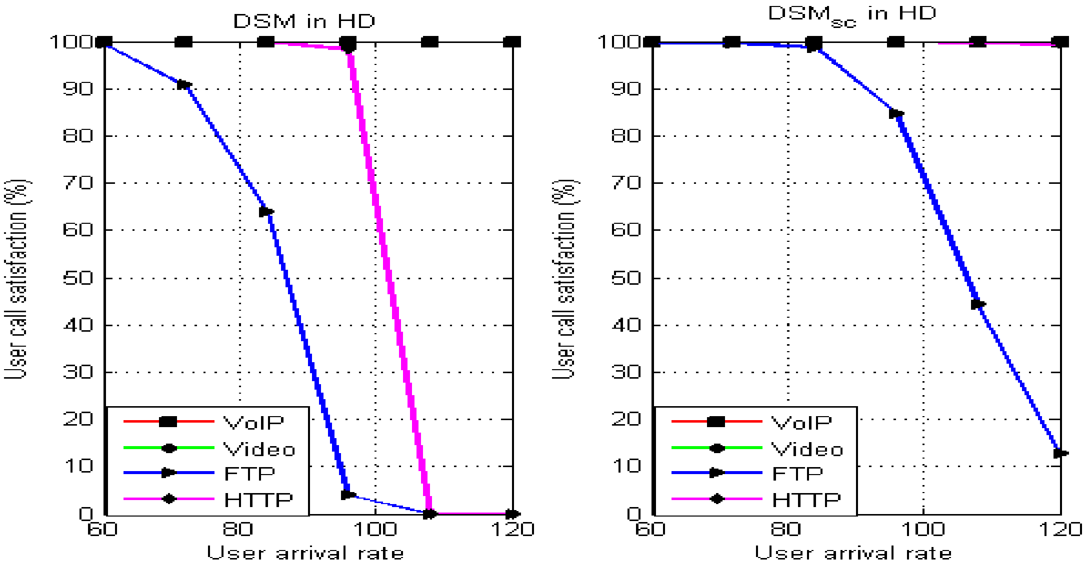


(b)

Figure 4.31: Average system throughput for multiplexed services in HD mode for (a) MQS and (b) DSM

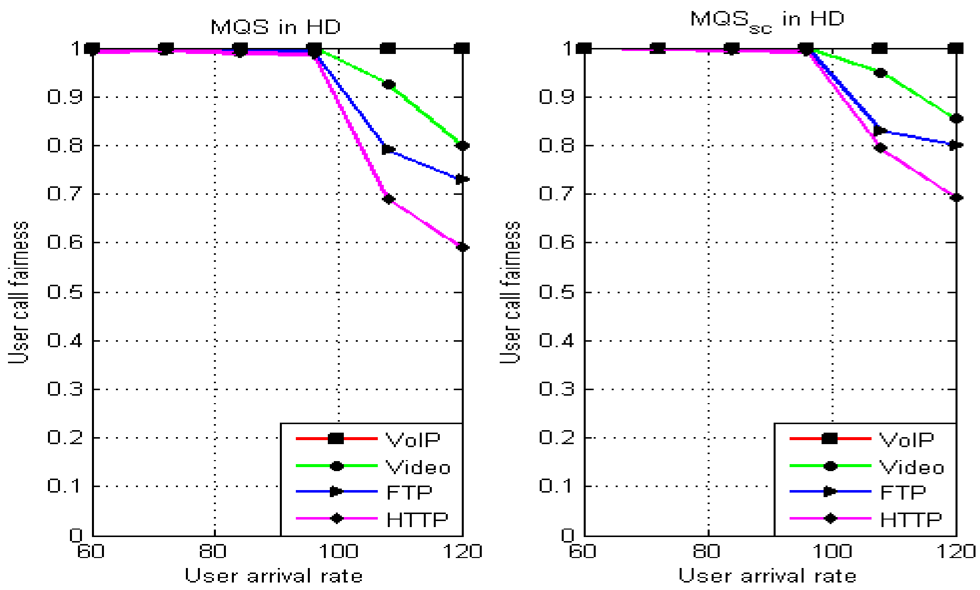


(a)

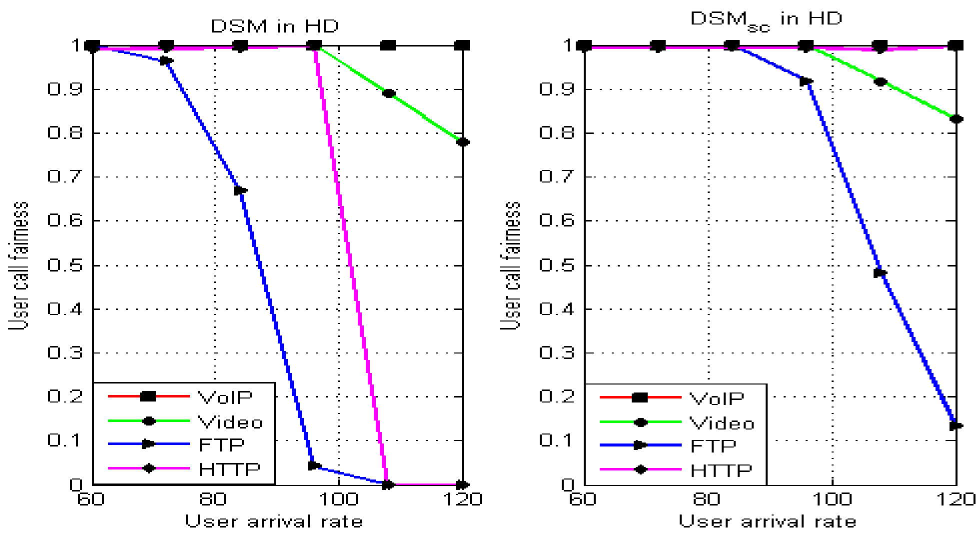


(b)

Figure 4.32: User call satisfaction for multiplexed services in HD mode for (a) MQS and (b) DSM

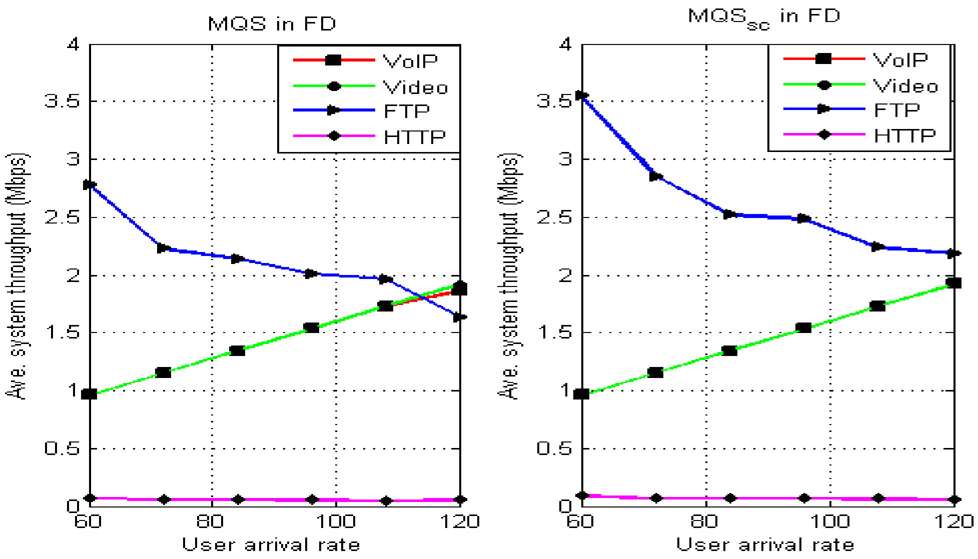


(a)

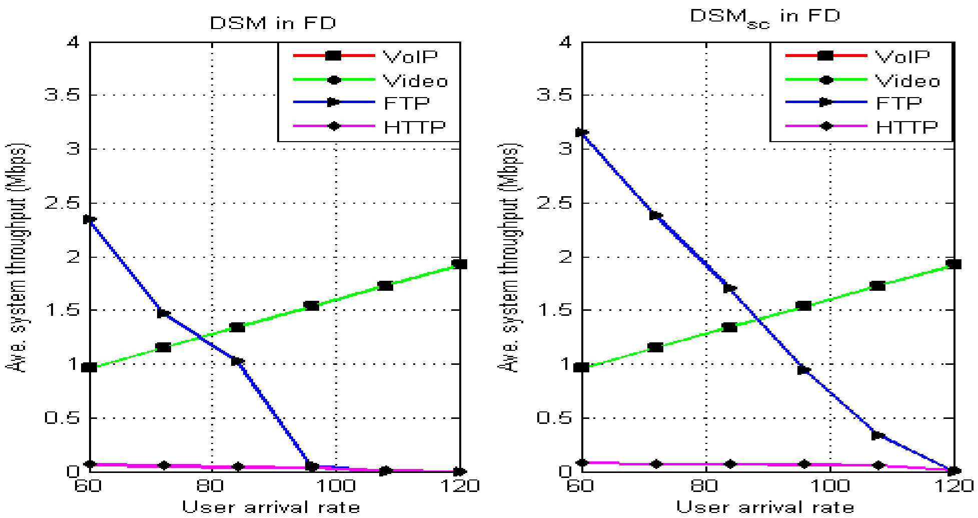


(b)

Figure 4.33: User call fairness for multiplexed services in HD mode for (a) MQS and (b) DSM

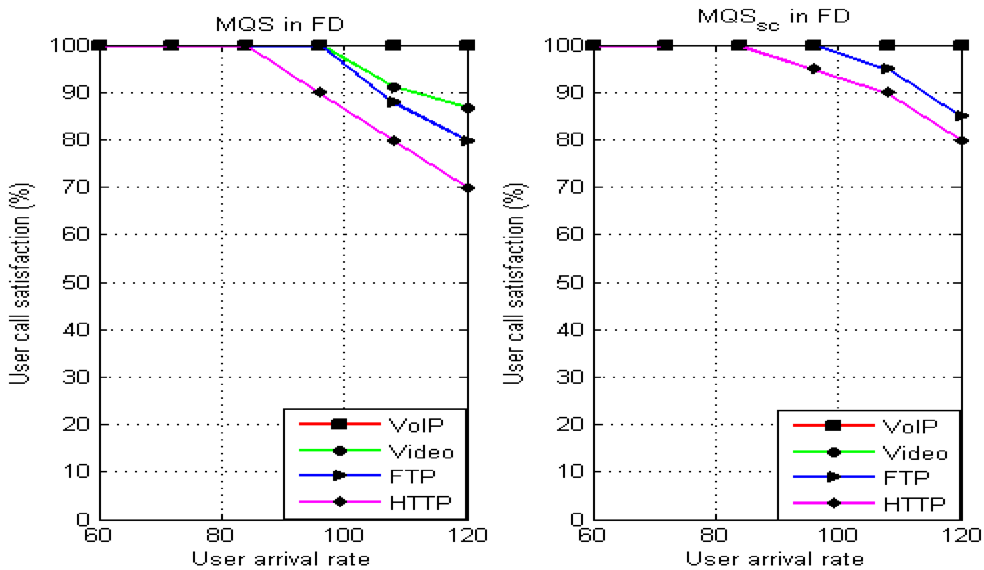


(a)

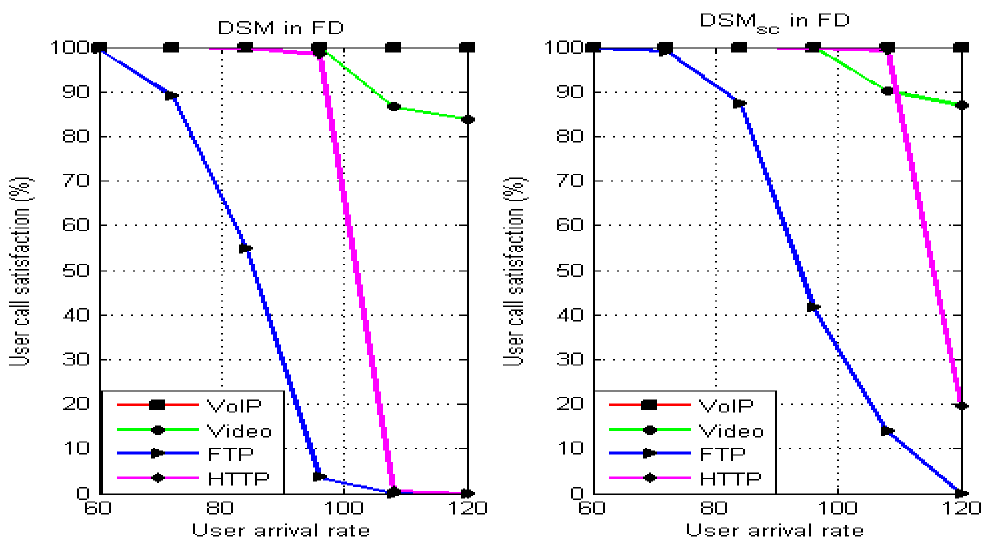


(b)

Figure 4.34: Average system throughput for multiplexed services in FD mode for (a) MQS and (b) DSM

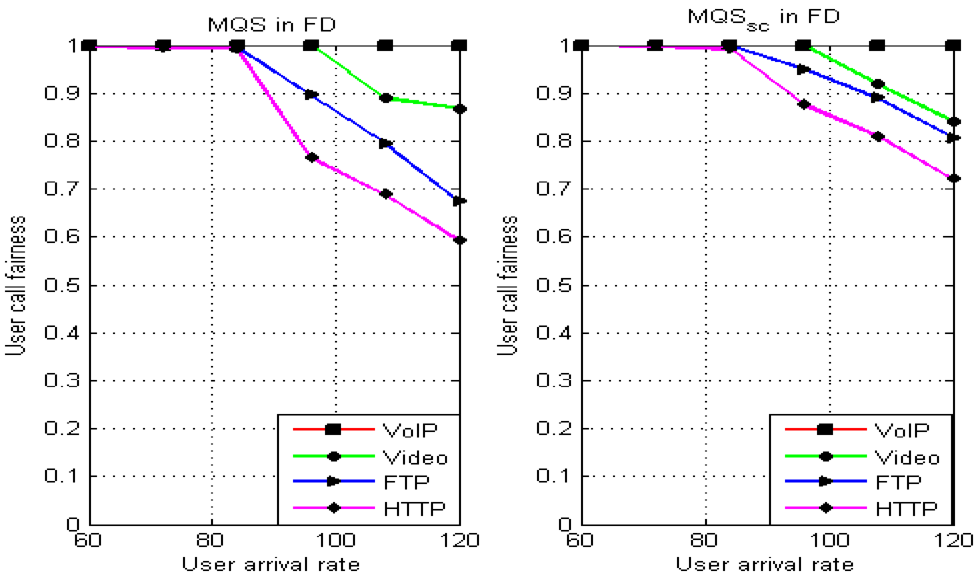


(a)

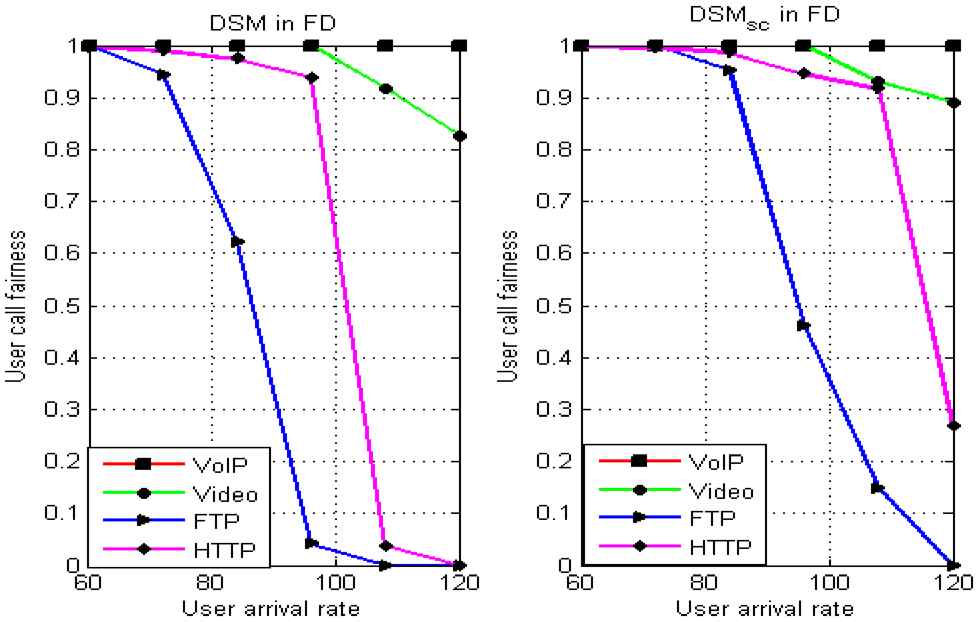


(b)

Figure 4.35: User call satisfaction for multiplexed services in FD mode for (a) MQS and (b) DSM



(a)



(b)

Figure 4.36: User call fairness for multiplexed services in FD mode for (a) MQS and (b) DSM

4.5 Chapter Summary

The performance evaluation for the different RRA algorithms under different scenarios supporting heterogeneous services has been presented. The performance comparison was done based on average system throughputs vs. user arrival rates, user call fairness vs. user arrival rates and user call satisfaction vs. user arrival rates. The simulation study shows that MQS achieves better average system throughputs than DSM/TSM, MSR and PF in that order both in RT and NRT traffic scheduling. For instance MQS achieves 0.8Mbps and 0.2Mbps higher in average system throughput at the maximum user arrival rate, respectively, in RT and NRT traffic compared with DSM/TSM. For the same traffics, the MQS achieve 8% and 2% higher in user call satisfaction and 0.02 and 0.2 higher in user call fairness indexes compared with DSM/TSM, all at the maximum user arrival rate. In the multiplexed VoIP, video, FTP and HTTP service scenario, MSR obtained 2.8Mbps, 52% and 0.35; PF achieved 2Mbps, 39% and 0.4; MQS achieved 1.7Mbps, 77% and 0.72; and DSM/TSM recorded 0.96Mbps, 54% and 0.52, respectively, in average system throughput, user call satisfaction and fairness. The analysis of the results also shows that the MQS provides the best QoS differentiation; as higher user call satisfaction and call fairness are allocated to VoIP, video streaming, FTP and HTTP in that order; which was not achievable by the other scheduling schemes with which it was compared.

Subcarrier allocation based on the proposed equalized AF scaling coefficient method used in scaling the backward SNR in the relaying link maximizes the user call satisfaction without sacrificing both the user call fairness and average system throughput as evaluated using MQS and DSM/TSM algorithms. The results, at the maximum user arrival rate, show that the MQS-sc achieves 1Mbps and 1Mbps higher in average system throughput, respectively, for Video and FTP when compared with MQS. It also achieves 2% and 29%

higher in user call satisfaction and 0.3 and 0.12 higher in user call fairness, respectively, for Video and FTP. The DSM-sc achieves 0Mbps and 0.5Mbps higher in average system throughput, respectively, for Video and FTP when compared with DSM. It also achieves 0% and 100% higher in user call satisfaction and 0.9 and 1.0 higher in user call fairness, respectively, for Video and FTP. It can be seen that the DSM-sc could only improve fairness for video users while maintaining same levels of user call satisfaction and average system throughput.

In the multiplexed services consisting of VoIP, video, FTP and HTTP operating in HD mode, the MQS achieved 1.42Mbps, 93% and 0.77 and, respectively, in average system throughput, user call satisfaction and fairness. However, the MQS-sc increased the throughput, user call satisfaction and fairness to 1.52Mbps, 95% and 0.84, respectively. In the case of DSM, the DSM_sc increased the throughput, user call satisfaction and fairness from 0.96Mbps, 50% and 0.44 to 1.1Mbps, 78% and 0.44, respectively. In the next chapter, the concluding remarks are presented and future works that can be done to enhance this work is presented

Chapter 5

Conclusions and Recommendations

5.1 Overview

This research has addressed one of the most essential aspects of wireless system which is radio resource allocation (RRA). The success of wireless systems in supporting multimedia services in high-data-rate communications is strongly tied to the performance of scheduling schemes. The originality of the work is the development of modified sigmoidal-type utility functions for scheduling multiplexed RT, NRT and BE services in a wireless system. The study also introduced a new method of subcarrier allocation in relaying using a relay station parameter.

This chapter presents a summary based on the discoveries and discussions earlier in chapter four. The chapter also describes the significances of the study and provides helpful information for researchers in this area. This chapter is arranged into the following subheadings: conclusion, research contribution and practical applications, limitations of the study and direction for future works.

5.2 Conclusion

The convergence of mobile and internet data has complicated the management of the scarce resources to be shared among network users. High efficient air interfaces that can support high data rates with high flexibilities are required. OFDMA is one of the preferred high performance physical layer air interfaces for next generation broadband wireless communication systems. Indeed, it has been adopted by several 4G standards including

WiMAX and LTE-A. However, efficient scheduling and resource allocation techniques are crucial in utilizing the resources and flexibilities offered by the access technologies and in particular OFDMA. Previous work used opportunistic policies to maximize the system efficiency through allocating the resources to only those users who maximize the system capacity. Although it optimized system efficiency, but its disadvantage was that it was unfair to users who could not maximize system efficiency; because they had poor channel conditions. This then called for scheduling schemes that can balance the trade-off between fairness in resource distribution and efficiency in resource usage. However, fairness is no longer a sufficient performance target when multimedia traffic with diverse QoS requirements has to be simultaneously handled in the same network. Therefore, to support these various services it has become inevitable for network operators to guarantee the satisfactory provision of the quality of services (QoS) in wireless links.

The objective of our study was to design a utility-based scheduling scheme that not only balances the trade-off between efficiency in resource usage and fairness in resource distribution; but which also assures QoS guarantee among network users. Three utility-based scheduling frameworks, called maximum QoS satisfaction (MQS), are developed: one for delay-sensitive RT traffic, another for throughput-sensitive NRT traffic and the third for BE traffic without QoS requirement. The delay-based utility function uses an increasing sigmoidal function based on HOL packet delay with inflection point in the users' users' HOL packet delay requirement, which is usually equal to the RT delay budget of the system. The throughput-based utility functions for NRT and BE services use a decreasing sigmoidal marginal utility function based on throughput with inflection point in the users' throughput requirement, respectively. However, for BE users' throughput requirement is zero.

The MQS was compared with MSR, PF and DSM/TSM scheduling schemes which have different optimization objectives: MSR only maximizes the system capacity without consideration for fairness, PF trades off system efficiency for fairness and DSM/TSM using a similar type of sigmoidal utility functions has the objective to maximize user call satisfaction in NRT or RT services. In MQS the objective is to maximize user call satisfaction in heterogeneous mixed RT and NRT services. Therefore, it is necessary to compare MQS with DSM/TSM, with which it has a similar design objective; and with PF and MSR, which are rate-based-only, to establish their suitability or otherwise in being used for QoS provisioning. The analysis of the results, in the multiplexed VoIP, video, FTP and HTTP services, show that the MSR algorithm maximizes the efficiency when it obtained the highest average system throughput, but which results in the lowest fairness among users. The PF algorithm compromises between efficiency and fairness in resource distribution, thus brings down the average system throughput in exchange for the higher call fairness compared with MSR. The MQS sacrificed the average system throughput to achieve the highest user call satisfaction and fairness. The DSM/TSM while also trading off the average system throughput for higher user fairness and user satisfaction fairness compared with MSR and PF, exhibited a situation whereby network resources were not allocated according to service priority.

Based on these results, it is concluded that a relationship do exist between system efficiency (throughput) in resource usage and fairness in resource distribution; and between fairness in resource distribution and user satisfaction (QoS). It is also observed that fairness alone (PF) without taking into account the delay requirements of RT users and rate requirements of NRT users does not guarantee that users are allocated resources according to their service priority.

Motivated by the need to provide efficient trade-off between system capacity and resource allocation fairness while assuring high QoS in relaying systems; studies were carried out on current approach to resource allocation. It is identified that the AF scaling coefficient as a potential relay station parameter can be modified within the objective constraints to provide distributed frequency multiplexing gains across subcarriers. Therefore, the AF scaling coefficients are derived and equalized for HD and FD systems based on the gain factor of the VG relay protocol. This is then applied to allocate resources. Two scheduling schemes are used to analyze and compare the performances for the scheme when conventional scaling coefficient is used as in MQS and DSM; and for the scheme using equalized scaling coefficient as in MQS-sc and DSM-sc.

The analysis of the results, in the multiplexed VoIP, video, FTP and HTTP services, shows that the MQS-sc increased the average system throughput, user call satisfaction and fairness of MQS. In similar manner, the DSM-sc increased the average system throughput, user call satisfaction and fairness of the DSM. In multiplexed Video and FTP services, the MQS-sc also achieves higher average system throughput, user call satisfaction and fairness compared with MQS while the DSM-sc was able to tremendously rescue the video users from resource starvation caused by the DSM. It is concluded, from these results, that the network operators can apply the equalized scaling coefficient technique in MQS and DSM/TSM scheduling algorithms to improve satisfaction and fairness levels among users, while increasing the average system capacity.

5.3 Contribution of the Study

The study contributes significantly by providing three modified sigmoidal-type utility functions for real-time, non-real-time and best- effort services, respectively, to maximize user satisfaction while providing a more efficient trade-off between average system throughput and user call fairness in heterogeneous mixed services compared with the benchmarks. Guaranteeing user satisfactions allow networks operators maintain a high number of subscribers, decrease churn, and attract new subscribers. Therefore, the developed utility-based frameworks can indeed be used not only to bridge the gap between system efficiency and fairness but to provide satisfaction among network users in the base stations of a wireless systems based on WiMAX and LTE-A standards.

The study also contributes immensely by developing an equalized scaling coefficient which is applied in subcarrier allocation to maximize user QoS at the same time that it improves both the system capacity and call fairness. Therefore, relay networks based on WiMAX and LTE-A standards can use the equalized scaling coefficients as a parameter in the radio resource management application to guarantee higher resource allocation fairness and satisfaction among users without compromising the system efficiency.

Some of the contributions of the thesis described above have been published or accepted for publication in different journals and international conferences. The complete list of publications associated with this thesis work is presented in Appendix A.

5.4 Limitation of the Study and Direction for Future Research

This thesis addressed resource allocation problems in OFDMA-based wireless systems in which only RT users' packets are dropped when they exceed their maximum packet delays (deadlines); connection admission control (CAC) is not considered. An efficient CAC involves a complete suspension of service and/or dropping of users if it becomes infeasible to provide the desired QoS to all users; so as not to degrade the system's efficiency. This approach provides the actual QoS guarantee rather than the statistical QoS assurance/guarantee considered in our implementation is recommended. For the resource allocation problem in the relaying systems, only a single relay was considered, the performance analyses can be extended to multiple relays to provide greater multi-user and spatial diversities. The equalized scaling coefficient used to solve the resource allocation problem was based on VG-AF relaying protocol Other AF relay protocols such CG, UG and FG are recommended for further investigation. The cell in the network model used for relaying was divided into two regions: RS and BS. In the implementation, only the users in the RS region were considered. Practical networks include users in the BS region as well; therefore, it is recommended that it is considered in future works. The solution of the optimization problem in this work was based on equal power allocation across all subcarriers; un-equal power allocation shall be considered in future works. To evaluate the performances of the scheduling and subcarrier allocation schemes, the simulation was performed over a duration corresponding to a simulation time of 20s; this could be higher to ensure better capturing of the dynamics of the system in the time domain as well as the traffic pattern of the users.

References

Abrar, M. , Gui, X. & Punchihewa, A.L (2012). Maximizing System Throughput in Multi-User Cooperative Relay Networks. IEEE Xplore. pp. 3310-3314.

Agarwal, R., Majjigi, V., Vannithamby, R. & Cioffi, J. (2007). Efficient Scheduling for Heterogeneous Services in OFDMA Downlink. In Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM), pp. 3235–3239.

Alasti, M., Neekzad, B., Hui, J. & Vannithamby, R. (2010). Quality of Service in WiMAX and LTE Networks. IEEE Communications Magazine. 48(5): 104-111.

Ali, S. & Zeeshan, M. (2012). A Utility Based Resource Allocation Scheme with Delay Scheduler for LTE Service-Class Support. IEEE Wireless Communication and Networking Conference: MAC and Cross-layer Design. pp. 1450-1455.

Ali, S.H, Lee, K-D. & Leung, V. C. M. (2007). Dynamic Resource Allocation in OFDMA Wireless Metropolitan Area Networks. IEEE Wireless Communications. 14(1): 6-13.

Al-Manthari, B., Hassanein, H., Ali, N.A & Nasser, N. (2009). Fair Class-Based Downlink Scheduling with Revenue Considerations in Next Generation Broadband Wireless Access Systems. IEEE Transactions On Mobile Computing. 8(6): 721-734.

Alouini, M. S. & Goldsmith, A. J. (2000). Adaptive Modulation over Nakagami Fading Channels. Kluwer Journal of Wireless Communications. 13(1–2): 119–143.

Alvi S.H. & Wyne, S. (2014). On Amplify-and-Forward Relaying over Hyper-Rayleigh Fading Channels. Radio Engineering. 23(4): 1226-1233.

Andrews, J.G., Ghosh, A. & Muhamed, R. (2007). Fundamentals of WiMAX Understanding Broadband Wireless Networking. New Jersey: Prentice Hall. pp. 5-83.

Andrews, M. (2004). Instability of the Proportional Fair Scheduling Algorithm for HDR. IEEE Transactions on Wireless Communications. 3(5): 1422-1426.

Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A., Whiting, P. & Vijayakumar, R. (2000). CDMA Data Qos Scheduling on the Forward Link with Variable Channel Conditions. Bell Labaratories, Lucent Technologies.

Andrews, M., Qian, L. & Stolyar, A. (2005). Optimal Utility-Based Multi-User Throughput Allocation Subject to Throughput Constraints. In Proceedings of the 24th IEEE Annual Joint Conference of the Computer and Communications Societies, pp. 2415–2424.

Andrews, M., Stolyar, A., Kumaran, K., Vijayakumar, R., Ramanan, K. & Whiting, P. (2001). Providing Quality of Service over a Shared Wireless Link. IEEE Communications Magazine. 39(2): 150-154.

Anghel, P.A & Kaveh, M. (2004). On the Diversity of Cooperative Systems. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 4: 577-580.

Balakrishnan, R. & Canberk, B. (2014). Traffic-Aware QoS Provisioning and Admission Control in OFDMA Hybrid Small Cells. IEEE Transaction on Vehicular Technology. 63(2): 802-810.

Beres, E. & Adve, R. (2008). Selection Cooperation in Multi-Source Cooperative Networks. IEEE Transactions on Wireless Communications. 7(1): 118–127.

Berger, S. & Wittneben, A. (2005). Cooperative Distributed Multiuser MMSE Relaying in Wireless Ad-Hoc Networks. In Proceedings of the 39th Asilomar Conference on Signals, Systems, and Computers, pp. 1072-1076.

Bi, S. & Zhang, Y.J. (2010). TDMA Achieves the Same Diversity Gain as Opportunistic Scheduling in Relay-Assisted Wireless Networks. In Proceedings of the IEEE Global Telecommunication Conference (GLOCOM), pp. 1-6.

Bletsas, A., Khisti, A., Reed, D.P. & Lippman, A. (2006). A simple cooperative diversity method based on network path selection. IEEE Journal on Selected Areas in Communications. 24: 659–672.

Boyer, J., Falconer, D. & Yanikomeroglu, H. (2004). Multihop Diversity in Wireless Relaying Channels. IEEE Transactions on Communications. 52(10): 1820–1830.

Broadband Wireless Access Networks. IEEE Transactions on Mobile Computing. 9(2): 259–269.

Chen, L., Wang, B., Chen, X., Zhang, X. & Yang, D. (2011). Utility-Based Resource Allocation for Mixed Traffic in Wireless Networks. In Proceedings of the IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM) Workshop, pp. 91–96.

Cheng, W., Zhang, X. & Zhang, H. (2012). Full/Half Duplex Based Resource Allocations for Statistical Quality of Service Provisioning in Wireless Relay Networks. In Proceedings of the IEEE INFOCOM, pp. 864-872.

Chia, S., Gill, T., Ibbetson, L., Lister, D., Pollard, A., Irmer, R., Almodovar, D., Holmes, N. & Pike, S. (2008). 3G Evolution. IEEE Microwave Magazine. 9(4): 52–63.

Cho, J.-W. & Chong, S. (2005). Utility Max-Min Flow Control Using Slope-Restricted Utility Functions. In Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM), 2: 819-824.

Choi, K.W., Jeon, W.S. & Jeong, D.G. (2009). Resource Allocation in OFDMA Wireless Communications Systems Supporting Multimedia Services. IEEE/ACM Transactions on Networking. 17(3): 926 - 935.

Chung, S. T. & Goldsmith, A. J. (2001). Degrees of Freedom in Adaptive Modulation: A unified View. IEEE Transactions on Wireless Communications. 49(9): 1561–1571.

Dahlman, E., Parkvall, S. & Skold, J. (2013). 4G LTE/LTE-Advanced for Mobile Broadband, 2nd Ed. Cambridge: Academic Press. pp. 20-80.

Dañobeitia, B. & Femenias, G. (2010). Dual Methods for Channel- and Qos-Aware Resource Allocation in MIMO-OFDMA Networks. IFIP Wireless Days (WD). pp. 1-5.

Dañobeitia, B. & Femenias, G. (2011). Cross-Layer Scheduling and Resource Allocation in OFDMA Wireless Networks. IFIP Wireless Days (WD), pp. 1-6.

Dhrona, P., Ali, N.A. & Hassanein, H. (2009). A Performance Study of Uplink Scheduling Algorithms in Point-To-Multipoint Wimax Networks. Computer Communications. 23 (3): 511-521.

Duval, O., Hasan, Z., Hossain, E., Gagnon, F. & Bhargava, V. K. (2010). Subcarrier Selection and Power Allocation for Amplify-and-Forward Relaying over OFDM Links," IEEE Transactions on Wireless Communications. 9(4): 1293 1297.

Emamian, V., Anghel, P. & Kaveh, M. (2002). Multi-User Spatial Diversity in a Shadow-Fading Environment. In Proceedings of the IEEE Vehicular Technology Conference, 1: 573-576.

Emmanouil, M., Per, Z., Christos, M. & Aris, S.L. (2009). Experimental Investigation of Cooperative Schemes on a Real-Time DSP-Based Testbed. EURASIP Journal on Wireless Communications and Networking. 1(956):15.

Enderle, N. & Lagrange, X. (2003). User Satisfaction Models and Scheduling Algorithms for Packet-Switched Services in UMTS. In Proceedings of the 57th IEEE Vehicular Technology Conference, pp. 1704–1709.

Erceg, V., Greenstein, L.J., Tjandra, S.Y., Parkoff, S.R., Gupta, A., Kulic, B., Julius, A.A & Bianchi, R. (1999). An Empirically Based Path Loss Model for Wireless Channels in Suburban Environments. IEEE Journal on Selected Areas in Communications. 17(7): 1205-1211.

Ermini, D. (2011). Traffic Scheduling in Point-to-Multipoint OFDMA-Based Systems (Doctoral Thesis), Universita Di Pisa, Italy.

Eryilmaz, A. & Srikant, R. (2007). Fair Resource Allocation in Wireless Networks using Queue-Length-Based Scheduling and Congestion Control. IEEE/ACM Transaction on Networks. 15(6): 1333–1344.

Fan, Y., Poor, H. V. & Thompson, J. S. (2008). Cooperative Multiplexing In Full-Duplex Multi-Antenna Relay Networks. In Proceedings of the IEEE Global Telecommunication. Conference, pp. 1–5.

Fareed, M.M. & Uysal, M. (2009). On Relay Selection for Decode-and-Forward Relaying. IEEE Transactions on Wireless Communications. 8: 3341–3346.

Femenias, G., Danobeitia, B. & Riera-Palou, F. (2012). Unified Approach to Cross-Layer Scheduling and Resource Allocation in OFDMA Wireless Networks. EURASIP Journal of Wireless Communications and Networking. 2012(1): 1-19.

Forkel, I., Klenner, H. & Kemper, A. (2005). High Speed Downlink Packet Access HSDPA- Enhanced Data Rates for UMTS Evolution. Elsevier Computer Networks. 49 (3): 325-340.

Fu,I., Chen, Y., Cheng, P., Yuk, Y. , Kim, Y. & Kwak, J. (2010). Multicarrier technology for 4G WiMAX system. IEEE Communication Magazine. 48(8): 50–58.

Ganz, A. & Wongthavarawat, K. (2003). Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems. International Journal of Communication Systems. 16(1): 81 – 96.

Girici, T., Zhu, C., Agre, J.R. & Ephremides, A. (2010). Proportional Fair Scheduling Algorithm in Ofdma-Based Wireless Systems with QoS Constraints. Journal of Communications and Networks. 12(1): 30-42.

Goldsmith, A. J. (1997). The Capacity of Downlink Fading Channels with Variable Rate and Power. IEEE Transactions on Vehicular Technology. 46(3): 569–580.

Goldsmith, A. J. & Chua, S.-G. (1997). Variable-Rate Variable-Power MQAM for Fading Channels. IEEE Transactions on Communications. 45(10): 1218–1230.

Goldsmith, A. J. & Chua, S.-G. (1998). Adaptive Coded Modulation for Fading Channels. IEEE Transactions on Communications. 46(5): 595–602.

Goldsmith, A., Jafar, S., Jindal, N. & Vishwanath, S. (2003). Capacity Limits of MIMO Channels. IEEE Journal on Selected Areas in Communications. 21(5): 684–702.

Hajjawi, A. & Ismail, M. (2015). QoS-based Scheduling Algorithm for LTE-A. Research Journal of Applied Sciences, Engineering and Technology. 11(8): 854-859.

Han, Z., Ji, Z. & Liu, K. (2005). Fair Multiuser Channel Allocation for OFDMA Networks using Nash Bargaining Solutions and Coalitions. IEEE Transactions on Communications. 53(8): 1366–1376.

Hasan, Z., Hossain, E. & Bhargava, V. (2011). Resource Allocation for Multiuser OFDMA-Based Amplify-and-Forward Relay Networks with Selective Relaying. IEEE International Conference on Communications, pp. 1 – 6.

Hasna, M. O. & Alouini, M. S. (2002). Performance Analysis of Two-Hop Relayed Transmissions over Rayleigh Fading Channels. In Proceedings of the IEEE Vehicular Technology Conference, 4: 1992–1996.

Hasna, M. O. & Alouini, M.-S. (2004). Harmonic Mean and End-to-End Performance of Transmission Systems with Relays. IEEE Transactions on communication: 52(1): 130-135.

Hasna, M.O. & Alouini, M.-S. (2004). A performance Study of Dual-Hop Transmissions with Fixed Gain Relays. IEEE Transactions on wireless communication. 3(6): 1963-1968.

Hindia, M.N., Reza, A.W. & Noordin, K.A. (2015). A Novel Scheduling Algorithm Based on Game Theory and Multicriteria Decision Making in LTE Network. International Journal on Distributed Sensor Networks. 2015: 1-8.

Hoo, L. M., Halder, B., Tellado, J. & Cioffi, J. M. (2004). Multiuser Transmit Optimization for Multicarrier Broadcast Channels: Asymptotic FDMA Capacity Region and Algorithms. IEEE Transactions on Wireless Communications. 52(6): 922-930.

Hui, S. & Yeung, K. (2003). Challenges in the migration to 4G mobile systems. IEEE Communications Magazine. 41(12): 54 – 59.

IEEE (2009). IEEE Standard 802.16j-2009 (Amendment to IEEE Std 802.16-2009) for Local and Metropolitan Area Networks Part 16: Air Interface for Broadband Wireless Access Systems Amendment 1: Multiple Relay Specification.

IEEE (2011).IEEE Standard 802.16m-2011 (Amendment to IEEE Std 802.16-2009) for Local and Metropolitan Area Networks Part 16: Air Interface for Broadband Wireless Access Systems Amendment 3: Advanced Air Interface.

Jain, R., Chiu, D. & Hawe, W. (1984). A Quantitative Measure of Fairness and Discrimination for Recourse Allocation in Shared Computer Systems. ACM Transaction on Computer Systems. DEC Research Report TR-301.

Jalali, A., Padovani, R. & Pankai, R. (2000). Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Communication Wireless System. In Proceedings of the 51st IEEE Vehicular Technology Conference, 3: 1854–1858.

Janani, M., Hedayat, A. , Hunter, T. E. & Nosratinia, A. (2004). Coded Cooperation in Wireless Communications: Space-Time Transmission and Iterative Decoding. IEEE Transactions on Signal Processing. 52(2): 362-371.

Jang, J. & Lee, K. B. (2003). Transmit Power Adaptation for Multiuser OFDM Systems. IEEE Journal on Selected Areas in Communications. 21(2): 171–178.

Javidi, S. & Kittipiyakul, T. (2004). Subcarrier Allocation in OFDMA Systems: Beyond Water-Filling. In Proceedings of the Asilomar Conference on Signals, Systems and Computers. 1: 334–338.

Jiang, Z., Ge, Y. & Li, Y. G. (2005). Max-Utility Wireless Resource Management for Best Effort Traffic. IEEE Transactions on Wireless Communications. 4(1): 100–111.

Juyeop, S. C., Dong-Ho, K., Sunghyun , C., Sangboh, C., Won-Hyoung, Y., Kiho, P., Kwon, K. T. & Lee, H. (2005). Design and Implementation of a Simulator Based on a Cross-Layer Protocol between MAC and PHY Layers in a Wibro Compatible IEEE 802.16e OFDMA System. IEEE Communications Magazine. 43(12):136 – 146.

Katoozian, M., Navaie, K. & Yanikomeroglu, H. (2009). Utility-Based Adaptive Radio Resource Allocation in OFDM Wireless Networks with Traffic Prioritization. IEEE Transactions on Wireless Communications. 8(1): 66–71.

Katoozian, M., Navaie, K. & Yanikomeroglu, H. (2008). Optimal Utility-Based Resource Allocation for OFDM Networks with Multiple Types of Traffic. In Proceedings of the 67th IEEE Vehicular Technology Conference, pp. 2223–2227.

Keeney, R. L. & Raiffa, H. (1994). Decisions with Multiple Objectives- Preferences and Value Tradeoffs. Cambridge and New York: Cambridge University Press. 39(2): 169-170.

Kelly, F. (1997). Charging and rate control for elastic traffic. European Transactions on Telecommunications. 8: 33–37.

Kelly, F. P., Maulloo, A. K. & Tan, D. K. H. (1998). Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability. Journal of Operational Research Society. 49: 237–252.

Kim, H. & Han, Y. (2005). A Proportional Fair Scheduling for Multicarrier Transmission Systems. IEEE Communications Letters. 9(3): 210–212.

Kim, J.B. & Kim, D. (2009). Comparison of Tightly Power-Constrained Performances for Opportunistic Amplify-and-Forward Relaying with Partial or Full Channel Information. IEEE Communications Letters. 13: 100–102.

Kim, Y. & Sichitiu, M. L. (2010). Fairness Schemes in 802.16j Mobile Multihop Relay Networks. In Proceedings of the IEEE GLOBECOM, pp. 1-5.

Knopp, R. & Humblet, P. A. (1995). Information Capacity and Power Control in Single-Cell Multiuser Communications. In Proceedings of the IEEE International Conference on Communications, pp. 331-335.

Kramer, G., Gastpar, M. & Gupta, P. (2005). Cooperative Strategies and Capacity Theorems for Relay Networks. IEEE Transactions on Information Theory. 51(9): 3037-3063.

Kullarni, G., Adlakha, S. & Srivastava, M. (2005). Subcarrier Allocation and Bit Loading Algorithms for OFDMA-Based Wireless Networks. IEEE Transactions on Mobile Computing. 4(6): 652–662.

Kuo, W.-H. & Liao, W. (2005). Utility-Based Optimal Resource Allocation in Wireless Networks. In Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM), pp. 3508–3512.

Kuo, W.-H. & Liao, W. (2007). Utility-based resource allocation in wireless networks. IEEE Transactions on Wireless Communications. 6(10): 3600–3606.

Kushner, H. & Whiting, P. (2004). Convergence of Proportional-Fair Sharing Algorithms under General Conditions. IEEE Journal on Wireless Communication, 3(4): 1250–1259.

Kwon, T., Choi, Y. & Das, S.K. (2002). Bandwidth Adaptation Algorithms for Adaptive Multimedia Services in Mobile Cellular Networks. Wireless Personal Communications. 22(3): 337-357.

Laneman, J. N. & Wornell, G. W. (2003). Distributed Space-Time Coded Strategies for Exploiting Cooperative Diversity in Wireless Networks. IEEE Transactions on Information Theory. 49(10): 2415-2525.

Laneman, J. N., Tse, D. N. C. & Wornell, G. W. (2004). Cooperative Diversity In Wireless Networks: Efficient Strategy and Outage Behavior. IEEE Transactions on Information Theory. 50(12): 3062-3080.

Laneman, J.N. & Wornell, G.W. (2000). Energy-Efficient Antenna Sharing and Relaying for Wireless Networks. In Proceedings of the IEEE Wireless Communications and Networking Conference, 1(1): 7–12.

Lataief, K. B. & Zhang, Y. J. (2006). Dynamic Multiuser Resource Allocation and Adaptation for Wireless Systems," IEEE Wireless Communications Magazine. 13(4): 38–47.

Lee, B.G., Park, D. & Seo, H. (2009). Wireless Communications Resource Management. IEEE Press, John Wiley & sons. pp. 30-64.

Lee, L.T.H., Chung-Ju, C., Yih-Shen, C. & Shen, S. (2005). A Utility-Approached Radio Resource Allocation Algorithm for Downlink in OFDMA Cellular Systems. In Proceedings of the 61st IEEE Vehicular Technology Conference, pp. 1798 - 1802.

Lee, V.-A. (2009). Performance of Relaying Protocols (Doctoral Thesis), Helsinki University of Technology, Finland.

Lee, W.C.Y. (1986). Elements of cellular mobile radio systems. IEEE Transactions on Vehicular Technology. 35(2): 48–56.

Lei, H., Zhang, L., Zhang, X. & Yang, D. (2007). A packet Scheduling Algorithm using Utility Function for Mixed Services in the Downlink of OFDMA Systems. In Proceedings of the 66th IEEE Vehicular Technology Conference, pp. 1664 - 1668.

Leke, A. & Cioffi, J. M. (1998). Multicarrier Systems with Imperfect Channel Knowledge. In Proceedings of the IEEE International Symposium PIMRC, 2: 549–553.

Li, F. & Jafarkhani, H. (2009). Resource Allocation Algorithms with Reduced Complexity in MIMO Multi-Hop Fading Channels. In Proceedings of the Wireless Communications and Networking Conference, pp. 1–6.

Li, Y.-B., Zhang, X., Ye, F. & Gao, Z.G. (2012). Utility Based Dynamic Subcarrier Allocation for QoS Guarantees in SC-FMDA System. RadioEngineering. 21(4): 1147-1155.

Lima, F.R.M., Rodrigues, E.B., Macial, T.F. & Nordberg, M. (2014). Resource Allocation and MIMO for 4G and Beyond. New York: Springer Science and Business Media. pp. 63-104.

Lin, B., Ho, P., Xie, L., Shen, X. & Tapolcai, J. (2010). Optimal Relay Station Placement in

Lin, Y.-N., Wu, C.-W., Lin, Y.-D. & Lai, Y.-C. (2008). A Latency and Modulation Aware Bandwidth Allocation Algorithm for WiMAX Base Stations. In Proceedings of the IEEE Wireless Communication and Networking Conference, pp. 1408-1413.

Liu, G., Zhang, J., Zhu, J. & Wang, W. (2007). Channel and Queue Aware Scheduling for Mixed Service in Multiuser MIMO OFDM System. In Proceedings of the IEEE Wireless Communication, Networking and Mobile Computing, pp. 500-503.

Liu, K.-H., Cai, L. & Shen, X. (2008). Multiclass Utility-Based Scheduling for UWB Networks. IEEE Transactions on Vehicular Technology. 57(2): 1176-1187.

Liu, X., Chong, E. & Shroff, N. (2001). Opportunistic Transmission Scheduling with Resource Sharing Constraints in Wireless Networks. IEEE Journal on Selected Areas in Communications. 19(10): 2053 – 2064.

Loa, K., Wu, C., Sheu, S., Yuan, Y., Chion, M., Huo, D. & Xu, L. (2010). IMT-Advanced Relay Standards [WiMAX/LTE Update]. IEEE Communication. Magazine. 48(8): 40–48.

Lu, N. (2007). Utility-based Bandwidth Adaptation for QoS Provisioning in Multimedia Wireless Networks (Doctoral Thesis), Queen Mary University of London, UK.

Madan, R., Boyd, S. & Lall, S. (2010). Fast algorithms for resource allocation in wireless cellular networks. IEEE/ACM Transactions on Networking. 18(3): 973–984.

Miao, G. & Himayat, N. (2008). Low Complexity Utility Based Resource Allocation for 802.16 OFDMA Systems. In Proceedings of the Wireless Communications and Networking Conference, pp. 1465-1470.

Miao, G., Himayat, N. & Li, G. Y. (2010). Energy-Efficient Link Adaptation in Frequency-Selective Channels," IEEE Transactions on Wireless Communications. 58(2): 545–554.

Mohanram, C. & Bhashyam, S. (2007). Joint Subcarrier and Power Allocation in Channel-Aware Queue-Aware Scheduling for Multiuser OFDM. IEEE Transactions on Wireless Communications. 6(9): 3208–3213.

Mokari, N., Javan, M. R. & Navaie, K. (2010). Cross-Layer Resource Allocation in OFDMA Systems for Heterogeneous Traffic with Imperfect CSI. IEEE Transactions on Vehicular Technology. 59(2): 1011–1017.

Musabe, R. & Larijani, H. (2014). Cross-layer scheduling and Resource Allocation for Heterogeneous Traffic in 3G LTE. Journal of Computer Networks and Communications. 2014: 1-13.

Mushtaq, M.S., Augustin, B. & Mellouk, A. (2014). QoE-Based LTE Downlink Scheduler for VoIP. In Proceeding of the IEEE Wireless Communications and Networking Conference, pp. 2190-2195.

Nabar, R. U., Oyman, O., Bolsckei, H. & Paulraj, J. (2003). Capacity Scaling in MIMO Wireless Networks. In Proceedings of Allerton Conference on Communication, Control and Computing, pp. 378-389.

Nabar, R., Bolcskei, H. & Kneubuhler, F. (2004). Fading Relay Channels: Performance Limits and Space-Time Signal Design. IEEE Journal on Selected Areas in communications. 22(6): 1099 – 1109.

Navaie, K., Valaee, S., Sharafat, A.-R. & Sousa, E. (2007). Optimum Model-Based Non-Real-Time Downlink Data Transmission in Heterogeneous DS-CDMA Cellular Networks. IEEE Transactions on Wireless Communications. 6(6): 2357–2367.

Ng, D.W. K, Lo, E. S. & Schober, R. (2012). Dynamic Resource Allocation in MIMO-OFDMA Systems with Full-Duplex and Hybrid Relaying. IEEE Transactions On Communications. 60(5):1291-1304.

Pantelidou, A. & Ephremides, A. (2008). What is optimal scheduling in wireless networks?. In Proceedings of the Annual International Conference on Wireless Internet, pp. 1–8.

Parag, P., Bhashyam, S. & Aravind, R. (2005). A subcarrier Allocation Algorithm for OFDMA using Buffer and Channel State Information. In Proceedings of the 62nd IEEE Vehicular Technology Conference, 1: 622–625.

Patel, C.S., Stuber, G.L. & Pratt, T.G. (2006). Statistical Properties of Amplify and Forward Relay Fading Channels. IEEE Transactions on Vehicular Technology. 55(1): 1–9.

Peter, S. & Heath, R. (2009). The future of WIMAX: Multihop Relaying with IEEE 802.16j," IEEE Communications Magazine. 47(1): 104–111.

Piro, G., Grieco, L.A., Boggia, G., Fortuna, R. & Camarda, P. (2011). Two-Level Downlink Scheduling for Real-Time Multimedia Services in LTE Networks. IEEE Transactions on Multimedia. 13(5): 1052-1065.

Rasouli, H. (2012). Cooperative Subcarrier and Power Allocation in OFDM Based Relaying System (Doctoral Thesis), Ryerson University, Canada.

Rhee, W. & Cioffi, J. M. (2000). Increase in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation. In Proceedings of the 51st IEEE Vehicular Technology Conference, 2: 1085 -1089.

Riihonen, T., Werner, S., Wichman, R. & Eduardo, Z. B. (2009). On the Feasibility of Full-Duplex Relaying in the Presence of Loop Interference. In Proceedings of the IEEE 10th Workshop on Signal Processing Advances in Wireless Communications, pp. 275–279.

Rodrigues, E. B. & Casadevall, F. (2009). Adaptive Radio Resource Allocation Framework for Multiuser OFDM. In Proceedings of the 69th IEEE Vehicular Technology Conference, pp. 1-6.

Rodrigues, E. B. & Casadevall, F. (2012). Rate Adaptive Resource Allocation with Fairness Control for OFDMA Networks. In Proceedings of the European Wireless Conference, pp. 18–20.

Ryu, S., Ryu, B., Seo, H. & Shin, M. (2005). Urgency and Efficiency Based Packet Scheduling Algorithm for OFDMA Wireless System. In Proceedings for the 61st IEEE Vehicular Technology Conference, 3: 1456-1462.

Saleh, A., Redana, S. & Raaf, B. (2009). Performance of Amplify-and-Forward and Decode-and-Forward in LTE-Advanced. In Proceedings of the IEEE Vehicular Technology Conference, pp. 1-5, 20-23.

Salem, M., Adinoyi, A., Yanikomeroglu, H. & Falconer, D. (2010). Opportunities and Challenges in OFDMA-Based Cellular Relay Networks: A Radio Resource Management Perspective. IEEE Transactions on Vehicular Technology. 59(5): 2496-2510.

Sediq, A., Gohary, R. & Yanikomeroglu, H. (2012). Optimal Tradeoff between Efficiency and Jain's Fairness Index in Resource Allocation. In Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 577–583.

Sediq, A., Gohary, R., Schoenen, R. & Yanikomeroglu, H. (2013). Optimal Tradeoff between Sum-Rate Efficiency and Jain's Fairness Index in Resource Allocation. IEEE Transactions on Wireless Communications. 12(7): 3496–3509.

Shakkottai, S. & Srikant, R. (2002). Scheduling Real-Time Traffic with Deadlines over a Wireless Channel. ACM/Baltzer Wireless Networks. 8(1): 13 - 26.

Shakkottai, S. & Stolyar, A.L. (2001). Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR. International Teletraffic Science and Engineering. 4: 793 – 804.

Sharifian, A. (2014). Utility-based Packet Scheduling and Resource Allocation Algorithms with Heterogeneous Traffic for Wireless OFDMA Networks (Doctoral Thesis). Carleton University Ottawa, Ontario, Canada.

Shen, Z., Andrews, J. G. & Evans, B. L. (2003). Optimal Power Allocation in Multiuser OFDM Systems. In Proceedings of the IEEE Global Communications Conference, 1: 337-341.

Shen, Z., Andrews, J. G. & Evans, B. L. (2005). Adaptive Resource Allocation in Multiuser OFDM Systems with Proportional Fairness. IEEE Transactions on Wireless Communications. 4(6): 2726–2737.

Shenker, S. (1995). Fundamental design issues for the future Internet. IEEE Journal on Selected Areas in Communications. 13(7): 1176–1188.

Shin, S. & Byung-Han, R. (2004). Packet Loss Fair Scheduling Scheme for Real-Time Traffic in OFDMA Systems. ETRI Journal. 26(5): 391 - 396.

Simoens, S., Munoz, O. & Vidal, J. (2007). Achievable Rates of Compress-and-Forward Cooperative Relaying on Gaussian Vector Channels. In Proceedings of the IEEE International Conference on Communications, pp. 4225–4231.

Sklar, B. (1997). Rayleigh fading channels in mobile digital communication systems characterization. IEEE Communications Magazine. 35(7):90–100.

Soldani, D. & Dixit, S. (2008). Wireless relays for broadband access. IEEE Communications Magazine. 46(3): 58–66.

Song, G. & Li, Y. G. (2003a). Adaptive Subcarrier and Power Allocation in OFDM Based on Maximizing Utility. In Proceedings of the IEEE Vehicular Technology Conference, 2: 905–909.

Song, G. & Li, Y. G. (2003b). Adaptive Resource Allocation Based on Utility Optimization in OFDM. In proceedings of the IEEE Global Telecommunication Conference (GLOCOM), 2: 586 – 590.

Song, G. & Li, Y. G. (2005a). Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Network. IEEE Communications Magazine. 43(12): 127-134.

Song, G. & Li, Y. G. (2005b). Cross-Layer Optimization for OFDM Wireless Network—Part I: Theoretical Framework," IEEE Transactions on Wireless Communications. 4(2): 614–624.

Song, G. & Li, Y. G. (2005c). Cross-Layer Optimization for OFDM Wireless Networks-Part II: Algorithm Development. IEEE Transactions on Wireless Communications. 4(2): 625–634.

Song, G. (2005). Cross-layer Resource Allocation and Scheduling in Wireless Multicarrier Networks (Doctoral Thesis), Georgia Institute of Technology, USA.

Song, G., Li, Y., & Cimini, L. J. (2009). Joint Channel- and Queue-Aware Scheduling for Multiuser Diversity in Wireless OFDMA Networks. IEEE Transactions on Communications. 57(7): 2109–2121.

Song, G., Li, Y., Cimini, L. J., & Zheng, H. (2004). Joint Channel-Aware and Queue-Aware Data Scheduling in Multiple Shared Wireless Channels. In Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), pp. 1939–1944.

Stratogiannis, D.G., Tsiropoulos, G. I., Kanellopoulos, J.D. & Cottis, P. G. (2010). 4G Wireless Networks: Architectures, QoS Support and Dynamic Resource Management. Wireless Network Traffic and Quality of Service Support: Trends and Standards, 1: 1-31.

Svedman, P., Wilson, S. K., Cimini, L. J. & Ottersten, B. (2007). Opportunistic Beamforming and Scheduling for OFDMA Systems. IEEE Transactions on Communications. 55(5): 941–952.

Tarokh, V., Jafarkhani, H. & Calderbank, A. (1999). Space-Time Block Coding for Wireless Communications: Performance Results. IEEE Journal on Selected Areas in Communications. 45(5): 1456–1467.

Tse, D. & Viswanath, P. (2005). Fundamentals of Wireless Communication. Cambridge University Press. pp. 10-37.

Tung, T-L. & Yao, K. (2002). Channel Estimation and Optimal Power Allocation for a Multiple-Antenna OFDM System. EURASIP Journal on Applied Signal Processing. 2002(3): 330-339.

Ukil, A. (2009). Cross-Layer Optimization in QoS Aware Next Generation Wireless Networks. In Proceedings of the 7th International Conference of Information, Communications and Signal Processing, pp. 1-5.

Vulpe, A., Halunga, S. & Fratu, O. (2014). QoS-Aware Downlink Scheduling in Multi-Carrier Communication Systems. In Proceedings of the 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace and Electronic Systems, pp. 1-5.

Wang, H. & Dittmann, L. (2010). Downlink Resource Management for QoS Scheduling in IEEE802.16 WiMAX Networks. Computer communications. 33(8): 940-953.

Wang, H. & Jia, W. (2010). An Optimized Scheduling Scheme in OFDMA Wimax Networks. International Journal of Communication Systems. 23(1): 23-39.

Wang, L., Ji, Y. & Liu, F. (2010). Joint Optimization for Proportional Fairness in OFDMA Relay-Enhanced Cellular Networks. In Proceedings of the Wireless Communication and Networking Conference, pp. 1-6.

Wang, X., Giannakis, G. & Marques, A. (2007). A Unified Approach to QoS-Guaranteed Scheduling for Channel-Adaptive Wireless Networks. In Proceedings of the IEEE. 95(12): 2410–2431.

Wang, Y., Lei, H. & Zhang, X. (2008). Real-Time Traffic Scheduling Algorithm for MIMO-OFDMA Systems. In Proceedings of the IEEE International Conference on Communications, pp. 4511 - 4515.

Wong, C. Y., Cheng, R. S., Letaief, K. B. & Murch, R. D. (1999). Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation. IEEE Journal on Selected Areas in Communication. 17(10): 1747–1758.

Yanikomeroglu, H. & Zhang, J. (2008). Beyond-4G Cellular Networks: Advanced Radio Access Network (RAN) Architectures, Advanced Radio Resource Management (RRM) Techniques, and other Enabling Technologies. In Proceedings of the WWRF Meeting, pp. 13–15.

Yao, Y. & Giannakis, G. B. (2005). Rate Maximizing Power Allocation in OFDM Based On Partial Channel Knowledge. IEEE Transactions on Wireless Communications. 4(3): 1073–1083.

Yin, H. & Alamouti, S. (2006). OFDMA: A Broadband Wireless Access Technology. IEEE Sarnoff Symposium, pp. 1-4.

Zhang, G. (2004). Subcarrier and Bit Allocation for Real-Time Services in Multiuser OFDM Systems. IEEE Communications Society. 5: 2985–2989.

Zhang, W., Xiang-Gen, X. & Letaief, K. B. (2007). Space-Time/Frequency Coding for OFDM in Next Generation Broadband Wireless Systems. IEEE Transactions on Wireless Communications. 14(2): 32–43.

Zhu, X., Huo, J., Xu, X., Xu, C. & Ding, W. (2008). QoS-Guaranteed Scheduling and Resource Allocation Algorithm for IEEE 802.16 OFDMA Systems. In Proceedings of the IEEE International Conference on Communication, pp. 3463-3468.

Appendix A

List of Publications

A1. Journal Publications

Bello, O., Zen, H., Othman, A. K. & Hamid, K. A. (2015). Computing Amplify-and-Forward Relay Amplification Factor to Improve Total Capacity at Destination. American Journal of Applied Sciences, 12(8): 572-580.

Bello, O., Zen, H., Othman, A. K. & Hamid, K. A. (2015). Enhancing Relay-Based Resource Allocation with Amplify-and-Forward Relaying. WSEAS Transactions on Communications. 14(48): 417-425.

Bello, O., Zen, H., Othman, A. K. & Hamid, K. A. (2015). Utility-based Scheduling Frameworks for Efficient Quality-of-Service Differentiation in a Mixture of Real-Time and Non-Real-Time Traffics. Journal of Computer Science, 11(7): 845-854.

A2. Conference Publications

Bello, O., Zen, H., Othman, A. K. & Hamid, K. A. (2015). Efficient and Low-Complexity Scheduling Algorithm for Multi-user Heterogeneous Traffic Scenario. 12th Malaysian IEEE International Conference on Communications, Kuching, 23-25 November.

Bello, O., Zen, H., Othman, A. K. & Hamid, K. A. (2015). Amplify-and-Forward Relaying to Improve Capacity and Error Rate at Destination Terminal. 12th Malaysian IEEE International Conference on Communications, Kuching, 23-25 November.

Bello, O., Zen, H., Othman, A. K. & Hamid, K. A. (2014). Utility Function in Broadband Wireless Network. In Proceedings of the Post Graduate Research Colloquium: Research Trends in Engineering and Business, Curtin University, Miri, Sarawak, pp. 29.

Appendix B

Algorithms for Subcarrier Assignment

B.1: Algorithm for Subcarrier Assignment for RT Traffic

$i = 1$; $\{Initialize\ iteration\ counter\}$

$\mathcal{N}_k^{(1)} = \mathcal{N}_k$ $\{Initialize\ set\ of\ non-allocated\ subcarrier\}$

$Q_m^{(1)}(n) = Q_m(n)\ \forall m$ $\{Initialize\ \ queue\ lengths\ in\ time\ slot\ n\}$

$P_T^{(1)} = P_T$ $\{Initialize\ \ available\ power\}$

$\boldsymbol{while}$ $\mathcal{N}_k^{(i)} \neq \emptyset$ and $\sum_{m=1}^{N_m} Q_m^{(i)} \neq 0$ and $D_{m_k^*}^{hol}[t] \leq D_m^{max}$ $\boldsymbol{do}$

$P_{m,k} = \begin{cases} P_T/N_k, & m = m_k^* \\ 0, & m \neq m_k^* \end{cases}$ $\{\ Allocate\ power\ \}$

$m_k^* = \arg max_{\Re_m} \left\{ w_m \min \left\{ r_{m,k}, \frac{Q_m^{(i)}}{T_s} \right\} \right\}$ $\{\ Allocate\ subcarrier\ \}$

$\mathcal{N}_k^{(i+1)} = \left\{ \mathcal{N}_k^{(i)} \backslash k \right\}$ $\{\ Update\ non-allocated\ subcarrier\ \}$

$Q_{m_k^*}^{(i+1)} = Q_{m_k^*}^{(i)} - R_{m_k^*} T_s$ $\{\ Update\ queue\ \}$

$i = i + 1$; $\{\ Update\ iteration\ counter\ \}$

$\boldsymbol{end\ while}$

B.2: Algorithm for Subcarrier Assignment for NRT Traffic

$i = 1;$ $\{Initialize\ iteration\ counter\}$

$\mathcal{N}_k^{(1)} = \mathcal{N}_k$ $\{Initialize\ set\ of\ non - allocated\ subcarrier\}$

$Q_m^{(1)}(n) = Q_m(n)\ \forall m$ $\{Initialize\ \ queue\ lengths\ in\ time\ slot\ n\}$

$P_T^{(1)} = P_T$ $\{Initialize\ \ available\ power\}$

$\boldsymbol{while}$ $\mathcal{N}_k^{(i)} \neq \emptyset$ and $\sum_{m=1}^{N_m} Q_m^{(i)} \neq 0$ $\boldsymbol{do}$

$P_{m,k} = \begin{cases} P_T/N_k, & m = m_k^* \\ 0, & m \neq m_k^* \end{cases}$ $\{\ Allocate\ power\ \}$

$m_k^* = \arg max_{\Re_m} \left\{ w_m \min \left\{ r_{m,k}, \frac{Q_m^{(i)}}{T_s} \right\} \right\}$ $\{\ Allocate\ subcarrier\ \}$

$\mathcal{N}_k^{(i+1)} = \left\{ \mathcal{N}_k^{(i)} \backslash k \right\}$ $\{\ Update\ non - allocated\ subcarrier\ \}$

$Q_{m_k^*}^{(i+1)} = Q_{m_k^*}^{(i)} - R_{m_k^*} T_s$ $\{\ Update\ queue\ \}$

$i = i + 1;$ $\{\ Update\ iteration\ counter\ \}$

$\boldsymbol{end\ while}$

Appendix C

MATLAB Codes for the simulations

C.1: Matlab code for Main program

```
close all; clear all; clc
start=tic;
%--------plotting against user arrival rate (users per time slot)
us=60;incr=12;maxusers=120;%
uuu=[us:incr:maxusers];
%store random values to distribute users randomly in cell
RandNo = rand(10,maxusers);
no_tfic=4;
ba=zeros(1,length(uuu));
dcolor1='-k*';
dcolor2='-go';
dcolor3='-bs';
dcolor4='-rv';
dcolor5='-md';
dcolor6='-cp';
% For plotting averaged results for different scheduling schemes
MQGdroppedT=zeros(no_tfic,length(uuu)); aveDelayMQG=zeros(no_tfic,length(uuu));
systhrMQG=zeros(no_tfic,length(uuu)); dmissMQG=zeros(no_tfic,length(uuu));
userthrMQG=zeros(no_tfic,length(uuu)); tfMQG=zeros(no_tfic,length(uuu));
delayMQG=zeros(no_tfic,length(uuu)); MQGcov=zeros(no_tfic,length(uuu));

DSMdroppedT=zeros(no_tfic,length(uuu)); aveDelayDSM=zeros(no_tfic,length(uuu));
systhrDSM=zeros(no_tfic,length(uuu)); dmissDSM=zeros(no_tfic,length(uuu));
userthrDSM=zeros(no_tfic,length(uuu)); tfDSM=zeros(no_tfic,length(uuu));
delayDSM=zeros(no_tfic,length(uuu)); DSMcov=zeros(no_tfic,length(uuu));
%----------------------------------
NT=1;NR=1; % NT (Transmit Antenna),NR (Receive Antenna)
```

```
N=10;  %Number of time slots
B=5*10^(6);%system banddwidth 3MHz for relay
Nb=1; %number of subband
Nsc=64; %no of subcarrier per subband
Bb=B/Nb; %subband bandwidth
K=Nsc*Nb; %Total no of subcarriers
Tcp=1.42*1e-5;%cyclic prefix duration
No=20;% Number of OFDM symbols per slot
To=91.4286*1e-6;% OFDM symbol duration
Fs=To*No;%Frame duration
Ts=2.0571e-3;%Ts=No*(To+Tcp)=2.0571e-3
bw=Ts;%1/1000; %window size of 10=0.1, 100=0.01, 1000=0.001
epsilon=0.001;% initial throughput to avoid division by zero
BER=10.^(-5);
Beta=-log(5*BER)/1.5;%SNR gap beta 3 or 1
Pt_dBm_BS=33.9897; % 0.5 wattdBm
Pt_dBm_RS=33.9897;%27; % dBm
Pt_dBm_P=30;%33.9897;%37;% dBm 30;%
N0=-123.24; %noise power in dBm/Hz
Noise_dBm_M=-142;% dBW
noiseM=dBw2w(Noise_dBm_M); %linear noise (watt) at MS per subcarrier
noiseR=dBw2w(Noise_dBm_M); %linear noise (watt) at RS per subcarrier
Ptot=dBm2w(Pt_dBm_P);
SUI_no=4; %SUI number
Nfading=1024; % Size of Doppler filter
Nfosf=4;%fading oversampling factor
[Delay_us, Power_dB, K_factor, Doppler_shift_Hz, Ant_corr, Fnorm_dB]...
    =SUI_parameters(SUI_no);%SUI parameters
radius=1000;% Cell radius =1Km
noRL=1;% no of relays
dThld=zeros(1,noRL);% relay location holder
IRdt=500;
```

```
for kl=1:noRL
  dThld(kl)=IRdt;% dThld=500 for equal user distribution within regions
  IRdt=IRdt+100;
end

PowerP=Ptot/K;% Power per subcarrier
%Compute Path Loss
fc=2e9;%in Hz
% Compute shadow variance
shadow_dB=8; shadowing=10^(-shadow_dB/10);%linear  (watt)
 % for implementing AMC
 %  SNR  thresholds     efficiency
              (dB)                   (bps/Hz)
 DRC=[  -9999       5.0           0;         %BPSK
                5.0         8.0         1.0;         %QPSK rate 1/2
                8.0         10.5        2.0;         %QPSK rate 2/3
              10.5        14.0        3.0;         %QPSK rate 3/4
              14.0        16.0        3.5;         %QPSK rate 7/8
              16.0        18.0        4.0;         %QAM16 rate 1/2
              18.0        20.0        4.5;         %QAM16 rate 2/3
              20.0        99999       6.0];        %QAM64 rate 1
ll=0;arrivalrate=5e+5;%Default assignments  (Dont change!)
PowerRS=zeros(noRL,uu*K); PowerBS=zeros(noRL,uu*K);

%----plotting against users
for uu=us:incr:maxusers
    ll=ll+1;
   for hh=1:noRL
       %Equal power allocation
       PowerRS(hh,:)=0.5*PowerP;
       PowerBS(hh,:)=0.5*PowerP;
   End
```

```
 %Conventional wireless networks parameters
 HHbs_ms_cell=zeros(1,NT*NR,uu*K);
 SNRbs_ms_cell=zeros(1,uu*K);
 dBS_MS_CELL=zeros(1,uu);
 BS_MS_CELLlocationSNR=zeros(noRL,uu);

%relay networks parameters
 HHbs_ms_rz=zeros(noRL,NT*NR,uu*K);  HHbs_ms_bz=zeros(noRL,NT*NR,uu*K);
 HHbs_rs=zeros(noRL,NT*NR,uu*K);  HHrs_ms=zeros(noRL,NT*NR,uu*K);
 Hbs_rs=zeros(noRL,uu*K);  Hrs_ms=zeros(noRL,uu*K);
 SNRbs_ms_rz=zeros(noRL,uu*K);  SNRbs_ms_bz=zeros(noRL,uu*K);
 SNRrs_ms=zeros(noRL,uu*K);  SNRbsrsHD=ones(noRL,uu*K);
 SNRbsrsFD=ones(noRL,uu*K);  Drm_AF_HD=zeros(noRL,uu*K);
 AFcoeff_HD=ones(noRL,uu*K);  SNR_AF_HD=zeros(noRL,uu*K);
 Drm_AFF_FD=zeros(noRL,uu*K);  AFFcoeff_FD=ones(noRL,uu*K);
 SNR_AFF_FD=zeros(noRL,uu*K);  Drm_AFF_HD=zeros(noRL,uu*K);
 AFFcoeff_HD=ones(noRL,uu*K);  SNR_AFF_HD=zeros(noRL,uu*K);
 Drm_AF_FD=zeros(noRL,uu*K);  AFcoeff_FD=ones(noRL,uu*K);
 SNR_AF_FD=zeros(noRL,uu*K);  YLI=zeros(noRL,uu*K);
 err_deco=zeros(noRL,uu*K);  DFcoeff_HD=zeros(noRL,uu*K);
 SNR_DF_HD=zeros(noRL,uu*K);  DFcoeff_FD=zeros(noRL,uu*K);
 SNR_DF_FD=zeros(noRL,uu*K);  dBS_RS=zeros(noRL,uu);
 dBS_MS_RZ=zeros(noRL,uu);  dBS_MS_BZ=zeros(noRL,uu);
 dRS_MS=zeros(noRL,uu); BS_RSlocationSNR=zeros(noRL,uu);
 BS_MS_BZlocationSNR=zeros(noRL,uu); BS_MS_RZlocationSNR=zeros(noRL,uu);
 RS_MSlocationSNR=zeros(noRL,uu);

 for j=1:noRL
  %For conventional network model untick below
  %dBS_MS_CELL(1,1:uu)=RandNo(1,1:uu).*radius;
```

```
 vm=3/4;  %2/4 puts half of users in RS and half in BS region
        %3/4 puts three-quarter of users in RS and one-quarter in BS
        %1/4 puts one-quarter of users in RS and three-quarter in BS
        %4/4 puts all users in RS region only (no users in BS)
%--- For Relay operation  untick below
  dBS_MS_CELL(j,1:uu*vm)=RandNo(j,1:uu*vm).*dThld(j)+(radius-dThld(j));
  dBS_MS_CELL(j,1+uu*vm:uu)=RandNo(j,1+uu*vm:uu).*dThld(j);
 aBM=find(dBS_MS_CELL(j,:)<dThld(j)); %users are located only in BS zone
 aRM=find(dBS_MS_CELL(j,:)>=dThld(j)); %users are located only in RS zone
 dBS_MS_BZ(j,aBM)=dBS_MS_CELL(j,aBM);% assign to BS_MS in BS region
 dBS_MS_RZ(j,aRM)=dBS_MS_CELL(j,aRM);% assign to BS_MS in RS region
 dBS_RS(j,aRM)=dThld(j);% assign to BS_RS in RS region
 dRS_MS(j,aRM)=dBS_MS_CELL(j,aRM)-dBS_RS(j,aRM); % assign to RS_MS in RS
region
end

for g=1:noRL
  RS_MSlocationSNR(g,1:uu)=PL_logdist_or_norm(fc,dRS_MS(g,1:uu),100,3);
  BS_RSlocationSNR(g,1:uu)=PL_logdist_or_norm(fc,dBS_RS(g,1:uu),100,3);
  BS_MS_RZlocationSNR(g,1:uu)=PL_logdist_or_norm(fc,dBS_MS_RZ(g,1:uu),100,3);
  BS_MS_BZlocationSNR(g,1:uu)=PL_logdist_or_norm(fc,dBS_MS_BZ(g,1:uu),100,3);
  BS_MS_CELLlocationSNR(1,1:uu)=PL_logdist_or_norm(fc,dBS_MS_CELL(1,1:uu),...
    100,3);
end
PathLossBS_MS_CELL=10.^(-BS_MS_CELLlocationSNR./10);
PathLossRS_MS=10.^(-RS_MSlocationSNR./10);
PathLossBS_RS=10.^(-BS_RSlocationSNR./10);
PathLossBS_MS_RZ=10.^(-BS_MS_RZlocationSNR./10);
PathLossBS_MS_BZ=10.^(-BS_MS_BZlocationSNR./10);

%--- If INF occurs because it is 0, set to 0
jj1=find(PathLossBS_MS_CELL==Inf);PathLossBS_MS_CELL(jj1)=0;
```

```
jj2=find(PathLossRS_MS==Inf);PathLossRS_MS(jj2)=0;
jj3=find(PathLossBS_RS==Inf);PathLossBS_RS(jj3)=0;
jj4=find(PathLossBS_MS_RZ==Inf);PathLossBS_MS_RZ(jj4)=0;
jj5=find(PathLossBS_MS_BZ==Inf);PathLossBS_MS_BZ(jj5)=0;

% specifying inputs for Uc1, Uc2, Uc3 and Uc4
%total values combined must be equal to 4 in the combination in Uc1 …Uc4
%0-VoIP, 0-Audio, 4-FTP, 0-BE  (selects on FTP)
%2-VoIP, 2-Audio, 0-FTP, 0-BE (multiplex only VoIP and Audio)
%0-VoIP, 2-Audio, 2-FTP, 0-BE (multiplex only Audio and FTP)
%1-VoIP, 1-Audio, 1-FTP, 1-BE (multiplex VoIP, Audio, FTP and HTTP)

Uc1=ceil(uu/4)*0; %VoIP (ertPS)
Uc2=ceil(uu/4)*0; %Audio or Video streaming (rtPS)
Uc3=ceil(uu/4)*4;  % FTP (nrtPS)
Uc4=ceil(uu/4)*0;% HTTP (BE)

QoSclass=zeros(1,uu); Rmin=zeros(1,uu); Rmax=zeros(1,uu);
QoSclass1=zeros(1,Uc1); QoSclass1(1,:)=1; QoSclass2=zeros(1,Uc2); QoSclass2(1,:)=2;
QoSclass3=zeros(1,Uc3); QoSclass3(1,:)=3; QoSclass4=zeros(1,Uc4); NoClass=0;
ClassNo=[ ];
if Uc1>0, NoClass=NoClass+1;ClassNo=[ClassNo 1];end
if Uc2>0, NoClass=NoClass+1;ClassNo=[ClassNo 2];end
if Uc3>0, NoClass=NoClass+1;ClassNo=[ClassNo 3];end
if Uc4>0, NoClass=NoClass+1;ClassNo=[ClassNo 4];end
Rmin1=zeros(1,Uc1);Rmax1=zeros(1,Uc1); Rmin2=zeros(1,Uc2);Rmax2=zeros(1,Uc2);
Rmin3=zeros(1,Uc3);Rmax3=zeros(1,Uc3); Rmin4=zeros(1,Uc4);Rmax4=zeros(1,Uc4);

%minimum required throughput
Rmin1(1,:)=25e3;%VoiP (ertPS)
Rmin2(1,:)=64e3;%audio streaming
Rmin3(1,:)=45e3;%video streaming=64e3
```

```
%maximum sustained throughput
Rmax1(1,:)=64e3;%VoiP (ertPS)
Rmax2(1,:)=500e3;%audio streaming (rtPS)
Rmax3(1,:)=500e3;% FTP (nrtPS)

for dt=1:Uc4
    Rmin4(1,dt)=1e3;%email/HTTP (BE)
    Rmax4(1,dt)=64e3;%email/HTTP (BE)
    QoSclass4(1,dt)=4;
end
Rmin(1,:)=[Rmin1,Rmin2,Rmin3,Rmin4];  Rmax(1,:)=[Rmax1,Rmax2,Rmax3,Rmax4];
QoSclass(1,:)=[QoSclass1,QoSclass2,QoSclass3,QoSclass4];

%Reordering Rmin, because different user are demanded different services
aux1QoSclass=QoSclass;  aux1Rmin=Rmin;  aux1Rmax=Rmax;
aux2Rmin=randperm(length(Rmin));

for f=1:length(Rmin)
   Rmin(f)=aux1Rmin(aux2Rmin(f));
   QoSclass(f)=aux1QoSclass(aux2Rmin(f));
   Rmax(f)=aux1Rmax(aux2Rmin(f));
end

vQQ=zeros(1,uu);  vQU1=zeros(1,uu);  latency=zeros(1,uu); QueueSize=zeros(1,uu);
PacketSize=zeros(1,uu); arrival=zeros(1,uu); ertPS=find(QoSclass==1);
rtPS=find(QoSclass==2); nrtPS=find(QoSclass==3); BE=find(QoSclass==4);

% maximun toreable delay (latency)
latency(ertPS)=20e-3;% ertPS (VoIP)
latency(rtPS)=30e-3;% rtPS (Audio/video streaming)
latency(nrtPS)=100e-3; %nrtPS (FTP)
latency(BE)=2000e-3; %BE (HTTP)
```

```
% For ertPS
% parameter assignment for MQS scheduler for ertPS
vQU1(ertPS)=1.0./(latency(ertPS));
% parameter assignment for DSMS scheduler ertPS
vQQ(ertPS)=log((1-0.01)/0.01)./(0.5.*latency(ertPS));
% For rtPS
vQU1(rtPS)=1.0./(latency(rtPS));
vQQ(rtPS)=log((1-0.01)/0.01)./(0.5.*latency(rtPS));
% For nrtPS
vQU1(nrtPS)=1.0./Rmin(nrtPS);
vQQ(nrtPS)=log((1-0.01)/0.01)./(0.5.*Rmin(nrtPS));
% For BE
vQU1(BE)=1.0./Rmin(BE);
vQQ(BE)=log((1-0.01)/0.01)./(0.5.*Rmin(BE));

MQGBffoverflow=zeros(N,uu); MQGarriveT=zeros(1,uu);
MQGNbitsDropped=zeros(N,uu); dlMQG=zeros(N,uu);
MQG_outage=zeros(N,uu); MQGHOLdelay=zeros(N,uu);
MQGHOLdelay(1,1:uu)=Ts; MQGaveRate=zeros(N,uu);
MQGaveRate(1,1:uu)=epsilon;
MQGrate=zeros(N,uu); MQGuser=zeros(N,uu); MQG_no_outage=zeros(N,uu);
MQGallocated=zeros(N,uu); MQGdropped=zeros(N,uu); MQGdroppedRate=zeros(N,uu);
MQGqdelay=zeros(N,uu); aThMQG=zeros(N,uu); aMQGqueues=zeros(N,uu);
aMQGwaitTime=zeros(N,uu); MQGqueues=zeros(N,uu); MQGqueueBegin=zeros(N,uu);
aThMQG(1,1:uu)=epsilon; %Initial average throughput
MQGdroppedRateS(1,1:uu)=epsilon; MQGqdelay(1,1:uu)=Ts;

DSMBffoverflow=zeros(N,uu);  DSMarriveT=zeros(1,uu);  DSMNbitsDropped=zeros(N,uu);
dlDSM=zeros(N,uu); DSM_outage=zeros(N,uu); DSMHOLdelay=zeros(N,uu);
DSMHOLdelay(1,1:uu)=Ts; DSMaveRate=zeros(N,uu); DSMaveRate(1,1:uu)=epsilon;
DSMrate=zeros(N,uu); DSMuser=zeros(N,uu); DSM_no_outage=zeros(N,uu);
```

```
DSMallocated=zeros(N,uu); DSMdropped=zeros(N,uu); DSMdroppedRate=zeros(N,uu);
aThDSM=zeros(N,uu);ThDSM=zeros(N,uu); aDSMqueues=zeros(N,uu);
aDSMwaitTime=zeros(N,uu); DSMqueues=zeros(N,uu); DSMqueueBegin=zeros(N,uu);
aThDSM(1,1:uu)=epsilon; %Initial average throughput
DSMqdelay(1,1:uu)=Ts;

for s=1:N      %iteration for time slot
%compute  SNR and relay parameters when relay is used for each time
%slot only since quasi-static flat fading is considered
    for jj=1:noRL  %noRL=1 (no of relay)
       HHbs_ms_cell(1,1,:)=SUI_fading(Power_dB, K_factor, Doppler_shift_Hz,...
            Fnorm_dB,K*uu,Nfading, Nfosf,NT*NR);
      HHbs_rs(jj,1,:)=SUI_fading(Power_dB, K_factor, Doppler_shift_Hz,...
           Fnorm_dB,K*uu,Nfading, Nfosf,NT*NR);
      HHbs_ms_rz(jj,1,:)=SUI_fading(Power_dB, K_factor, Doppler_shift_Hz,...
          Fnorm_dB,K*uu,Nfading, Nfosf,NT*NR);
     HHbs_ms_bz(jj,1,:)=SUI_fading(Power_dB, K_factor, Doppler_shift_Hz,...
          Fnorm_dB,K*uu,Nfading, Nfosf,NT*NR);
     HHrs_ms(jj,1,:)=SUI_fading(Power_dB, K_factor, Doppler_shift_Hz,...
         Fnorm_dB,K*uu,Nfading, Nfosf,NT*NR);
end
 p=0;
for i=1:uu
    for k=1:K
        p=p+1;
       %compute channel gains
        Hrs_ms(jj,p)=HHrs_ms(jj,:,p)^2*PathLossRS_MS(jj,i)*shadowing;%NLOS
       Hbs_rs(jj,p)=HHbs_rs(jj,:,p)^2*PathLossBS_RS(jj,i);%LOS
       %SNR on relay links
      SNRbsrsHD(jj,p)=PowerBS(jj,p)*HHbs_rs(jj,:,p)^2*PathLossBS_RS(jj,i)/noiseM;
      SNRbs_ms_rz(jj,p)=PowerBS(jj,p)*HHbs_ms_rz(jj,:,p)^2*...
      PathLossBS_MS_RZ(jj,i)*shadowing/noiseM;%SNR on BS_MS in RZ
```

```
 SNRbs_ms_bz(jj,p)=PowerBS(jj,p)*HHbs_ms_bz(jj,:,p)^2*...
 PathLossBS_MS_BZ(jj,i)*shadowing/noiseM;%SNR on BS_MS in BZ
%SNR in conventional network (when relay is not used)
SNRbs_ms_cell(1,p)=PowerBS(1,p)*HHbs_ms_cell(1,:,p)^2*...
PathLossBS_MS_CELL(1,i)*shadowing/noiseM;%SNR on BS_MS in CELL
if Hrs_ms(jj,p)~=0 || Hbs_rs(jj,p)~=0
    YLI(jj,p)=dB2w(10);% average loop  interference power in FD (dB)
   err_deco(jj,p)=1;%decoding error is 1 (perfect decoding)
   %SNR for Decode-and-Forward (DF)at Destination
   SNR_DF_HD(jj,p)=min((PowerRS(jj,p).*Hrs_ms(jj,p)./noiseM),...
              (err_deco(jj,p)*PowerBS(jj,p)*Hbs_rs(jj,p)./noiseR));
   SNR_DF_FD(jj,p)=min((PowerRS(jj,p)*Hrs_ms(jj,p)/noiseM),...
       (err_deco(jj,p)*PowerBS(jj,p)*Hbs_rs(jj,p)/noiseR)/(PowerBS(jj)*YLI(jj,p)+1));
 %Amplify-and-Forward (AF) in HD mode
   AFcoeff_HD(jj,p)=((noiseR*Drm_AF_HD(jj,p)^2)*PowerRS(jj,p).*...
   Hrs_ms(jj,p)/noiseM)/((noiseR*Drm_AF_HD(jj,p).^2)...
         *PowerRS(jj,p)*Hrs_ms(jj,p)/noiseM +1);
  % SNR at Destination
   SNR_AF_HD(jj,p)=(PowerBS(jj,p)*Hbs_rs(jj,p)/noiseM)*AFcoeff_HD(jj,p);

%Amplify-and-Forward (AF) in FD mode
 Drm_AF_FD(jj,p)=sqrt(1/(PowerBS(jj,p)*Hbs_rs(jj,p)+noiseR+...
      PowerRS(jj,p)*YLI(jj,p)*noiseR));
AFcoeff_FD(jj,p)=((noiseR*Drm_AF_FD(jj,p).^2)*PowerRS(jj,p).*...
         Hrs_ms(jj,p)./noiseM)./(((noiseR.*Drm_AF_FD(jj,p).^2)...
             *PowerRS(jj,p)*Hrs_ms(jj,p)/noiseM)+1/(PowerRS(jj,p)*YLI(jj,p)+1));
SNR_AF_FD(jj,p)=((PowerBS(jj,p).*Hbs_rs(jj,p)./noiseM)./...
        (PowerRS(jj,p)*YLI(jj,p)+1))*AFcoeff_FD(jj,p);

% Equalied Amplify-and-Forward (AF) in HD mode
Drm_AFF_HD(jj,p)=Drm_AF_HD(jj,p)+1/(Drm_AF_HD(jj,p)*PowerRS(jj,p)*...
       Hrs_ms(jj,p)*noiseR./noiseM);
```

```
AFFcoeff_HD(jj,p)=((noiseR.*Drm_AFF_HD(jj,p).^2).*PowerRS(jj,p).*...
          Hrs_ms(jj,p)./noiseM)./((noiseR.*Drm_AFF_HD(jj,p).^2)*...
                PowerRS(jj,p).*Hrs_ms(jj,p)/noiseM +1);%scaling coefficient
SNR_AFF_HD(jj,p)=(PowerBS(jj,p).*Hbs_rs(jj,p)./noiseM).* AFFcoeff_HD(jj,p);

% Equalized Amplify-and-Forward (AF) in FD mode
Drm_AFF_FD(jj,p)=Drm_AF_FD(jj,p)+1/(Drm_AF_FD(jj,p)*PowerRS(jj,p)*...
       Hrs_ms(jj,p)*noiseR/noiseM);
AFFcoeff_FD(jj,p)=((noiseR*Drm_AFF_FD(jj,p)^2)*PowerRS(jj,p)*...
     Hrs_ms(jj,p)/noiseM)/(((noiseR*Drm_AFF_FD(jj,p)^2).*...
            PowerRS(jj,p)*Hrs_ms(jj,p)/noiseM)+1/(PowerRS(jj,p)*YLI(jj,p)+1));
SNR_AFF_FD(jj,p)=((PowerBS(jj,p).*Hbs_rs(jj,p)./noiseM)./...
          (PowerRS(jj,p).*YLI(jj,p)+1)).*AFFcoeff_FD(jj,p);
end
end %for k
end % for uu

%re-formatting data
SNRDF_HD=squeeze(SNR_DF_HD);SNRDF_FD=squeeze(SNR_DF_FD);
SNRAF_HD=squeeze(SNR_AF_HD);SNRAF_FD=squeeze(SNR_AF_FD);
SNRAFF_HD=squeeze(SNR_AFF_HD);SNRAFF_FD=squeeze(SNR_AFF_FD);
SNRBS_MS_RZ=squeeze(SNRbs_ms_rz);SNRBS_MS_BZ=squeeze(SNRbs_ms_bz);
SNRBS_MS_CELL=squeeze(SNRbs_ms_cell);
SNR0=SNRBS_MS_CELL; %SNR in conventional network

 % maximum ratio combining (MRC) at the destination to provide
 % spatial diversity in relay network
SNR10=SNRAF_HD+SNRBS_MS_RZ+SNRBS_MS_BZ; %SNR for existing AF in HD
SNR20=SNRAFF_HD+SNRBS_MS_RZ+SNRBS_MS_BZ; %SNR for proposed AFF in HD
SNR30=SNRDF_HD+SNRBS_MS_RZ+SNRBS_MS_BZ;%SNR for DF in HD
SNR40=SNRAF_FD+SNRBS_MS_RZ+SNRBS_MS_BZ; %SNR for existing AF in FD
SNR50=SNRAFF_FD+SNRBS_MS_RZ+SNRBS_MS_BZ; %SNR for proposed AF in FD
```

```
SNR60=SNRDF_FD+SNRBS_MS_RZ+SNRBS_MS_BZ;%SNR for DF in full-duplex (D)

% Initialize users
users=1:uu;
%********************************** schedulers
% Traffic models
   %for VoIP
PacketSize(ertPS)=512; arrival(ertPS)=Rmax(ertPS); QueueSize(ertPS)=128*1024*8;
   %for video
PacketSize(rtPS)=640; arrival(rtPS)=Rmin(rtPS); QueueSize(rtPS)=128*1024*8;
    %FTP
PacketSize(nrtPS)=1200; arrival(nrtPS)=Rmax(nrtPS); QueueSize(nrtPS)=128*1024*8;
%HTTP
PacketSize(BE)=512; arrival(BE)=Rmax(BE); QueueSize(BE)=128*1024*8;

%DRA- discrete rate adaptation
%CRA countinuous rate adaptation
RA=0; % Rate adaptation   RA=1 (DRA) RA=0 (CRA)
D=1; DD=2;% mode=1 (FD), mode=2 (HD)

% MQS Scheduler subprogram
[MQGNbitsDropped,MQGBffoverflow,MQGHOLdelay,MQG_outage,MQGarriveT,...
 MQGtotalRate,MQGsuballo,MQGrate,dlMQG,MQGaveRate,MQGuser,MQGqdelay,...
 aMQGwaitTime,MQG_no_outage,MQGdroppedRate,MQGdropped,aThMQG,...
 MQGqueueBegin,MQGqueues,aMQGqueues,MQGallocated]...
 =MQGalgorithmLKC(uu,users,K,B,DD,To,N,arrival,Rmin,Rmax,Beta,QoSclass,...
 SNR10,PacketSize,QueueSize,Ts,bw,latency,vQU1,kVGhd,MQGBffoverflow,...
 MQGHOLdelay,MQG_outage,MQGarriveT,MQGrate,dlMQG,MQGaveRate,MQGuser,...
MQGqdelay,aMQGwaitTime,MQG_no_outage,MQGdroppedRate,MQGdropped,aThMQG,..
MQGqueueBegin,MQGqueues,aMQGqueues,MQGallocated,DRC,s,MQGNbitsDropped,RA)
;
```

```
% DSM/TSM Scheduler subprogram
[DSMNbitsDropped,DSMBffoverflow,DSMHOLdelay,DSM_outage,DSMarriveT,...
    DSMtotalRate,DSMsuballo,DSMrate,dlDSM,DSMaveRate,DSMuser,DSMqdelay,...
    aDSMwaitTime,DSM_no_outage,DSMdroppedRate,DSMdropped,aThDSM,...
    DSMqueueBegin,DSMqueues,aDSMqueues,DSMallocated]...
=DSMalgorithmLKC(uu,users,K,B,DD,To,N,arrival,Rmin,Rmax,Beta,...
    QoSclass,SNR10,PacketSize,QueueSize,Ts,bw,latency,vQQ,kVGhd,...
    DSMBffoverflow,DSMHOLdelay,DSM_outage,DSMarriveT,DSMrate,dlDSM,...
    DSMaveRate,DSMuser,DSMqdelay,aDSMwaitTime,DSM_no_outage,...
    DSMdroppedRate,DSMdropped,aThDSM,DSMqueueBegin,DSMqueues,...
    aDSMqueues,DSMallocated,DRC,s,DSMNbitsDropped,RA);
 end     %end of N (time slots)
 % Computing the performance averages
% For MQS
% Average packet delay based on HOL packet
userdelay=MQGHOLdelay(N,:);
delayMQG(1,ll)=mean(userdelay(ertPS)); if arrivalrate==0,delayMQG(1,ll)=0;end
delayMQG(2,ll)=mean(userdelay(rtPS)); if arrivalrate==0,delayMQG(2,ll)=0;end
delayMQG(3,ll)=mean(userdelay(nrtPS)); if arrivalrate==0,delayMQG(3,ll)=0;end
delayMQG(4,ll)=mean(userdelay(BE)); if arrivalrate==0,delayMQG(4,ll)=0;end

%average system throughput and average user throughput
aveMQG=mean(MQGrate); systhrMQG(1,ll)=mean(sum(MQGrate(:,ertPS).'));
userthrMQG(1,ll)=mean(mean(MQGrate(:,ertPS))); if arrivalrate==0,systhrMQG(1,ll)=0;end
systhrMQG(2,ll)=mean(sum(MQGrate(:,rtPS).'));
userthrMQG(2,ll)=mean(mean(MQGrate(:,rtPS))); if arrivalrate==0,systhrMQG(2,ll)=0;end
systhrMQG(3,ll)=mean(sum(MQGrate(:,nrtPS).'));
userthrMQG(3,ll)=mean(mean(MQGrate(:,nrtPS))); if arrivalrate==0,systhrMQG(3,ll)=0;end
systhrMQG(4,ll)=mean(sum(MQGrate(:,BE).'));
userthrMQG(4,ll)=mean(mean(MQGrate(:,BE)));if arrivalrate==0,systhrMQG(4,ll)=0;end

%Jain's throughput fairness
```

```
tfMQG(1,ll)=JNfairnessN(aveMQG(ertPS),arrivalrate,zeros(1,length(ertPS)));
tfMQG(2,ll)=JNfairnessN(aveMQG(rtPS),arrivalrate,zeros(1,length(rtPS)));
tfMQG(3,ll)=JNfairnessN(aveMQG(nrtPS),arrivalrate,Rmin(nrtPS));
tfMQG(4,ll)=JNfairnessN(aveMQG(BE),arrivalrate,Rmin(BE));

%User satisfactin in percent
MQGcov(1,ll)= mean(mean(MQG_no_outage(:,ertPS))).*100;
if arrivalrate==0,MQGcov(1,ll)=100;end
MQGcov(2,ll)= mean(mean(MQG_no_outage(:,rtPS))).*100;
if arrivalrate==0,MQGcov(2,ll)=100;end
MQGcov(3,ll)= mean(mean(MQG_no_outage(:,nrtPS))).*100;
if arrivalrate==0,MQGcov(3,ll)=100;end
MQGcov(4,ll)= mean(mean(MQG_no_outage(:,BE))).*100;
if arrivalrate==0,MQGcov(4,ll)=100;end
%For DSM/TSM
userdelay=DSMHOLdelay(N,:);
delayDSM(1,ll)=mean(userdelay(ertPS)); if arrivalrate==0,delayDSM(1,ll)=0;end
delayDSM(2,ll)=mean(userdelay(rtPS)); if arrivalrate==0,delayDSM(2,ll)=0;end
delayDSM(3,ll)=mean(userdelay(nrtPS)); if arrivalrate==0,delayDSM(3,ll)=0;end
delayDSM(4,ll)=mean(userdelay(BE)); if arrivalrate==0,delayDSM(4,ll)=0;end
aveDSM=mean(DSMrate); systhrDSM(1,ll)=mean(sum(DSMrate(:,ertPS).'));
userthrDSM(1,ll)=mean(mean(DSMrate(:,ertPS))); if arrivalrate==0,systhrDSM(1,ll)=0;end
systhrDSM(2,ll)=mean(sum(DSMrate(:,rtPS).'));
userthrDSM(2,ll)=mean(mean(DSMrate(:,rtPS))); if arrivalrate==0,systhrDSM(2,ll)=0;end
systhrDSM(3,ll)=mean(sum(DSMrate(:,nrtPS).'));
userthrDSM(3,ll)=mean(mean(DSMrate(:,nrtPS))); if arrivalrate==0,systhrDSM(3,ll)=0;end
systhrDSM(4,ll)=mean(sum(DSMrate(:,BE).'));
userthrDSM(4,ll)=mean(mean(DSMrate(:,BE))); if arrivalrate==0,systhrDSM(4,ll)=0;end
tfDSM(1,ll)=JNfairnessN(aveDSM(ertPS),arrivalrate,zeros(1,length(ertPS)));
tfDSM(2,ll)=JNfairnessN(aveDSM(rtPS),arrivalrate,zeros(1,length(rtPS)));
tfDSM(3,ll)=JNfairnessN(aveDSM(nrtPS),arrivalrate,Rmin(nrtPS));
tfDSM(4,ll)=JNfairnessN(aveDSM(BE),arrivalrate,Rmin(BE));
```

```
DSMdroppedT(1,ll)= mean(DSMdroppedRate(N,ertPS)).*100;
if arrivalrate==0,DSMdroppedT(1,ll)=0;end
DSMdroppedT(2,ll)= mean(DSMdroppedRate(N,rtPS)).*100;
if arrivalrate==0,DSMdroppedT(2,ll)=0;end
DSMdroppedT(3,ll)= mean(DSMdroppedRate(N,nrtPS)).*100;
if arrivalrate==0,DSMdroppedT(3,ll)=0;end
DSMdroppedT(4,ll)= mean(DSMdroppedRate(N,BE)).*100;
if arrivalrate==0,DSMdroppedT(4,ll)=0;end
DSMcov(1,ll)= mean(mean(DSM_no_outage(:,ertPS))).*100;
if arrivalrate==0,DSMcov(1,ll)=100;end
DSMcov(2,ll)= mean(mean(DSM_no_outage(:,rtPS))).*100;
if arrivalrate==0,DSMcov(2,ll)=100;end
DSMcov(3,ll)= mean(mean(DSM_no_outage(:,nrtPS))).*100;
if arrivalrate==0,DSMcov(3,ll)=100;end
DSMcov(4,ll)= mean(mean(DSM_no_outage(:,BE))).*100;
if arrivalrate==0,DSMcov(4,ll)=100;end
clc
ll
end %End uu loop

%sample plot
figure (1)
plot( uuu,mean(systhrMQG(1:4,:)./1e+6),dcolor1);
hold on
plot( uuu,mean(systhrDSM(1:4,:)./1e+6),dcolor2);
hold on
legend('MQG','DSM'); title('Total average system Capacity');
xlabel('User arrival rate'); ylabel('system throughput (Mbps)');
grid
finish=toc(start)
```

C.2: Matlab code for the subprogram for MQS rule

```
function [NbitsDropped,Bffoverflow,MQGHOLdelay,MQG_outage,...
    MQGarriveT,MQGtotalRate,MQGsuballo,MQGrate,dlMQG,MQGaveRate,...
    MQGuser,MQGqdelay,aMQGwaitTime,MQG_no_outage,MQGdroppedRate,...
    MQGdropped,aThMQG,MQGqueueBegin,MQGqueues,aMQGqueues,...
    MQGallocated]=MQGalgorithmLKC(uu,users,K,B,D,To,N,arrival,Rmin,...
    Rmax,Beta,QoSclass,SNR,PacketSize,QueueSize,Ts,bw,latency,ap,rl,...
    Bffoverflow,MQGHOLdelay,MQG_outage,MQGarriveT,MQGrate,dlMQG,...
    MQGaveRate,MQGuser,MQGqdelay,aMQGwaitTime,MQG_no_outage,...

MQGdroppedRate,MQGdropped,aThMQG,MQGqueueBegin,MQGqueues,aMQGqueues,...
    MQGallocated,DRC,s,NbitsDropped,RA)
 if s==1
   MQGqueues(1,1:uu)=arrival(1:uu).*Ts;%Npackets;%
   aMQGqueues(1,1:uu)=(1-bw).*aMQGqueues(1,1:uu)+bw.*MQGqueues(1,1:uu);
   aMQGwaitTime(1,1:uu)=aMQGqueues(1,1:uu)./arrival(1:uu);    MQGqdelay(1,1:uu)=Ts;
end
MQGtotalRate=zeros(uu,K); MQGsuballo=zeros(uu,K); MQGactiveUser=ones(K,uu);
MQGqueueBegin(s,users)=MQGqueues(s,users); occupy=zeros(1,K);
unoccupy=find(not(occupy));            while            not(isempty(unoccupy))            &&
sum(MQGqueues(s,users))>0
 c=unoccupy(1);  MQGauxrateK=log2(1+SNR(c:K:K*uu)./Beta).*B/(D*K);
 %computing AMC for DRA
 SINR=20.*log10(SNR(c:K:K*uu));  ncodes=size(DRC,1);
 if RA==1
   for ul=1:uu
     for uj=1:ncodes
       if SINR(1,ul)>=DRC(uj,1) && SINR(1,ul)<DRC(uj,2),
          MQGauxrateK(1,ul)=DRC(uj,3).*B/(D*K);
       end
     end
```

```
  end
 %MQGauxrateK=2.*round(log2(1+SNR(c:K:K*uu)./Beta)./2).*B/(D*K);
end
%Utility function
wm=zeros(1,length(users));
for up=users
 if QoSclass(up)==1 || QoSclass(up)==2 %VoIP and Video/audio
       wm(up)=(exp(ap(up)*(MQGHOLdelay(s,up)-latency(up))))/...
            (exp(ap(up)*(MQGHOLdelay(s,up)-latency(up)))+exp(-ap(up)*...
                  (MQGHOLdelay(s,up)-latency(up))));
  elseif QoSclass(up)==3 % FTP
       wm(up)=(exp(-ap(up)*(MQGaveRate(s,up)-Rmin(up))))/...
          (exp(ap(up)*(MQGaveRate(s,up)-Rmin(up)))+exp(-ap(up)*...
                  (MQGaveRate(s,up)-Rmin(up))));
  elseif  QoSclass(up)==4 %BE
       wm(up)=(exp(-ap(up)*(MQGaveRate(s,up)-Rmin(up))))/...
         (exp(ap(up)*(MQGaveRate(s,up)-Rmin(up)))+exp(-ap(up)*...
             (MQGaveRate(s,up)-Rmin(up))));
  end
end
wmqg=wm(users);%
MQGauxrate=MQGactiveUser(c,users).*MQGauxrateK(users);
MQGauxqueue=MQGactiveUser(c,users).*MQGqueues(s,users)./Ts;
minMQG=min(MQGauxrate(users),MQGauxqueue(users));
[maxMQG,MQGindex]=max(wmqg.*minMQG);

if sum(minMQG)<=0,break;end
if minMQG(MQGindex)>0 && (MQGqueues(s,MQGindex)-minMQG(MQGindex)*Ts)>=0
  MQGtotalRate(MQGindex,c)=MQGtotalRate(MQGindex,c)+minMQG(MQGindex);
  MQGsuballo(MQGindex,c)=1;
  MQGqueues(s,MQGindex)=MQGqueues(s,MQGindex)-minMQG(MQGindex)*Ts;
  MQGallocated(s,MQGindex)=MQGallocated(s,MQGindex)+1;
```

```
 if MQGallocated(s,MQGindex)==1,MQGarriveT(s,MQGindex)=((c-1)/K)*Ts;end
  unoccupy(1)=[];
 else
  MQGactiveUser(c,MQGindex)=0;
end
if isempty(unoccupy)
 %Check if RT users have exceeded delay budget and drop HOL  packet
 for jj=users
  if QoSclass(jj)==1 || QoSclass(jj)==2
     if MQGallocated(s,jj)>0 && MQGHOLdelay(s,jj)>latency(jj)
        MQGqueues(s,jj)=MQGqueues(s,jj)+sum(MQGtotalRate(jj,:))*Ts;
        NbitsDropped(s,jj)=min(MQGqueues(s,jj),PacketSize(jj));
        MQGqueues(s,jj)=MQGqueues(s,jj)-NbitsDropped(s,jj);
       [a,b]=find(MQGtotalRate(jj,:)>0);
        MQGsuballo(jj,b)=0;MQGtotalRate(jj,b)=0;MQGactiveUser(b,jj)=0;
       unoccupy=[unoccupy b];MQGallocated(s,jj)=0;
     end
  end
 end
end
if  isempty(unoccupy)|| sum(sum(MQGactiveUser))==0, break;end %
end % while
% Compute performance metrics at the end of a time slot
for jj=1:uu
  MQGrate(s,jj)=sum(MQGtotalRate(jj,:));
  MQGuser(s,jj)=sum(MQGsuballo(jj,:));
  MQGaveRate(s+1,jj)=mean(MQGrate(1:s,jj));
  aThMQG(s+1,jj)=(1-bw)*aThMQG(s,jj)+bw*MQGrate(s,jj);
  MQGqueues(s+1,jj)=MQGqueues(s,jj)+arrival(jj)*Ts;
  aMQGqueues(s+1,jj)=(1-bw)*aMQGqueues(s,jj)+bw*MQGqueues(s+1,jj);
  % queue dropped bcos of buffer overflow
  Bffoverflow(s,jj)=min(0,QueueSize(jj)-MQGqueues(s+1,jj));
```

```
% queue dropped bcos of buffer overflow
MQGdropped(s+1,jj)=NbitsDropped(s,jj)-Bffoverflow(s,jj);
MQGdroppedRate(s+1,jj)=MQGdropped(s+1,jj)/(arrival(jj)*Ts);
MQGqueues(s+1,jj)=min(QueueSize(jj),MQGqueues(s+1,jj));
MQGHOLdelay(s+1,jj)=MQGHOLdelay(s,jj)+Ts-MQGrate(s,jj)*Ts/arrival(jj);
aMQGwaitTime(s+1,jj)=aMQGqueues(s+1,jj)/arrival(jj);
%compute deadline miss
dlMQG(s+1,jj)=(sum(MQGdropped(1:s,jj))/(sum(MQGdropped(1:s,jj))+...
 sum(MQGrate(1:s,jj)*Ts)));if isnan(dlMQG(s+1,jj)), dlMQG(s+1,jj)=0;end

%computer user call satisfaction
if QoSclass(jj)==3 || QoSclass(jj)==4
  if  MQGaveRate(s,jj)>=Rmin(jj)
      MQG_no_outage(s,jj)=1;
  end
elseif QoSclass(jj)==1 || QoSclass(jj)==2
  if  MQGHOLdelay(s,jj)<=latency(jj)
      MQG_no_outage(s,jj)=1;
  end
 end
end %end jj
```