Performance Analysis of Packet Scheduling Algorithms for Long Term Evolution (LTE)

A Thesis
submitted to
University of Technology, Sydney
by

Minjie Xue

In accordance with
the requirements for the Degree of

Master of Engineering (Research)

Faculty of Engineering and Information Technology
University of Technology, Sydney
New South Wales, Australia
August 2010

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

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

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

Signature of Candidate

Production Note:
Signature removed prior to publication.

ACKNOWLEDGMENT

First and foremost, I would like to express my deepest gratitude to my supervisor, A/Prof Kumbesan Sandrasegaran, who has given me great supports and valuable guidance throughout my research work in UTS. He gave me the opportunity to get involved into the research area of telecommunication and offered me valuable suggestions and criticisms with his profound knowledge and research experience.

I would like to extend my appreciation to Huda Adibah Mohd Ramli who has helped me to gain the fundamental knowledge within my research scope and given me many suggestions on my research.

I also would like to thank Riyaj Basukala and Rachod Patachaianand, who have given me a lot of comments to improve my research work.

I would like to thank my fellow officemates: Cheng-Chung Lin, Leijia Wu and Lu Chen, who have given me a lot of encouragement and made the office a very pleasant and comfortable place.

Finally, I would like to give grateful thank to my parents, Jianyou Xue and Lili Wu, who has given me the excellent support and encouragement to achieve the best education.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The third generation partnership project long term evolution (3GPP LTE) system is proposed as a new radio access technology in order to support high-speed data and multimedia traffic. The 3GPP LTE system has a flat radio access network architecture consisting of only one node, known as eNodeB, between user and core network. All radio resource management (RRM) functions are performed at the eNodeB. As one of the essential RRM functions, packet scheduling is responsible for the intelligent allocation of radio resources for active users. Since there is a diversity of the traffic types in wireless systems, active users may have different Quality of Service (QoS) requirements. In order to satisfy various QoS requirements and efficiently utilize the radio resources, a packet scheduler adopts a specific packet scheduling algorithm when making decisions. Several packet scheduling algorithms have been proposed in the literature.

The objective of this thesis is to evaluate the performance of the well-known and some recently proposed packet scheduling algorithms and identify the suitability of these algorithms in the downlink LTE system. The performance evaluation of packet scheduling algorithms based on both computer simulation and theoretical analysis is provided in this thesis.

The performance of packet scheduling algorithms is evaluated in three scenarios including 100% RT scenario, 100% NRT scenario and 50% RT and 50% NRT scenario under the downlink LTE simulation environment. The simulation results for well-known packet scheduling algorithms show that Maximum-Largest Weighted Delay First (M-LWDF) outperforms other algorithms in the 100% RT scenario, while Exponential/Proportional Fair (EXP/PF) is comparatively more suitable in the 50% RT and 50% NRT scenario. In the 100% NRT scenario, Proportional Fair (PF) and Maximum Rate (Max-Rate) achieve a good throughput and resource block (RB)

utilization performance while Round Robin (RR) has the best fairness performance. Additionally, two recently proposed algorithms are evaluated and can be considered as the packet scheduling candidates. The simulation results show that Sun Qiaoyun's Algorithm is more appropriate than Jeongsik Park's Algorithm for the downlink LTE supporting the real-time traffic.

The mathematical model for performance evaluation of the packet scheduling algorithms in the downlink LTE system is discussed in this thesis. The theoretical delay analysis for OFDMA system and the theoretical throughput analysis of PF algorithm is studied and validated in detail. This thesis moves further to theoretical performance analysis of M-LWDF and obtains the analytical result of the expected throughput of M-LWDF.

Chapter 1

INTRODUCTION

1.1 Brief History

Based on the successful deployment of Global System for Mobile Communications (GSM), the Third Generation Partnership Project (3GPP) standardization body finalized the specification of Universal Mobile Telecommunications System (UMTS) in the first 1999 release. As the air interface technology of UMTS, the wideband code-division multiple access (WCDMA) technology along with high-speed packet access (HSPA) technology provided 3GPP with a highly competitive radio access technology. WCDMA/HSPA is being widely deployed all over the world and has become a leading third generation (3G) technology.

However, with the increasing requirements and expectations from users and emergence of competing radio access technologies [2], such as IEEE 802.16 (WiMAX) standard, it is crucial for 3GPP to enhance the existing WCDMA/HSPA technology, in order to maintain the competitiveness in the market. Consequently, 3GPP proposed the Evolved UMTS Terrestrial Radio Access Network (E-URTAN) Long Term Evolution (LTE), which aims

> "to develop a frame work for the evolution of the 3GPP radio-access technology towards a high-data-rate, low-latency and packet-optimized radio-access technology [1]".

The 3GPP releases are shown in Figure 1-1.

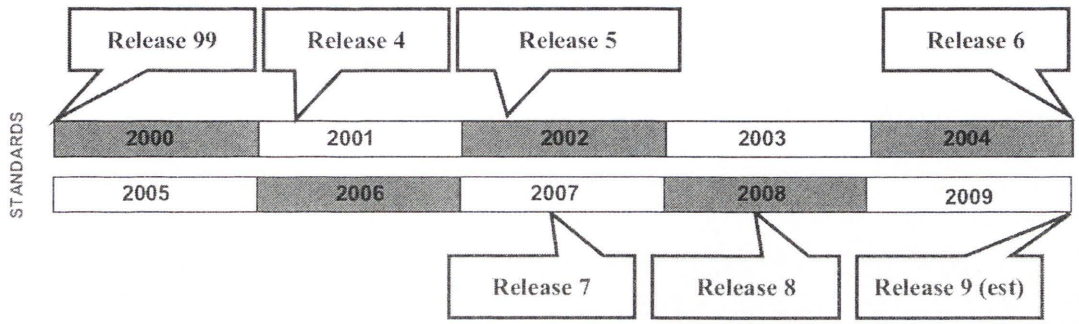

Figure 1-1 History of the 3GPP Releases [3]

Figure 1-2 illustrates the evolution of the 3GPP standards. As shown, with the development of 3GPP technologies, the peak data rate of wireless systems has been greatly improved and LTE is expected to achieve a significant throughput enhancement compared with the earlier communication networks. The maximum speed of GPRS first launched in 1998 was 40 kbps. WCDMA 2002 can support up to 384 kbps while HSPA and HSPA Evolution supported 3.6-14.4 Mbps and 21-42 Mbps, respectively. The current LTE offers 150 Mbps, which is more than 3000 times the data rate achievable 10 years ago.

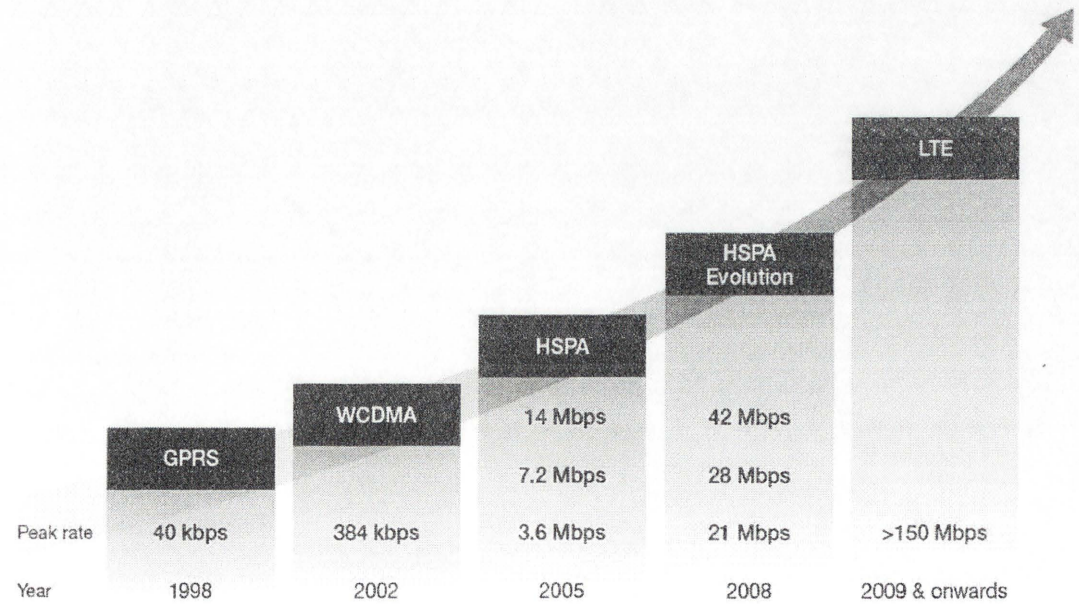

Figure 1-2 Evolution of 3GPP Technologies [4]

The evolution of 3GPP radio access technologies is given in Figure 1-3. After evaluations of the existing radio access technologies in mid-1980s, TDMA has been chosen for GSM, followed by WCDMA and HSPA that were built on CDMA technology. The recently proposed LTE system adopts OFDMA for the downlink transmission.

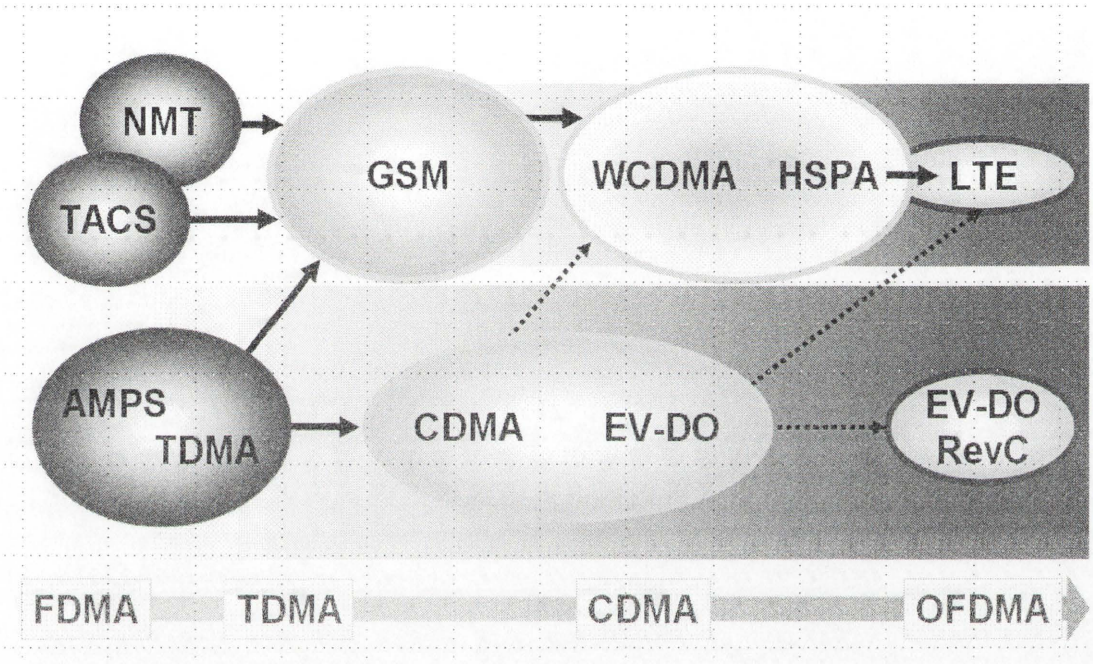

Figure 1-3: Development of 3GPP Radio Access Technologies[5]

1.2 LTE

LTE proposed by 3GPP brings significant improvements to 3G mobile systems. LTE not only provides a significant evolution in radio access technologies, but also uses simplified network architecture. Both aspects will be further discussed in Chapter 2.

LTE adopts different radio access schemes for the downlink direction and uplink direction. Orthogonal Frequency Division Multiple Access (OFDMA) is used for the downlink transmission, while Single Carrier Frequency Division Multiple Access (SC-FDMA) technology has been chosen for the uplink transmission.

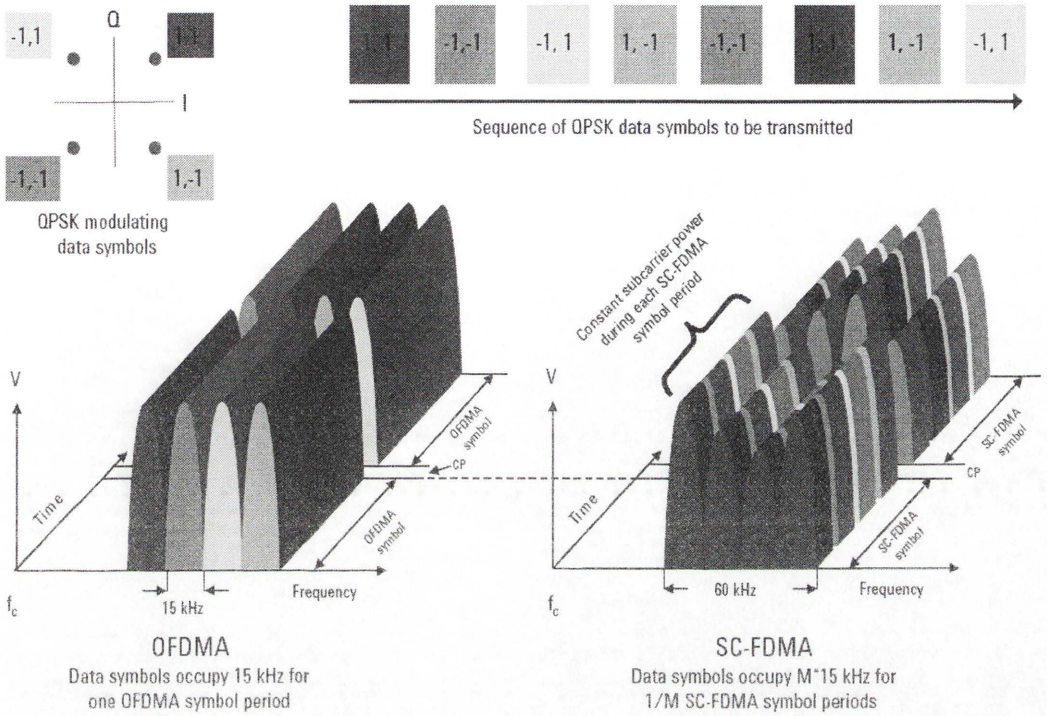

Figure 1-4: Difference between OFDMA and SC-FDMA for the Transmission of a Sequence of QPSK Data Symbols [6]

A comparison of OFDMA and SC-FDMA for the transmission of a sequence of Quadrature Phase Shift Keying (QPSK) data symbols is given Figure 1-4. In both schemes, the available bandwidth is divided into M consecutive 15 kHz subcarriers, and the same symbol length of 66.7 μsecs is used. The cyclic prefix (CP) is inserted every 66.7 μs. For OFDMA, all subcarriers will be allocated to different data symbols and every M data symbols among the data symbol sequence will be transmitted at the same time. On the contrary, SC-FDMA will allocate the whole bandwidth to just one data symbol at each point of time and data symbols will be transmitted sequentially. Correspondingly, the transmission time of each data symbol for SC-FDMA is 1/M of that for OFDMA. Moreover, OFDMA uses the same transmission power for all subcarriers while transmission power on each subcarrier might be different for SC-FDMA.

Compared with other technologies used in earlier mobile networks, e.g. Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA) and Code

Division Multiple Access (CDMA), OFDMA can support higher data rates, multi-user diversity and provide good performance in frequency selective fading channels. A detailed discussion will be given in Section 2.2.

1.3 Wireless Spectrum

Spectrum is the most critical physical resource in cellular communications. Consequently, the spectrum allocation has drawn a lot of attention. As an emerging technology, a key design goal of LTE is to support spectrum flexibility. Accordingly, LTE should enable the deployment in various spectrum environments in terms of duplex mode, frequency bands as well as achievable bandwidths.

LTE is able to support both paired spectrum allocation and unpaired spectrum allocation. Figure 1-5 illustrates the operation of both paired and unpaired spectrum allocation. The paired spectrum allocation uses one frequency band for the uplink direction and another frequency band for the downlink direction; while for the unpaired spectrum, the same frequency band is employed for both uplink transmission and downlink transmission at different times. Therefore, both Time Division Duplex (TDD) and Frequency Division Duplex (FDD) will be operated by LTE to fully make use of the paired and unpaired spectrum.

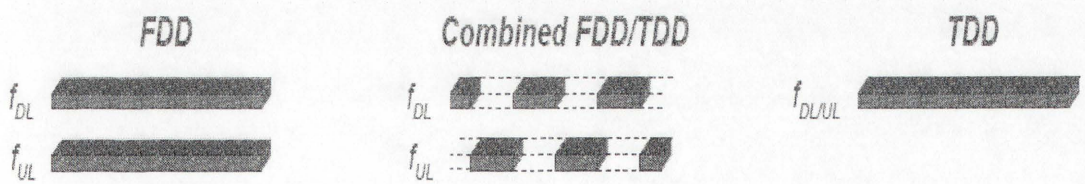

Figure 1-5 FDD/TDD in Paired and Unpaired Spectrum Allocation [7]

LTE enables operation in different frequency bands. Figure 1-6 gives the LTE operating bands in 3GPP specifications. There are 15 FDD bands and 8 TDD bands. More operating bands will be added with the standardization process.

E-UTRA operating band	Uplink (UL) operating band BS receive UE transmit	Downlink (DL) operating band BS transmit UE receive	Duplex mode
	$F_{UL\_low} - F_{UL\_high}$	$F_{DL\_low} - F_{DL\_high}$	
1	1920 – 1980 MHz	2110 – 2170 MHz	FDD
2	1850 – 1910 MHz	1930 – 1990 MHz	FDD
3	1710 – 1785 MHz	1805 – 1880 MHz	FDD
4	1710 – 1755 MHz	2110 – 2155 MHz	FDD
5	824 – 849 MHz	869 – 894 MHz	FDD
6	830 – 840 MHz	875 – 885 MHz	FDD
7	2500 – 2570 MHz	2620 – 2690 MHz	FDD
8	880 – 915 MHz	925 – 960 MHz	FDD
9	1749.9 – 1784.9 MHz	1844.9 – 1879.9 MHz	FDD
10	1710 – 1770 MHz	2110 – 2170 MHz	FDD
11	1427.9 – 1452.9 MHz	1475.9 – 1500.9 MHz	FDD
12	698 – 716 MHz	728 – 746 MHz	FDD
13	777 – 787 MHz	746 – 756 MHz	FDD
14	788 – 798 MHz	758 – 768 MHz	FDD
...			
17	704 – 716 MHz	734 – 746 MHz	FDD
...			
33	1900 – 1920 MHz	1900 – 1920 MHz	TDD
34	2010 – 2025 MHz	2010 – 2025 MHz	TDD
35	1850 – 1910 MHz	1850 – 1910 MHz	TDD
36	1930 – 1990 MHz	1930 – 1990 MHz	TDD
37	1910 – 1930 MHz	1910 – 1930 MHz	TDD
38	2570 – 2620 MHz	2570 – 2620 MHz	TDD
39	1880 – 1920 MHz	1880 – 1920 MHz	TDD
40	2300 – 2400 MHz	2300 – 2400 MHz	TDD

Figure 1-6 Operating Bands of E-UTRAN [6]

Some of the frequency bands are used by other technologies, e.g. the 1800 and 1900 MHz frequency bands for GSM in Europe as well as in Asia. The spectrum allocation of the International Mobile Telecommunications-2000 (IMT-2000) is given in Figure 1-7.

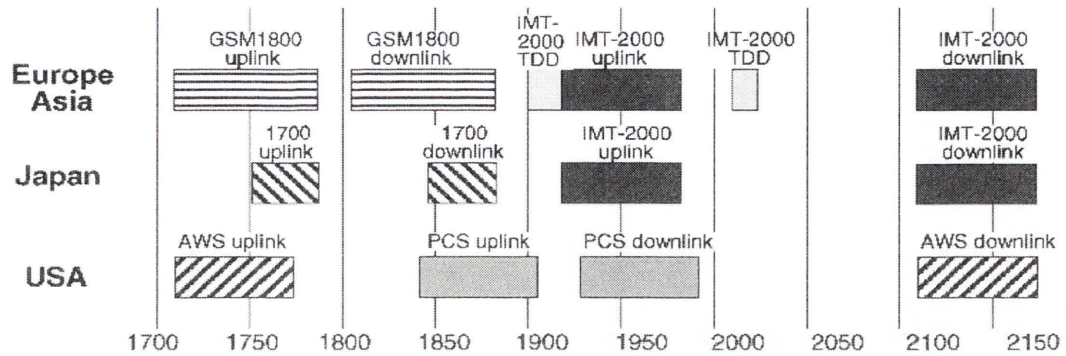

Figure 1-7 Spectrum Allocation of IMT-2000 [8]

LTE can coexist with these technologies. LTE will likely start by deploying within the newly released 2.6 GHz frequency band and refarming to the existing 900 and 1800 MHz bands [3].

As inter-networking with other 3G systems is an important requirement, LTE allows bandwidth flexibility. The narrowband spectrum allocation is quite flexible in the LTE system [1]. As the example given in Figure 1-8, the spectrum allocation for LTE system can begin with a small bandwidth and be increased gradually with the growing number of users switching to LTE system [9]. Transmissions for LTE can be operated in the bandwidth within a range of 1.25 MHz to 20 MHz [1]. The bandwidth flexibility in LTE system is illustrated in Figure 1-9.

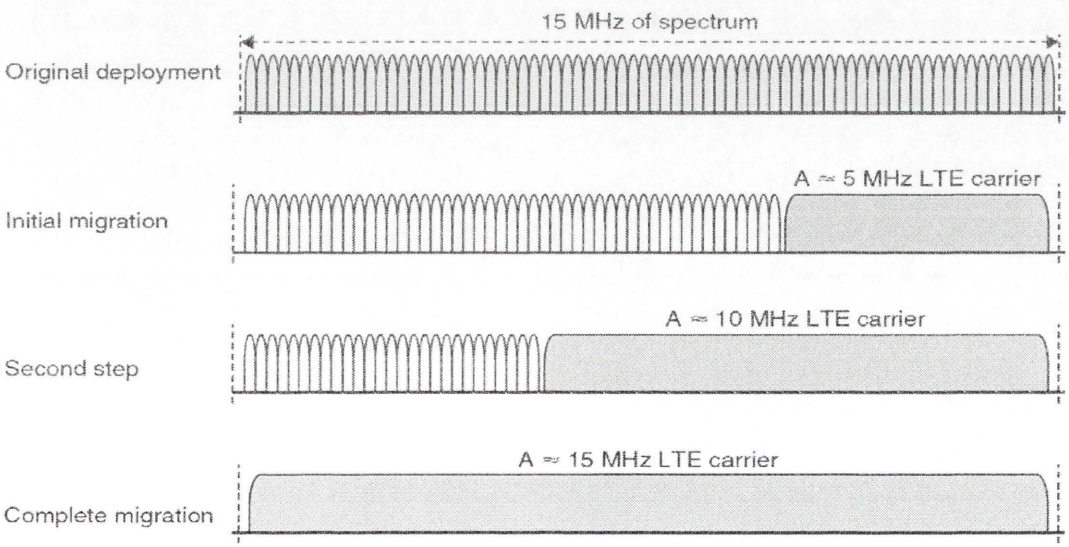

Figure 1-8 Migration of Spectrum Allocation from GSM Deployment to LTE [10]

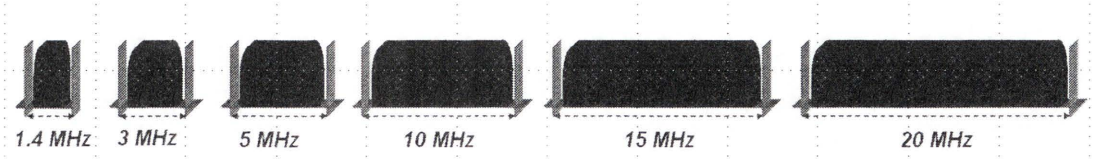

Figure 1-9 Bandwidth Flexibility in LTE System [5]

1.4 Packet Scheduling in Downlink LTE System

Due to the scarcity of frequency resources, a set of mechanisms, known as Radio Resource Management (RRM), is designed to optimize the efficient usage of the limited radio spectrum resources. RRM schemes include admission control, power control, congestion control, packet scheduling (PS), handover control and link adaptation. Packet scheduling, as one of the most important RRM functions of LTE, is the emphasis of this thesis.

Packet scheduling is responsible for the intelligent allocation of radio resources for active users. Active users refer to users with packets waiting in the buffer and competing for transmission. Since there is a diversity of the traffic types in wireless systems, active users may have different Quality of Service (QoS) requirements. In order to satisfy various QoS requirements and efficiently utilize the radio resources, a packet scheduler adopts a specific packet scheduling algorithm when making decisions. The discussion of Packet scheduling algorithms will be given in detail in Chapter 3.

Figure 1-10 illustrates a generalized PS model for the downlink LTE system. Each active user will be allocated one buffer within the eNodeB. Packet scheduler will allocate the available radio resources to the active users based on certain scheduling criteria. The scheduling criteria may take various factors into consideration, such as channel condition, amount of packets waiting in the user's buffer, delay of the waiting packets, type of services and so on.

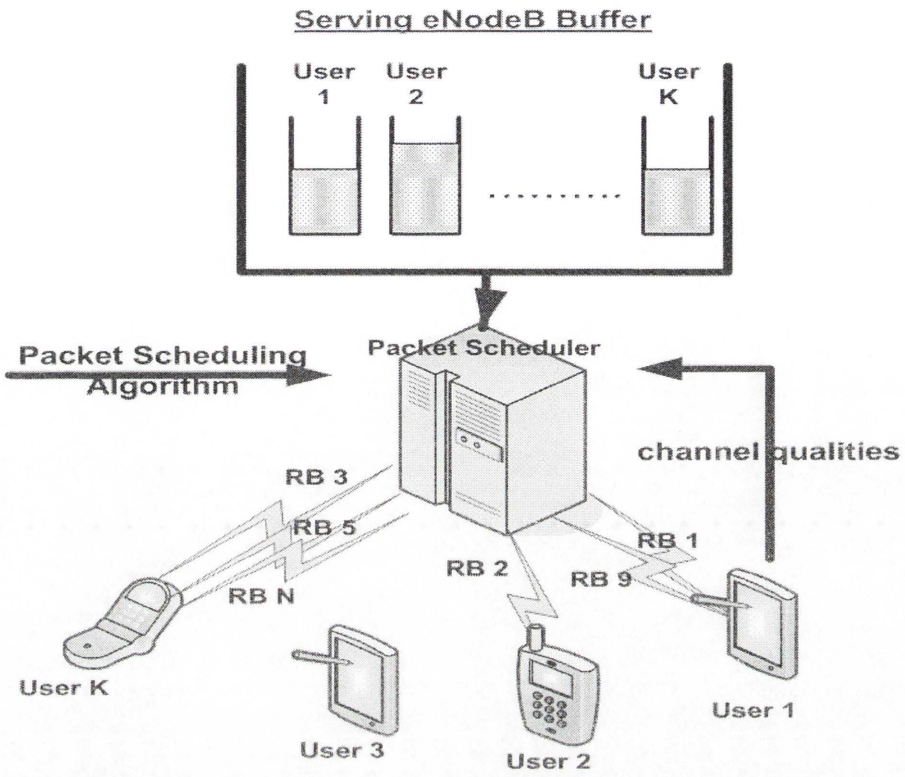

Figure 1-10 Generalized PS Model for Downlink LTE System[9]

1.5 Problem Statements and Research Objectives

Packet scheduling in downlink LTE system is an important area of research and has attracted much research interests. Several packet scheduling algorithms have been proposed for the downlink LTE system. However, most of the performance evaluations of these packet scheduling algorithms are based on simulation results and very little work related to theoretical performance analysis has been published in the literature.

The objectives of the thesis are given as follows:

- To model, simulate, validate and evaluate current well-known and new packet scheduling techniques for the 3GPP LTE.
- To develop analytical/mathematical models of the performance of packet scheduling algorithms and compare it with the simulation results.
- To identify the suitability of various packet scheduling algorithms.

1.6 Thesis Outline

This thesis is organized as follows:

Chapter 1 gives an introduction of this thesis. The brief information of cellular systems, LTE, wireless spectrum and packet scheduling in the downlink LTE system is provided. The problem statement and research objectives are given in this chapter.

In Chapter 2, background knowledge of LTE network architecture, the resource block and OFDMA technology is discussed in detail. After that, an introduction of six main Radio Resource Management (RRM) mechanisms and the radio propagation model used in this thesis is given.

Chapter 3 reviews a number of packet scheduling algorithms and discusses five performance metrics that are designed for the performance analysis of these algorithms. The performance of these packet scheduling algorithms is evaluated under the downlink LTE simulation environment and the performance comparison of these algorithms is given at the end of this chapter.

Theoretical delay analysis of the OFDMA system with Voice-over-IP (VoIP) traffic is discussed in Chapter 4. The Hybrid-Automatic Repeat Request (HARQ) is employed to improve system performance. A brief introduction of VoIP and HARQ is provided. The analytical model for delay is divided into two levels: the talk spurt level and the voice packet level. The analysis of both levels is explained in detail. The simulation result of talk spurt assignment latency distribution $F_{wb}(t)$ is provided.

Chapter 5 discusses the theoretical throughput analysis of packet scheduling algorithms. After the step-by-step derivations, the mathematical expressions of the expected throughput for proportional fair (PF) algorithm and M-LWDF algorithm are obtained. The visualisation results for the throughput analysis of both algorithms are provided.

Chapter 6 concludes this thesis and provides the plans for future research work.

1.7 Original Contribution

The following contributions included in this thesis are considered original.

Chapter 3

- Performance comparison of eight packet scheduling algorithms under three different simulation scenario.

Chapter 5

- New theoretical throughput analysis model of M-LWDF algorithm by combining the existing analysis model of throughput which was proposed for PF algorithm and the analytical approaches for delay.

1.8 Publication

The following conference paper has been published based on the contributions included in this thesis.

I. M. Xue, K. Sandrasegaran, H. A. Mohd Ramli, and C.-C. Lin, "Performance Analysis of Two Packet Scheduling Algorithms in Downlink 3GPP LTE System," in 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops Perth, Australia, 2010, pp. 915-919.

Summary: This paper evaluates the performance of two simple packet scheduling algorithms for real-time traffic in the third generation partnership project long term evolution (3GPP LTE) system. These algorithms, known as Algorithm 1 and Algorithm 2 for this paper, were proposed to support real-time traffic in orthogonal frequency division multiple access (OFDMA) system. Simulation results show that Algorithm 1 outperforms Algorithm2 by achieving a lower packet delay and packet loss rate while having almost similar throughput and fairness performance.

Chapter 2

BACKGROUND

This chapter gives background knowledge of LTE system. Three aspects of LTE are discussed which include the network architecture, the minimum resource allocation unit (resource block) and radio access technology. Thereafter, introductions of the radio resource management (RRM) mechanisms and the radio propagation model used in this thesis are provided.

2.1 LTE Architecture

As LTE is designed as the packet-optimizing technology which requires the seamless network connectivity, a flat radio access network architecture with less evolved nodes is adopted by the LTE system so that the network latencies can be reduced. The radio access network for LTE is known as Evolved-UTRAN (E-UTRAN) which comprises of only one node, known as eNodeB, between user and core network.

The network architecture comparison between UTRAN and E-UTRAN is given in Figure 2-1. Previously, NodeBs in UMTS were connected via the Radio Network Controller (RNC) that is responsible for NodeB management and radio resource allocation. As shown in the figure, LTE architecture has omitted RNC. Instead most of the RNC functions are now performed by the eNodeB which is directly connected to the core network. Additionally, E-UTRAN supports the interfaces (X_2) between eNodeBs, which facilitate the execution of some radio related functions such as handover preparation, interactions with neighbouring eNodeBs and so on. Therefore, LTE has a much more simplified network architecture than UMTS.

As shown in Figure 2-1, E-UTRAN is connected to evolved packet core (EPC) which is the packet-only core network of LTE via interface S1. Specifically, interface S1 connects eNodeBs to mobility management entity (MME) and serving gateway (S-GW) / packet data network (PDN) gateway (P-GW). MME is the control node which is responsible for functions related to bearer management and connection management [11, 12]. S-GW and P-GW are the gateways that terminate the packet data interface towards E-UTRAN and PDN, respectively.

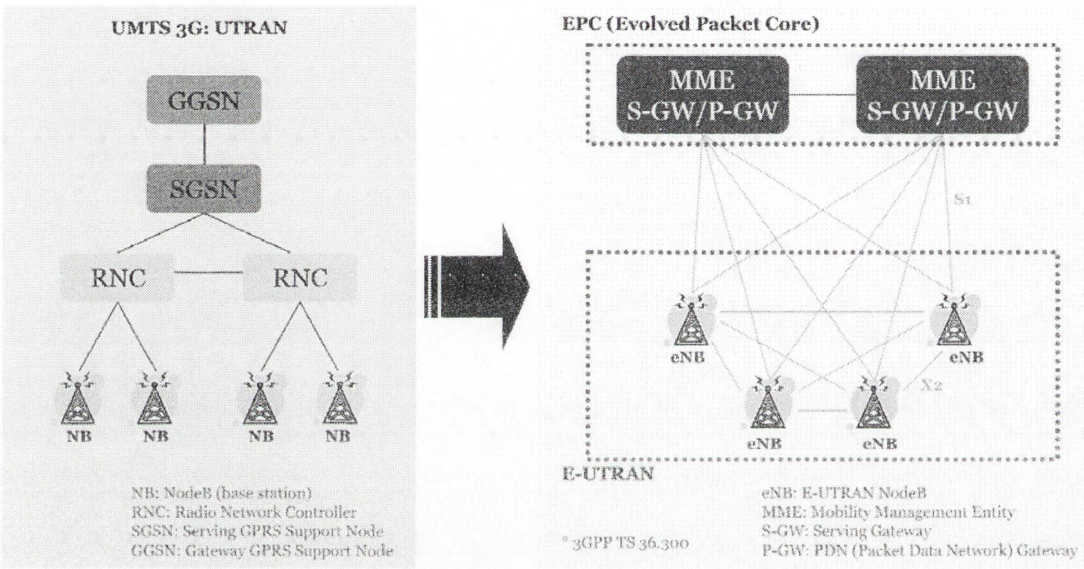

Figure 2-1: Network Architectures of UTRAN and E-UTRAN [4]

2.2 Resource Block

The minimum radio resource allocation unit in LTE is defined as the Resource Block (RB).

The RB in downlink LTE system is illustrated in Figure 2-2. The RB consists of both frequency domain and time domain. In frequency domain, every 12 consecutive sub-carriers, with total bandwidth of 180 kHz, are grouped as one sub-band. Furthermore, one sub-band and one time slot of 0.5 ms duration serve as a RB. A time slot contains either 6 or 7 OFDM symbols, depending on whether long or short Cyclic Prefix (CP) is used. Therefore, each RB contains $12 \times 7 = 84$ radio resource elements when normal CP is used.

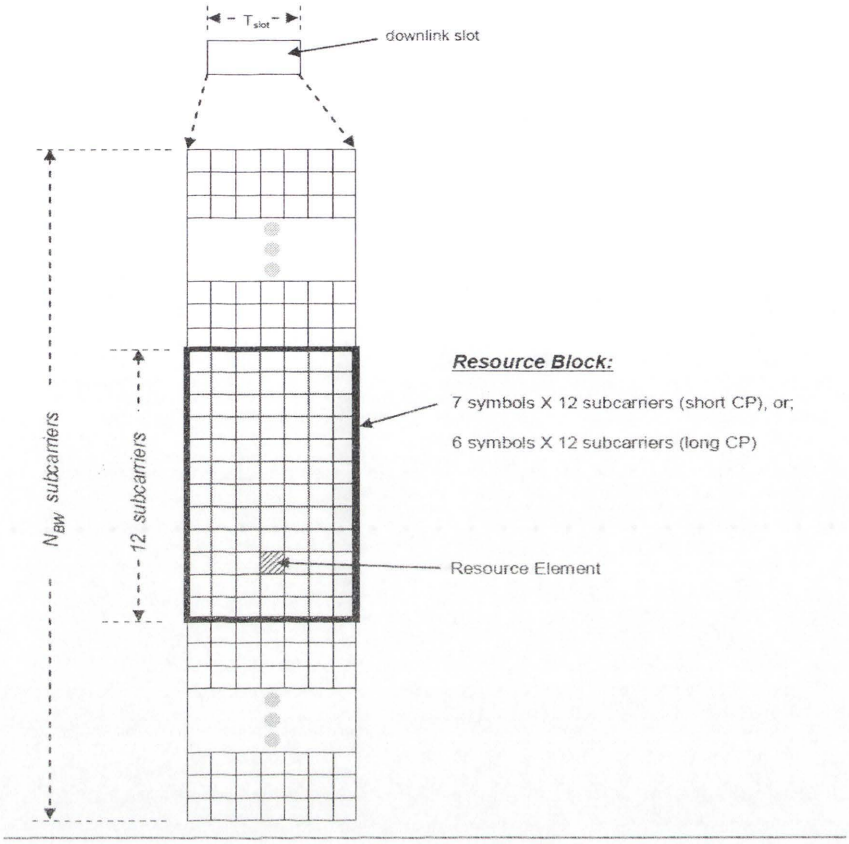

Figure 2-2: The Downlink LTE Resource Block [13]

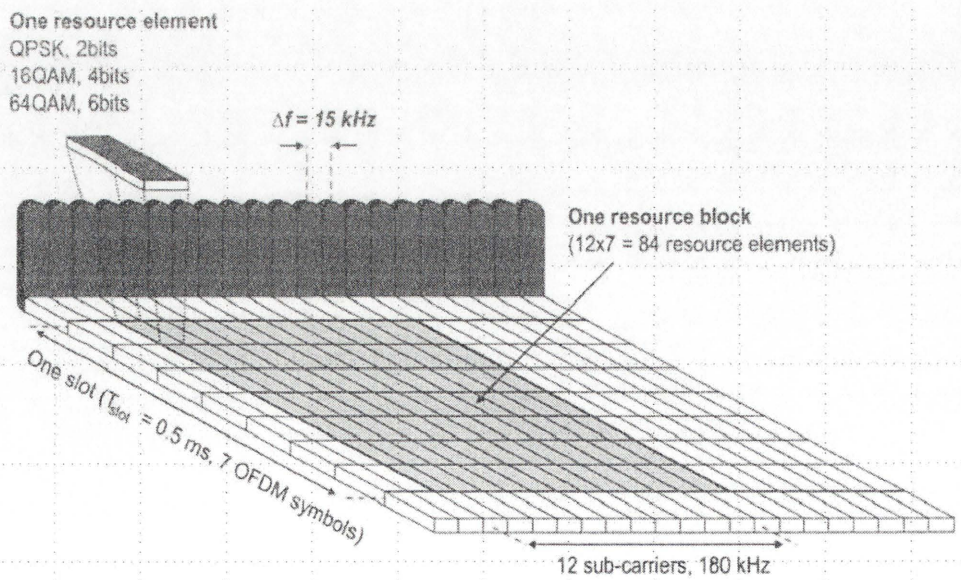

Figure 2-3: Radio Resource Block for the Downlink LTE [5]

Figure 2-3 gives a graphical representation of the downlink LTE RB.

The PS for the downlink 3GPP LTE system is operated at the Transmit Time Interval (TTI), which comprises two time slots and makes up an interval with 1 ms duration. In each TTI for each sub-band, the packet scheduler assigns two consecutive RBs in time domain to one user, as illustrated in Figure 2-4.

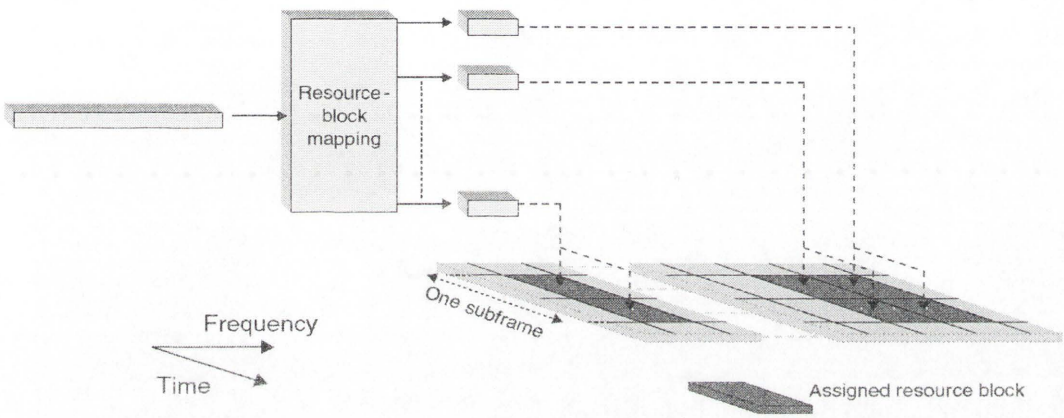

Figure 2-4: RB Assignment for the Downlink LTE [10]

The number of RBs in the downlink 3GPP LTE is determined by the available downlink system bandwidth, as shown in Table 2-1.

Table 2-1. Number of Available RBs Depending on Downlink Bandwidth [6]

Bandwidth (MHz)	1.25	3	5.0	10	15.0	20.0
Number of available RBs	6	15	25	50	75	100
Sub-carrier bandwidth (kHz)	15					
RB bandwidth (kHz)	180					

2.3 OFDMA

OFDMA technology is a variant of Orthogonal Frequency Division Multiplex (OFDM) technology. OFDM divides the system radio resource (bandwidth) into multiple narrowband orthogonal subcarriers with equal frequency spacing. A subcarrier spacing

of 15 kHz is adopted by LTE. The subcarriers' orthogonality ensures that at a sampling point for each single subcarrier, all the other subcarriers have zero crossings, as shown in Figure 2-5.

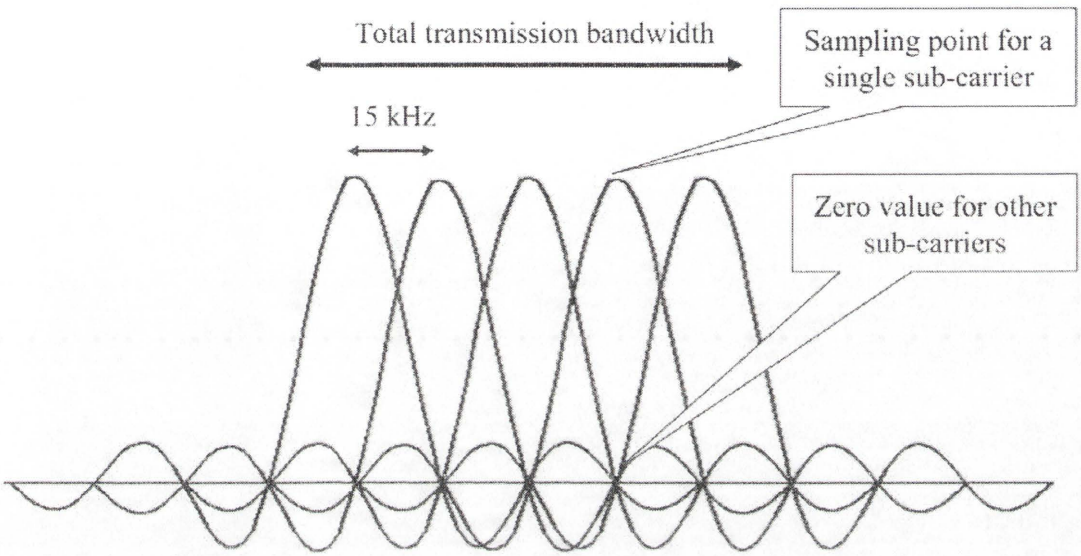

Figure 2-5: Maintaining the Subcarriers' Orthogonality [3]

Figure 2-6 illustrates an OFDM symbol given in both frequency domain and time domain. Guard intervals are inserted between each of the symbols in time domain so as to combat the inter-symbol interference caused by the delay spread of multi-path channels [11].

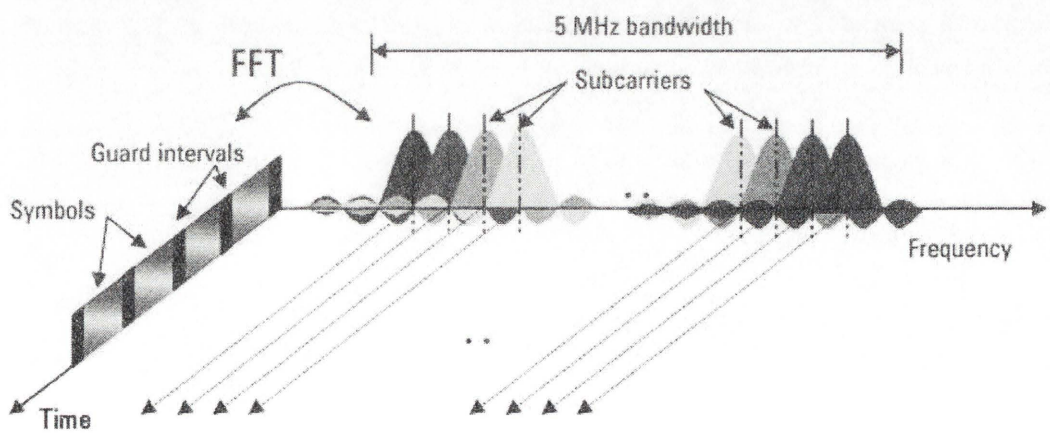

Figure 2-6: OFDM Symbol in both Frequency Domain and Time Domain [6]

Figure 2-7 illustrates the difference between OFDM and OFDMA subcarrier allocation. As shown, OFDM assigns each subcarrier to one specific user for the duration of a session while OFDMA allows subsets of subcarriers to be allocated dynamically among different users at each time interval, as TDMA technology is embedded into OFDMA. Due to time-domain statistical multiplexing, OFDMA further improves the OFDM robustness to frequency-selective fading and interference [6].

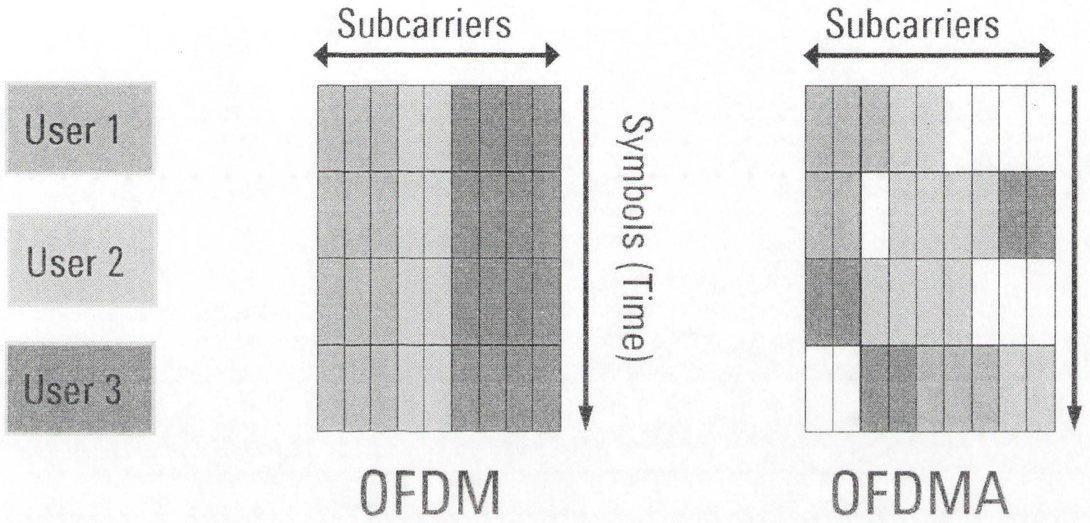

Figure 2-7: A Comparison of OFDM and OFDMA [6]

OFDMA, rather than WCDMA and TDMA, is chosen as the radio access technology of the downlink LTE system due to the following properties [3, 6, 11, 12, 14]:

- The orthogonality between narrow band subcarriers ensures a high spectral efficiency.
- The introduce of guard intervals between symbols can remove the delay spread of multi-path channels so that the inter-symbol interference can be limited.
- As OFDMA signals are represented in the frequency domain rather than in the time domain, OFDMA requires a much simpler base-band receiver than other technologies.

- It is easy to cooperate OFDMA with advanced receiver and antenna technologies, e.g. multiple-input and multiple-output (MIMO) technology, which will further enhance the throughput performance and spectral efficiency.

2.4 Radio Resource Management

Radio Resource Management (RRM) is a set of mechanisms designed to optimize the efficient usage of the limited radio spectrum resources.

RRM is required by the 3G systems to guarantee the target Quality of Service (QoS), maximize the system efficiency, maintain the planned coverage area and offer high capacity. These objectives may be contradictory and trade-offs have to be made. As shown in the following figure, radio network planning (RNP) offers the rough tuning of the objectives while the introduction of RRM enables a perfect match of these objectives.

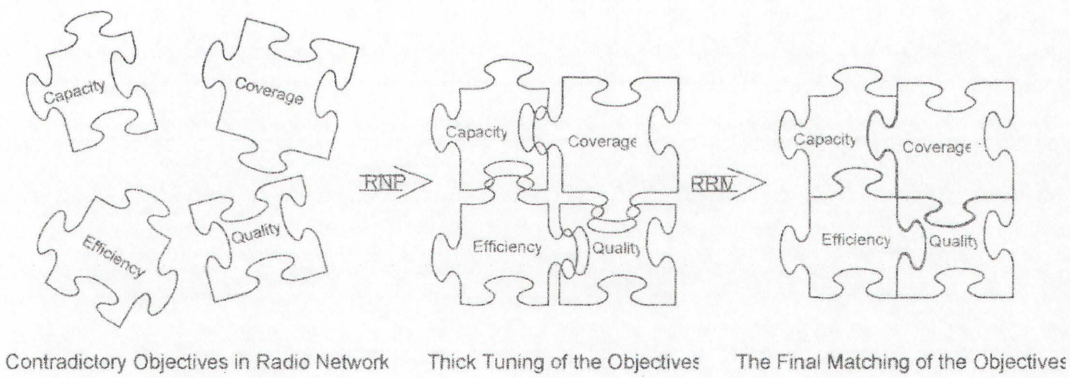

Figure 2-8 Objectives of Quality of Service [15]

RRM schemes include admission control, congestion control, handover control, packet scheduling, power control and link adaptation. Every RRM mechanism will be discussed in more details in the following sections.

2.4.1 Admission Control

Admission control decides whether a new call request will be admitted or rejected.

When there is a connection establishment request, the admission controller will make the admission decision based on the available resources and quality requirement of ongoing connections. The QoS requirement of a connection includes average data rate, Eb/N_0, etc. In UMTS, If the QoS requirement of the new request can be satisfied and QoS of the ongoing connections will not be degraded below planned levels by the admission of the new request, the new connection request can be admitted; otherwise, it will be rejected.

2.4.2 Congestion Control

Congestion control, also referred to as load control, is used to prevent the system from getting overloaded. If the overload occurs, congestion controller is also responsible for network recovery from the congested situation.

Some possible actions can be adopted to achieve this. The first case is that some connections in the congested cell will be handed over to a neighbouring cell with lower traffic load than the current cell. Additionally, most of the new connection admission requests may be blocked when the congestion happens. In WCDMA, another possible action is to reduce the transmission powers, as it will lead to a decrease of transmission data rate. These actions can reduce the load placing on the congesting cell and ensure the stability of the network.

2.4.3 Handover Control

Handover is the process of switching the service provision for a mobile user from one cell to another or from one system to another.

Handover can be triggered for many reasons [16]. First, handover can be performed to deal with the mobility of the users. The main target of handover is to ensure that the connection of a user can be maintained with a guaranteed QoS when the user is moving from the coverage of one cell to that of another. Second, a user may be handed over to neighbour cells when the current cell reaches its maximum capacity or is overloaded. Third, handover may be triggered if the user is switching between networks with different services, such as handover between WCDMA and GSM 900/1800, handover between WCDMA/FDD and WCDMA/TDD and so on.

Handover can be generally classified into two categories: Hard Handover and Soft Handover. For the hard handover, the connection to the current cell is released before the user is connected to the target cell. Connecting with at most one cell simplifies the design of the handset and makes it cheaper. On the contrary, soft handover enables user's connection to the target cells while retaining the connection with current cell. The user can connect to multiple cells simultaneously. Connection to a certain cell will be dropped if the received signal level from this cell is lower than a given threshold. Signals from all connecting cells will be combined to provide a better quality of connection. Thus, soft handover enhances the reliability of connection. All handovers in LTE are hard handovers.

2.4.4 Packet Scheduling

As discussed in Section 1.4, packet scheduling is responsible for the intelligent allocation of radio resources for active users. Packet scheduling is introduced to support various types of services with different QoS requirements and efficiently utilize the radio resources. Several packet scheduling algorithms were proposed to facilitate the allocation of radio resources. Further discussion of packet scheduling algorithms will be given in Chapter 3.

2.4.5 Power Control

Power control is a strategy used to optimise the level of transmission power in order to improve capacity, coverage, and received user quality and decrease interference.

On the one hand, higher transmission power for a specific user brings better performance to the user, such as higher signal-to-noise ratio (SNR), lower bit error rate, greater spectrum efficiency, etc. On the other hand, increasing transmission power will raise the overall transmission power consumption as well as the interference to the other users in the same frequency band.

Power control is designed to ensure that with the selected transmission power the receiver will have an adequate signal level for its requirement without creating unnecessary amount of interference.

2.4.6 Link Adaptation

Link adaptation is a technique used to make the most of instantaneous channel quality [5].

Figure 2-9 illustrates the model of link adaptation. The receiver reports the channel conditions to the transmitter. Then according to the receiver's feedback, the transmitter adjusts system parameters in order to match the current channel conditions. Two parameters that can be adjusted are transmit power and Modulation and Coding Scheme (MCS). The corresponding link adaptation mechanisms are known as power control and rate control, respectively.

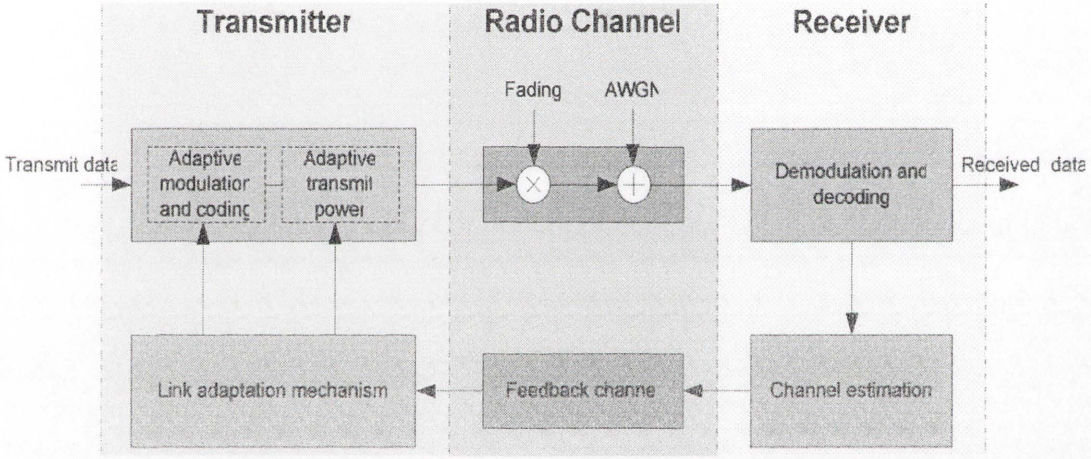

Figure 2-9 Model of Link Adaptation [15]

A comparison between power control and rate control is given in Figure 2-10. Power control adjusts the transit power to combat the channel fading and maintains a designed data rate regardless of channel qualities; rate control keeps the transmit power at a constant level and adjusts the data rate by choosing the appropriate MCS which depends on channel variations.

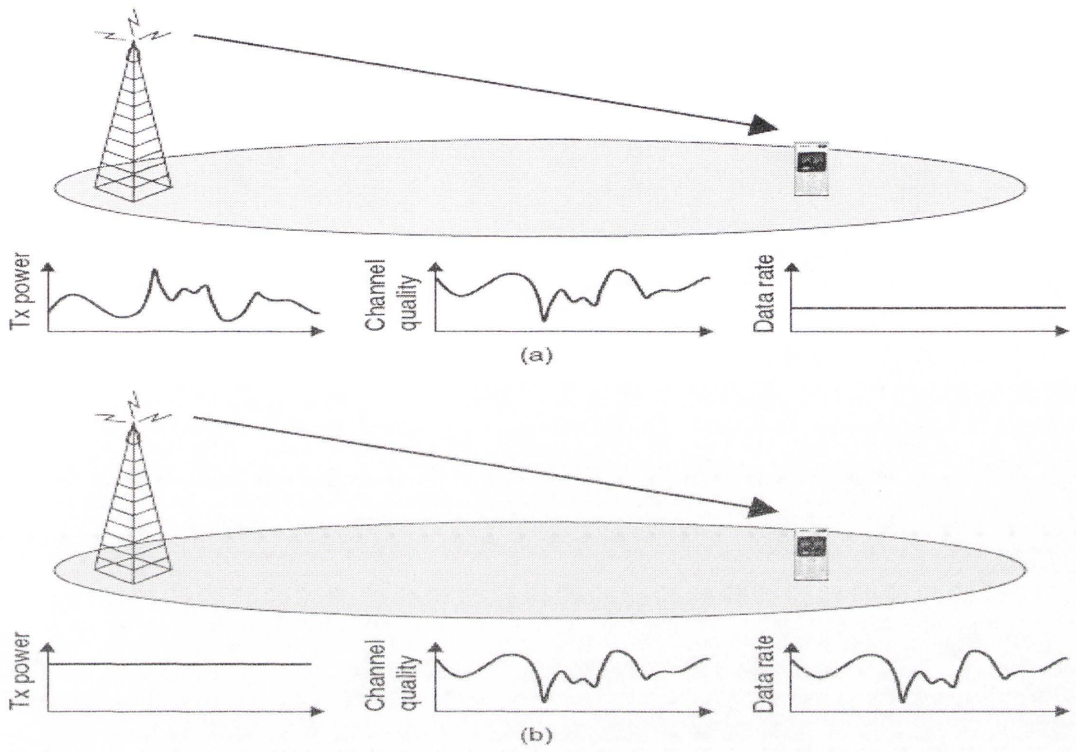

Figure 2-10 (a) Power Control and (b) Rate Control [10]

2.5 Radio Propagation Model

Radio propagation refers to how radio signals are propagated from one place to another. It might be affected by reflection, refraction, diffraction, absorption, polarization and scattering [17]. Several simplified mathematical models have been proposed to model the radio propagation.

One of the most widely accepted models is given in Figure 2-11. The overall propagation effect on the signal is denoted as a parameter called "channel gain", g_{total} [18]. The received signal power can be calculated by the sum of the product of the transmit signal power and the path gain and the thermal noise power. R, S and P_N represent the power of the received signal, the transmitted signal and the thermal noise respectively.

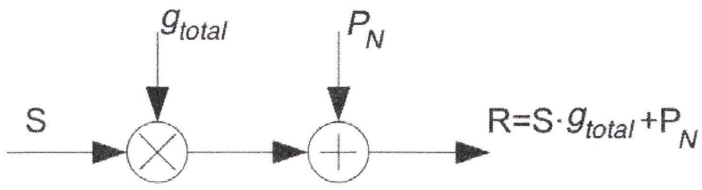

Figure 2-11 Radio Propagation Model [16]

The channel gain is further modelled as a combination of path loss [8], shadow fading [19] and multi-path fading [20] propagation gains, given as

$$g_{total} = g_{pl} \times g_s \times g_{mpath} . \tag{2-1}$$

2.5.1 Path Loss

Path loss is the large-scale fading depending on the distance between the transmitter and the receiver.

The Extended COST-231 Hata Model for urban environment proposed in [21] is used in this thesis. This model is proposed for the frequency of 2 GHz and simplifies the calculation of path loss as a closed form formula [22]. The path loss in dB is given as

$$\begin{aligned} pl_i(t) &= 46.3 + 33.9 \times \log_{10}(f_c) - 13.82 \times \log_{10}(h_t) \\ &+ a(h_r) + (44.9 - 6.55 \times \log_{10}(h_t)) \times \log_{10}(d_i(t)) + C_m \end{aligned}, \tag{2-2}$$

where

$$a(h_r) = (1.1 \times \log_{10}(f_c) - 0.7) \times h_r - (1.56 \times \log_{10}(f_c) - 0.8), \tag{2-3}$$

in which $pl_i(t)$ and $d_i(t)$ denote the path loss and distance (in *km*) of user i at time t, respectively. f_c is the frequency of the transmission in MHz, h_t is the height of base station or transmitter in meters and h_r is the height of the mobile or receiver in meters.

2.5.2 Shadow Fading

Shadow fading refers to signal attenuations caused by signal reflection, diffraction and shielding phenomenon from obstructions such as building, trees, and rocks.

The approach proposed in [23] models the shadow fading as a correlated log-normal distribution with zero mean (in dB) and a specific standard deviation (in dB). The mathematical expression for the shadow fading (in dB) is given as

$$\xi_i(t+1) = \rho_i(t) \times \xi_i(t) + \sigma \times \left(\sqrt{1-\rho_i(t)^2}\right) \times W(t), \tag{2-4}$$

where $\xi_i(t)$ is the shadow fading gain in dB of user i and $W(t)$ is a random Gaussian variable at time t. σ represents the given standard deviation for shadow fading. $\rho_i(t)$ denotes the autocorrelation function of shadow fading of user i at time t, and can be calculated by

$$\rho_i(t) = \exp\left(\frac{-v_i}{d_0}\right), \tag{2-5}$$

in which d_0 is the correlation distance of shadow fading and v_i is the speed of user i.

2.5.3 Multi-path Fading

The multi-path fading refers to the addition of multi-path components caused by the reflection and scattering of the radio signal. The received signals from different path have different attenuations and delays, which result in fluctuations of the received signal.

In this thesis, the multi-path fading is approximated as a complex random Gaussian process $\mu(t)$, which is given as

$$\mu(t) = \sqrt{\mu_1^2(t) + \mu_2^2(t)}, \tag{2-6}$$

where $\mu_1(t)$ and $\mu_2(t)$ are uncorrelated filtered white Gaussian noises with zero means $E[\mu_i(t)] = = 0$ and identical variances $Var[\mu_i(t)] = \sigma_{\mu i}^2 = \sigma_{\mu 0}^2$, $i=1,2$.

As discussed in [24], the approximation of each Gaussian process $\mu_i(t)$ ($i=1,2$) can be expressed as a finite sum of properly weighted sinusoids with evenly distributed phases, i.e.

$$\tilde{\mu}_i(t) = \sum_{n=1}^{N_i} c_{i,n} \cos(2\pi f_{i,n} t + \theta_{i,n}), \qquad i = 1,2 \quad , \tag{2-7}$$

where N_i, $c_{i,n}$, $f_{i,n}$ and $\theta_{i,n}$ denote the number of sinusoids, Doppler coefficient, discrete Doppler frequency and Doppler phase of the ith process, respectively.

The Monte Carlo Method (MCM) [25] is deployed to determine the value of parameters $c_{i,n}$ and $f_{i,n}$. Then the approximated Gaussian process can be rewrote as

$$\tilde{\mu}_i(t) = \sum_{n=1}^{N_i} \sigma_{\mu_0} \sqrt{\frac{2}{N_i}} \cos(2\pi f_{\max} \sin(\frac{\pi}{2}\mu_n)t + \theta_{i,n}), \qquad i = 1,2\,. \tag{2-8}$$

in which f_{max} is the maximum Doppler frequency.

The envelope of the Gaussian process $\mu(t)$ is a Rayleigh process $\xi(t)$ [26], which is expressed as

$$\xi(t) = |\mu(t)| \,. \tag{2-9}$$

The model of multi-path fading is given in Figure 2-12.

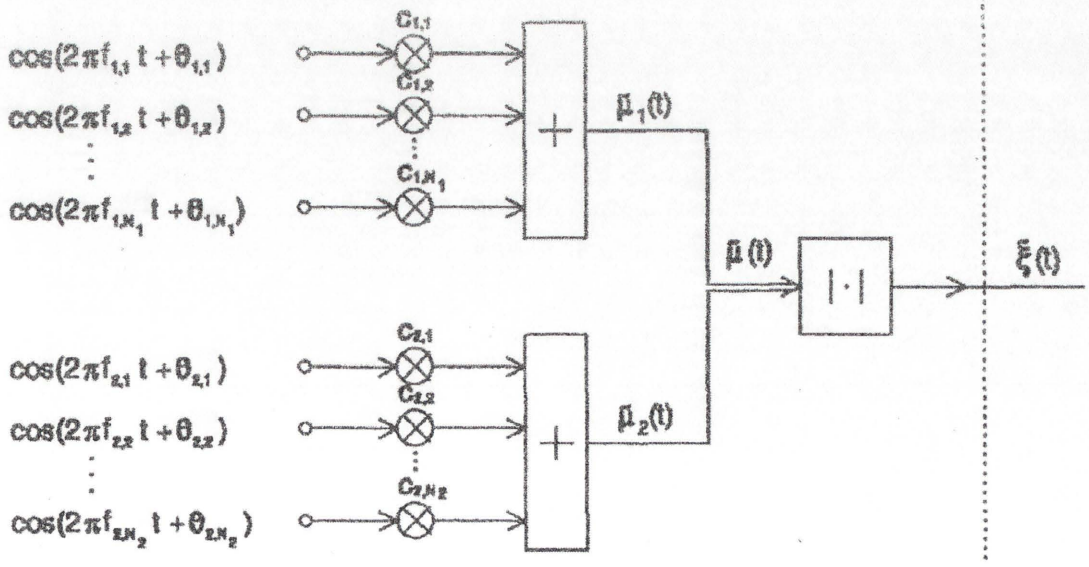

Figure 2-12:Model of Multi-path Fading [25]

2.5.4 SNR to Data Rate Mapping

As discussed earlier, channel gain is determined by the path loss, shadow fading and multi-path fading. Then for the downlink LTE system, the channel gain of user i on RB j at time t, denoted as $Gain_{i,j}(t)$, is given as

$$Gain_{i,j}(t) = 10^{\left(\frac{pl_i(t)}{10}\right)} \times 10^{\left(\frac{\xi_i(t)}{10}\right)} \times 10^{\left(\frac{mpath_{i,j}(t)}{10}\right)}, \tag{2-10}$$

where $pl_i(t)$ and $\xi_i(t)$ denotes the path loss and shadow fading (both in dB) of user i at time t respectively and $mpath_{i,j}(t)$ represent the multi-path fading (in dB) of user i on RB j at time t.

Then the instantaneous downlink SNR value of user i on RB j at time t ($\gamma_{i,j}(t)$) can be obtained through the approach discussed in [27-29], as shown below:

$$\gamma_{i,j}(t) = \frac{P_{total}/N \times Gain_{i,j}(t)}{I + N_o}, \tag{2-11}$$

where P_{total} is the total eNodeB downlink transmission power, N is the number of available RBs, I and N_0 represent the inter-cell interference power level and the noise power level within each RB, respectively. N_0 is the noise power after performing matched filter detection at the receiver. As the inter-cell interference refers to the cell interference caused by the neighbouring cells and the assumption in this thesis is based on one cell scenario, we can say that the inter-cell interference is not applicable in this thesis and I can be set to zero.

The number of bits per symbol of user i on a subcarrier within RB j at time t ($nbits_{i,j}(t)/symbol$) can be computed according to the approach discussed in [27-29]. The achievable data rate for user i at time t ($date\_rate_i(t)$) can be obtained by

$$data\_rate_i(t) = \frac{nbits_{i,j}(t)}{symbol} \times \frac{nsymbols}{slot} \times \frac{nslots}{TTI} \times \frac{nsc}{N}, \quad (2\text{-}12)$$

where *nsymbols/slot* is the number of symbols per time slot, *nslot/TTI* is the number of time slots per Transmission Time Interval (TTI), *nsc/N* is the number of subcarriers per RB and *N* is the number of available RBs.

Therefore, based on the computed SNR value given in (2-11), the achievable data rate can be determined by (2-12), and an appropriate modulation and coding scheme (MCS) can be chosen according to Table 2-2.

Table 2-2. Mapping Table of Downlink SNR to Data Rate [29]

Minimum Instantaneous Downlink SNR Value (dB)	Modulation and Coding	Data Rate (kbps)
1.7	QPSK (1/2)	168
3.7	QPSK (2/3)	224
4.5	QPSK (3/4)	252
7.2	16 QAM (1/2)	336
9.5	16 QAM (2/3)	448
10.7	16 QAM (3/4)	504
14.8	64 QAM (2/3)	672
16.1	64 QAM (3/4)	756

2.6 Summary

In this chapter, background knowledge of LTE architecture, resource block and OFDMA technology has been given. Six RRM functions have been briefly explained. A detailed radio propagation model used in this thesis has also been discussed.

Chapter 3

PACKET SCHEDULING ALGORITHMS

This chapter discusses the performance metrics that are designed for the performance analysis of packet scheduling algorithms and gives brief introduction on several packet scheduling algorithms. The performances of these algorithms are evaluated under the downlink LTE simulation environment and performance comparison of packet scheduling algorithms is provided at the latter part of this chapter.

3.1 Performance Metrics of Packet Scheduling Algorithms

The LTE system is designed as a packet-optimized network supporting both Real-Time (RT) and Non-Real-Time (NRT) traffics. Packet scheduling plays an important role in guaranteeing the system performance. The vital target of packet scheduling algorithms is to meet the QoS and fairness requirements of each user while ensuring the efficient usage of the available radio resources. In this thesis, the performances of packet scheduling algorithms are evaluated in terms of performance metrics such as system throughput, average system HOL delay, Packet Loss Ratio (PLR), fairness and Resource Block (RB) utilization.

The system throughput is a measure of the average transmission rate of the system. We assume that there are no transmission errors. Then the system throughput is defined as the sum of transmitted packet size of all users per second, which is given by

$$system\ throughput = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} ptransmit_i(t), \tag{3-1}$$

where K is the total number of users, T represents the total simulation time, and $ptransmit_i(t)$ denotes the number of transmitted bits of user i at time t.

Average system Head of Line (HOL) delay is one of the QoS requirements. The HOL delay is defined as the time duration from arrival time of the first packet waiting in the buffer to current time. Average system HOL delay describes the average HOL delay of all the user buffers throughout the simulation time, which is given as follows:

$$Average\ System\ HOL\ Delay = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{K}\sum_{i=1}^{K}W_i(t), \tag{3-2}$$

where $W_i(t)$ denotes the HOL delay of user i at time t.

RT users and NRT users require different delay deadlines for the packet transmission. A packet will be discarded once the waiting time of the packet exceeds the user's delay deadline. Packet Loss Ratio (PLR) is defined as the proportion of total discarded packet size to total arrived packet size. PLR is mathematically expressed as:

$$PLR = \frac{\sum_{t=1}^{T}\sum_{i=1}^{K} pdiscard_i(t)}{\sum_{t=1}^{T}\sum_{i=1}^{K} psize_i(t)}, \tag{3-3}$$

in which $psize_i(t)$ and $pdiscard_i(t)$ denotes the total received packet size and the total discarded packet size of user i at time t, respectively.

Fairness measures whether users are receiving a fair resource block allocation. Fairness evaluates the difference between the users who have the most and least transmitted packet size. The mathematical expression of fairness is given as:

$$fairness = 1 - \frac{\max(ptotaltransmit(i)) - \min(ptotaltransmit(i))}{\sum_{t=1}^{T}\sum_{i=1}^{K} psize_i(t)}, \tag{3-4}$$

and

$$ptotaltransmit(i) = \sum_{t=1}^{T} ptransmit_i(t), \tag{3-5}$$

where $psize_i(t)$ denotes the total received packet size of user i at time t and $ptotaltransmit(i)$ is the total transmitted packet size of user i throughout the simulation time.

The resource block (RB) utilization is defined as the ratio of average number of RBs that have been used for transmission at each TTI to the total number of RBs, which can be mathematically expressed as:

$$RB\_utilization = \frac{\frac{1}{T}\sum_{t=1}^{T} total_{RB\_used}(t)}{N}, \tag{3-6}$$

where $total_{RB\_used}(t)$ is the total number of RBs that have been used for transmission at time t and N is the total number of RBs.

3.2 Review of Packet Scheduling Algorithms

This section provides a review of existing packet scheduling algorithms. Several well-known and recently proposed Packet scheduling algorithms will be discussed in the following subsections.

3.2.1 Round Robin (RR)

The Round Robin (RR) algorithm [10] assigns equal portions of packet transmission time to each user in a circular order. The index number of the user who is selected for transmission at time t is denoted as $k(t)$ and can be updated by

$$k(t) = \begin{cases} 1 & t = 1 \\ k(t-1)+1 & t > 1 \ and \ k(t-1) < K \\ 1 & t > 1 \ and \ k(t-1) = K \end{cases}. \tag{3-7}$$

where K is the total number of users.

RR algorithm achieves the best fairness performance if the users have similar arrival packet sizes and instantaneous achievable data rate. Since RR algorithm does not take channel conditions for each user into consideration, it may have a comparatively worse throughput performance.

3.2.2 First-In-First-Out (FIFO)

The First-In-First-Out (FIFO) algorithm gives transmission priority to the user with the highest HOL delay at each time slot, as given in (3-8).

$$k = \arg\max W_i(t). \tag{3-8}$$

For the similar reasons as RR algorithm, FIFO algorithm has a good fairness performance but a low throughput performance.

3.2.3 Maximum Rate (Max-Rate)

The Maximum Rate (Max-Rate) algorithm [10] transmits the packets of the user with highest instantaneous achievable data rate, as given in (3-9).

$$k = \arg\max r_i(t), \tag{3-9}$$

where $r_i(t)$ is the instantaneous achievable data rate of user i at time t which depends on the reported SNR value, as discussed in (2-12).

Max-Rate algorithm maximizes the system throughput since it always selects users with the best channel conditions. On the contrary, users with low SNR values might never be selected for transmission, which leads to the poor fairness performance of Max-Rate algorithm.

3.2.4 Proportional Fair (PF)

Proportional Fair (PF) algorithm [30] is proposed to provide a balanced performance between the fairness and system throughput. The metric k is given as

$$k = \arg\max \frac{r_i(t)}{R_i(t)}, \tag{3-10}$$

and

$$R_i(t) = (1 - \frac{1}{t_c}) \times R_i(t-1) + \frac{1}{t_c} \times r_i(t-1), \tag{3-11}$$

where $r_i(t)$ and $R_i(t)$ are the instantaneous achievable data rate and the average data rate of user i at time t, respectively. Parameter t_c is the update window size and controls the latency of the system.

As PF algorithm incorporates the instantaneous achievable data rate with the average data rate of each user at every time slot, it achieves a good balance between throughput and fairness performance.

3.2.5 Maximum-Largest Weighted Delay First (M-LWDF)

The Maximum-Largest Weighted Delay First (M-LWDF) algorithm [31] is proposed to support RT services. The scheduling criterion is given as follows:

$$k = \arg\max a_i W_i(t) \frac{r_i(t)}{R_i(t)}, \tag{3-12}$$

where

$$a_i = -\frac{(\log \delta_i)}{\tau_i}, \tag{3-13}$$

in which $W_i(t)$ is the HOL packet delay of user i at time t, τ_i is the delay threshold of user i and δ_i denotes the maximum probability for HOL packet delay of user i to exceed the delay threshold of user i.

Since M-LWDF jointly considers HOL delay along with the instantaneous data rate and average data rate of each user, it obtains a good throughput and fairness performance along with a relatively low PLR.

3.2.6 Exponential/Proportional Fair (EXP/PF)

The Exponential/Proportional Fair (EXP/PF) [32, 33] is designed to support multi-media applications. The scheduling criterion k for either RT or NRT services of each user is defined as

$$k = \arg\max \begin{cases} \exp\dfrac{a_i W_i(t) - \overline{aW(t)}}{1+\sqrt{\overline{aW(t)}}} \dfrac{r_i(t)}{R_i(t)} & i \in RT \\ \dfrac{w(t)}{M(t)} \dfrac{r_i(t)}{R_i(t)} & i \in NRT \end{cases}, \tag{3-14}$$

and

$$\overline{aW(t)} = \frac{1}{N_{RT}} \sum_{i \in RT} a_i W_i(t), \tag{3-15}$$

$$w(t) = \begin{cases} w(t-1) - \varepsilon & W_{\max} > \tau_{\max} \\ w(t-1) + \dfrac{\varepsilon}{k} & W_{\max} < \tau_{\max} \end{cases}, \tag{3-16}$$

where $M(t)$ is the average number of waiting packets for all RT services at time t, ε and k are constant, and W_{max} and τ_{max} are the maximum HOL packet delay out of RT service users and maximum delay constraint of all RT service users, respectively.

The EXP/PF algorithm gives a higher priority to the RT service users whose packets are approaching the transmission deadline than NRT service users.

3.2.7 Jeongsik Park's Algorithm

Jeongsik Park's Algorithm [34] is divided into two steps. Step 1 allocates the available RBs to users whose packets are approaching the transmission deadline. Whenever there are remaining RBs after Step 1 has been executed, the algorithm will allocate the remaining RBs based on throughput enhancement (as described in Step 2).

Step 1

The users' queue state information can be updated at every TTI (Transmit Time Interval) by

$$Bcurr_i(t+1) = Bcurr_i(t) + (psize_i(t) - ptransmit_i(t) - pdrop_i(t)), \quad (3\text{-}17)$$

in which $Bcurr_i(t)$ represents the number of bits in user i's buffer at time t; $psize_i(t)$, $ptransmit_i(t)$ and $pdrop_i(t)$ denotes the number of received bits, the number of transmitted bits and the number of dropped bits of user i at time t, respectively.

The delay constraint for real-time service is given by TFT (Time for Transmission). TFT is defined as the maximum acceptable time duration from packet arrival in the buffer to departure within which the packet will not be dropped. In other words, packet will be dropped once the delay of the packet exceeds its assigned TFT. In order to meet the assigned TFT, the following condition needs to be satisfied:

$$\sum_{k=t}^{t+TFT+1} (ptransmit_i(t) + pdrop_i(t)) \geq psize_i(t). \quad (3\text{-}18)$$

We rewrite (3-18) as

$$\begin{aligned} ptransmit_i(t) \geq \max[&Bcurr_i(t-TFT+1) \\ &- \sum_{i=t+TFT+1}^{t-1} (ptransmit_i(t) + pdrop_i(t)), 0], \\ &= \gamma_i(t) \end{aligned} \quad (3\text{-}19)$$

where $\gamma_i(t)$ denotes the number of urgent packets for user i that need to be transmitted at time t in order to avoid packet loss.

Based on $\gamma_i(t)$, users are put into two groups: patient (S_p) and impatient (S_{ip}) groups.

$$S_p = \{i | \gamma_i(t) = 0\}, \quad (3\text{-}20)$$

$$S_{ip} = S - S_p = \{i | \gamma_i(t) > 0, 1 \leq i \leq K\}, \quad (3\text{-}21)$$

where K is the total number of users.

As the HOL packets of users in the impatient group are approaching the assigned TFT, the scheduler allocates resource blocks to impatient group prior to patient group. Only users in impatient group will have the chance to compete for resource blocks in Step 1.

Step 2

If there are still some resource blocks available after Step1, the remaining resource blocks will be assigned to active users based on the following procedure.

Assume users periodically report the Channel State Information (CSI) to the base station. The reported CSI at time t is recorded by the channel matrix H which is defined as

$$H(t)=\begin{pmatrix} C_{11}(t) & C_{12}(t) & C_{13}(t) & \cdots & C_{1K}(t) \\ C_{21}(t) & C_{22}(t) & C_{23}(t) & \cdots & C_{1K}(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{N1}(t) & C_{N2}(t) & C_{N3}(t) & \cdots & C_{NK}(t) \end{pmatrix}, \tag{3-22}$$

where N and K are the total number of RBs and users respectively. $C_{ij}(t)$ represents the channel state of user i on sub-band j at time slot t and has an integer value between 1 (worst channel state) and 9 (best channel state).

$C_{avg}(t)$ denotes the average value of $C_{ij}(t)$ and can be calculated by

$$C_{avg}(t)=\sum_{i=1}^{K}\sum_{j=1}^{N}\frac{C_{ij}(t)}{K\times N}. \tag{3-23}$$

Index matrix I denotes whether users' reported channel state values on each RB at time t are above average and is defined as

$$I(t)=\begin{pmatrix} a_{11}(t) & a_{12}(t) & a_{13}(t) & \cdots & a_{1K}(t) \\ a_{21}(t) & a_{22}(t) & a_{23}(t) & \cdots & a_{1K}(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{N1}(t) & a_{N2}(t) & a_{N3}(t) & \cdots & a_{NK}(t) \end{pmatrix}, \tag{3-24}$$

in which $a_{ij}(t)$ is given by

$$a_{ij}(t) = \begin{cases} 1 & if\ C_{ij}(t) > C_{avg}(t), \\ 0 & otherwise. \end{cases} \tag{3-25}$$

The average channel state value on RB *i* experienced by all users can be calculated by

$$M_{Ri}(t) = \frac{\sum_{j=1}^{K} C_{ij}(t)}{K}, \tag{3-26}$$

Similarly, the average channel state value of all RBs experienced by user *j* is given as

$$M_{Cj}(t) = \frac{\sum_{j=1}^{N} C_{ij}(t)}{N}. \tag{3-27}$$

We denote $C_{\max}(t) = \max\lfloor C_{ij}(t) \rfloor$ for $1 \le i \le N$ and $1 \le j \le K$.

According to [34], the Sub-band Discrimination Factor (SDF) at time *t* and the User Discrimination Factor (UDF) at time *t* are defined to facilitate the resource allocation. The mathematical expression for SDF and UDF are given in (3-28) and (3-29), respectively.

$$SDF_i(t) = \frac{\sqrt{\frac{\sum_{j=1}^{K} a_{ij}(t)(C_{\max}(t) - C_{ij}(t))^2}{b_i(t)}}}{M_{Ri}(t)}, \tag{3-28}$$

$$UDF_j(t) = \frac{\sqrt{\frac{\sum_{j=1}^{N} a_{ij}(t)(C_{\max}(t) - C_{ij}(t))^2}{g_j(t)}}}{M_{Cj}(t)}, \tag{3-29}$$

where $b_i(t) = \sum_{j=1}^{K} a_{ij}(t)$ and $g_j(t) = \sum_{j=1}^{N} a_{ij}(t)$.

$SDF_i(t)$ is defined as a variance of the difference between C_{max} and channel state values on RB *i* experienced by users whose channel state value on RB *i* are above C_{avg} at time *t*. A RB with the smaller SDF value indicates that the overall channel states of the RB are closer to C_{max}. In other words, RB with smaller SDF value has more users whose

channel state values are good enough to be chosen for transmission. To enhance the throughput performance, RBs with higher SDF values should be considered for transmission prior to those with lower SDF values.

Similarly, $UDF_j(t)$ is defined as a variance of the difference between C_{max} and channel state values of user j on the RBs on which channel state value of user j are above C_{avg} at time t. A user with the smaller UDF value has better alternative RBs. Users with higher UDF values should be given higher priority than those with lower UDF values.

For each TTI, according to the maximum channel state value in each sub-band, sub-bands are classified into several channel state groups. The scheduler gives priority to groups with larger maximum channel state value over groups with smaller maximum channel state value. In each group, the sub-band with largest SDF will be considered first. The scheduler will assign the selected sub-band to the user with the best channel state. If there is more than one user who has the best channel state in the selected sub-band, the user with the highest UDF will be selected. This procedure is repeated until all sub-bands have been allocated.

As channel quality is the critical criterion for the allocation decision-making in Step 2, a throughput enhancement is achieved by Jeongsik Park's Algorithm.

Since Jeongsik Park's Algorithm gives higher priority to users approaching the transmission deadline in Step 1 and considers the channel quality in Step 2, it achieves a good PLR and throughput performance. But Jeongsik Park's Algorithm requires a comparatively longer time for scheduling decision making as the scheduler needs two steps to make the allocation decision.

3.2.8 Sun Qiaoyun's Algorithm

Sun Qiaoyun's Algorithm [35] allocates resource blocks by jointly considering the Channel State Information (CSI), users' Quality of Service (QoS) requirement and the Queue State Information (QSI). These three factors will be described in the following subsections and a priority metric $\mu_i(t)$ will be defined based on these factors. The user with the highest priority metric will be selected for transmission.

The CSI factor

A Channel Quality Indicator (CQI) is defined to provide CSI from users to the scheduler. Sun Qiaoyun ' s Algorithm considers the CQI based on the famous proportional fairness (PF) [30] scheduling, which is given as:

$$f_1(CQI_i(t)) = \frac{r_i(t)}{R_i(t)}, \tag{3-30}$$

where $r_i(t)$ and $R_i(t)$ are the current achievable data rate and the average data rate of user i at time t respectively. $R_i(t)$ can be updated by

$$R_i(t+1) = (1-\frac{1}{t_c}) * R_i(t) + \frac{1}{t_c} * r_i(t). \tag{3-31}$$

The QoS factor

As Sun Qiaoyun's Algorithm is proposed to support RT services, QoS requirement of RT services is a significant scheduling criterion. Two important QoS parameters will be considered, which are PLR and Head-of-Line (HOL) packet delay. For each user, both PLR and HOL delay should meet the transmission constraint, given as follows:

$$PLR_i(t) \le PLR_{req,i}, \tag{3-32}$$

and

$$W_i(t) < W_{\max,i}, \tag{3-33}$$

where $PLR_i(t)$ and $W_i(t)$ represent the packet loss rate and HOL packet delay of user i at time t respectively, and $PLR_{reg,i}$ and $W_{max,i}$ denote the PLR threshold and maximum allowable HOL delay of user i respectively.

The QoS factor is utilized to optimize the QoS performance and is defined as

$$f_2(QoS_i(t)) = \frac{PLR_i(t)}{PLR_{req,i}} \cdot \frac{W_i(t)}{W_{\max,i}}. \tag{3-34}$$

The QSI factor

Queue status of users is another important factor that affects the system performance. The QSI factor is utilized to provide QSI to the scheduler and is denoted as

$$f_3(QSI_i(t)) = \frac{Bcurr_i(t)}{Bcurr\_avg_i(t)}, \tag{3-35}$$

where $Bcurr_i(t)$ is the queue length (in bits) of user i at time t and $Bcurr\_avg_i(t)$ is the average queue length of all the K users as given in $Bcurr\_avg_i(t) = \frac{1}{K}\sum_i Bcurr_i(t)$.

By jointly considering these three factors, the priority metric for each sub-band at timeslot t is defined as:

$$\begin{aligned} \mu_i(t) &= f(CQI_i(t), QoS_i(t), QSI_i(t)) \\ &= f_1(CQI_i(t)) \cdot f_2(QoS_i(t)) \cdot f_3(QSI_i(t)) \\ &= \frac{r_i(t)}{R_i(t)} \cdot \frac{PLR_i(t)}{PLR_{req,i}} \cdot \frac{W_i(t)}{W_{\max,i}} \cdot \frac{Bcurr_i(t)}{Bcurr\_avg_i(t)} \end{aligned} \tag{3-36}$$

For each sub-band at each TTI, the scheduler allocates the resource blocks to the user with the highest priority metric.

As Sun Qiaoyun's Algorithm allocates RBs based on the channel conditions, users' QoS requirement and queue status of each user, it has a good throughput and fairness performance and can support a certain number of RT users with the desired QoS requirement.

3.3 Performance Comparison of Packet Scheduling Algorithms

To analyse the performance of packet scheduling algorithms, the simulation of downlink LTE system was set up in a Matlab environment.

The channel model and the traffic model are mostly taken from [35] and [36]. The relevant parameters are given in Table 3-1, Table 3-2 and Table 3-3, respectively.

Table 3-1. Downlink LTE System Parameters[28, 29]

Parameters	Values
Carrier Frequency	2 GHz
Bandwidth	5 MHz
Number of Sub-carriers	300
Number of RBs	25
Number of Sub-carriers per RB	12
Sub-Carrier Spacing	15 kHz
Slot Duration	0.5 ms
Scheduling Time (TTI)	1 ms
Number of OFDM Symbols per Slot	7

Table 3-2. Parameters of a RT Video Streaming Application [28, 37]

Information types	Distribution	Distribution Parameters
Inter-arrival time between the beginning of successive frames	Deterministic (Based on 20fps)	50ms
Number of packets (slices) in a frame	Deterministic	8
Packet (slice) size	Truncated Pareto (Mean=100bytes, max=125bytes)	K=40bytes, α=1.2
Inter-arrival time between packets (slices) in a frame	Truncated Pareto (Mean=6ms, Max=12.5ms)	K=2.5ms, α=1.2

Table 3-3. Parameters of a NRT Web Browsing Application [28, 37]

Component	Distribution	Distribution Parameters
Main object size (S_M)	Truncated Lognormal (Min=100 bytes, Max=20 Kbytes)	σ=25032 bytes, μ=10710 bytes
Embedded object size (S_E)	Truncated Lognormal (Min=50 bytes, Max=20 Kbytes)	σ=126168 bytes, μ=7758 bytes
Number of embedded objects per page (N_d)	Truncated Pareto (Mean=5.64, Max=53)	K=2, m=55, α=1.1
Parsing time (T_p)	Exponential (Mean=0.13 sec)	χ=7.69
Reading time (D_{pc})	Exponential (Mean=30 sec)	χ=0.33

A radio cell with a centralized eNodeB and K active wireless users is considered. The 5MHz system bandwidth is divided into 25 RBs. The carrier frequency is 2GHz. The users are uniformly distributed within the cell with speeds between 1-100 km/h at random directions. The schedulers can allocate multiple RBs to active users at each TTI, which is 1 ms in LTE system.

As discussed in Section 2.5, the achievable data rate can be determined based on the SINR value through the approach proposed in [8] and [19], and an appropriate MCS can be chosen according to Table 2-2.

Both Real-Time (RT) video streaming application and Non-Real-Time (NRT) web browsing application are tested in the simulation, as shown in Table 3-2 and Table 3-3, respectively.

For the RT users, the source video data rate is 128 kbps [28, 29, 37, 38]. The delay threshold is 20 ms [39] and the requested packet loss rate is set to 0.01 [40-42]. We use the assumption as given in [28] for the RT users. Assume that the buffer for the RT user is full at the beginning of the simulation and able to store 5s of video streaming service [37]. The mean running time of a video streaming is assumed to be around 23s [28]. Then in order to avoid the buffer running empty, the RT users require a minimum throughput of 100 kbps [28].

The NRT users are assumed to have infinite buffers.

3.3.1 Performance Comparison of Well-Known Packet Scheduling algorithms

In this subsection, performance of five well-known packet scheduling algorithms is evaluated. These algorithms are RR, PF, Max-Rate, M-LWDF and EXP/PF.

These well known algorithms are tested in three scenarios, which are 100% RT scenario, 100% NRT scenario and 50% RT and 50% NRT scenario.

100% RT Scenario

In this sub-section, the performance of system with up to 110 RT users will be tested. Four packet scheduling algorithms are evaluated, which are RR, PF, Max-Rate and M-LWDF. The simulation results are given in the following figures.

Figure 3-1 compares the system throughput of the four packet scheduling algorithms. M-LWDF achieves the highest system throughput while RR has the lowest system throughput. When the number of RT users is larger than 70, M-LWDF outperforms Max-Rate and PF in terms of system throughput, as M-LWDF considers not only channel conditions but also the average system HOL delay. As M-LWDF gives higher priority to the user with larger HOL delay, it achieves a lower PLR and a higher probability of successful packet transmission so that M-LWDF has a comparatively better throughput performance.

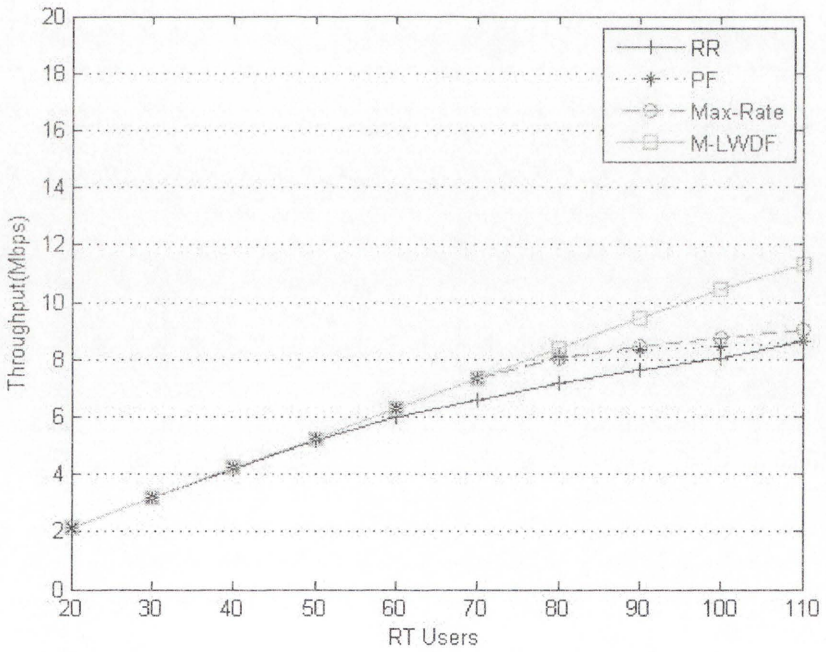

Figure 3-1: System Throughput vs. Number of RT Users

As shown in Figure 3-2 and Figure 3-3, M-LWDF has the lowest delay and PLR, as it takes the delay and PLR requirements of RT users into consideration. Max-Rate and PF have similar but relatively worse delay and PLR performance when compared with M-

LWDF. As the requested PLR is set to 0.01, M-LWDF can support more RT users (110 RT users) than Max-Rate (70 RT users) and PF (70 RT users). Moreover, RR has the highest delay and PLR and can only support 40 RT users for a PLR of 0.01.

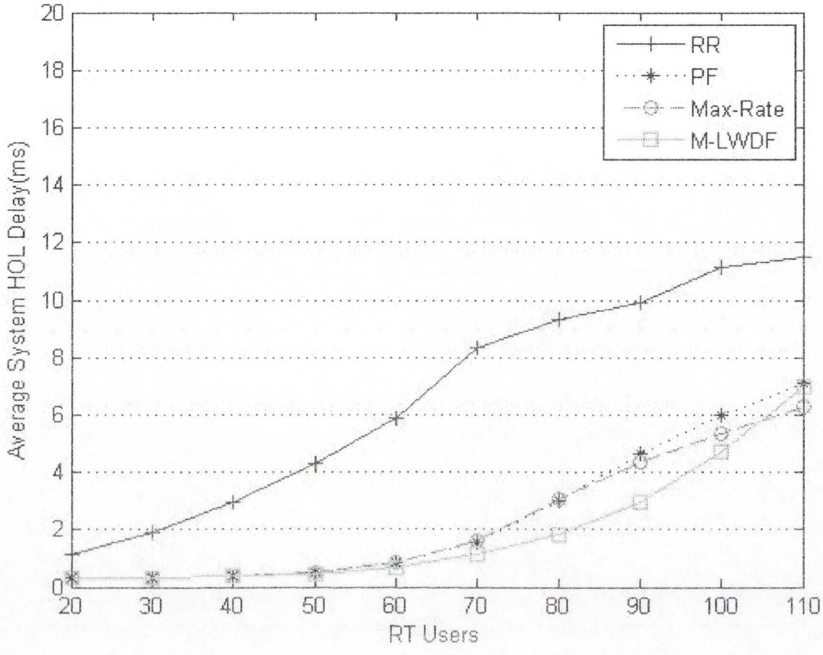

Figure 3-2: Average System HOL Delay vs. Number of RT Users

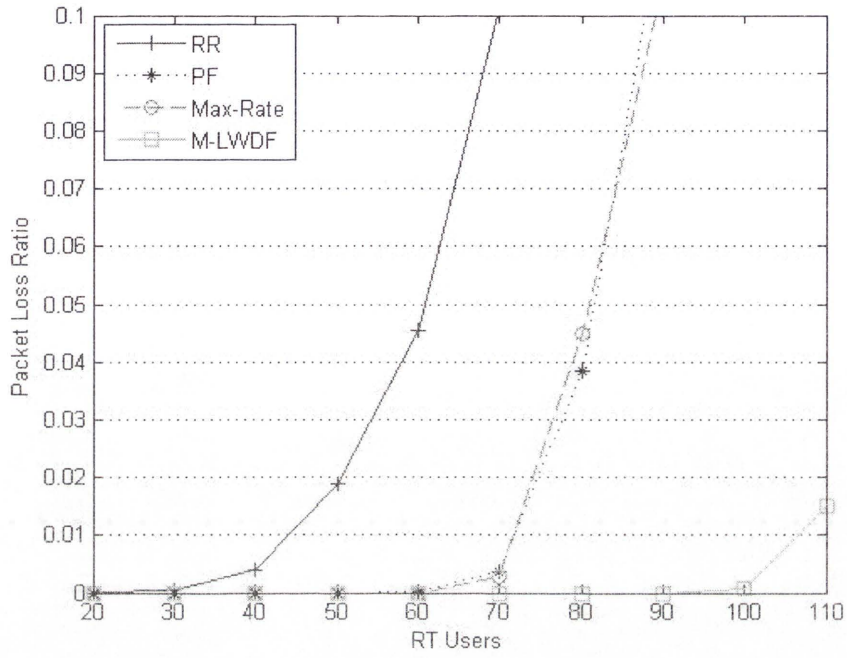

Figure 3-3: PLR vs. Number of RT Users

The RB utilization is given in Figure 3-4. The RB utilization for PF, Max-Rate and M-LWDF are similar and much better than that of RR. As RR allocates RBs to users in a circular order and doesn't take channel condition into consider, it doesn't make fully use of RBs and has a comparatively worse RB utilization than other algorithms.

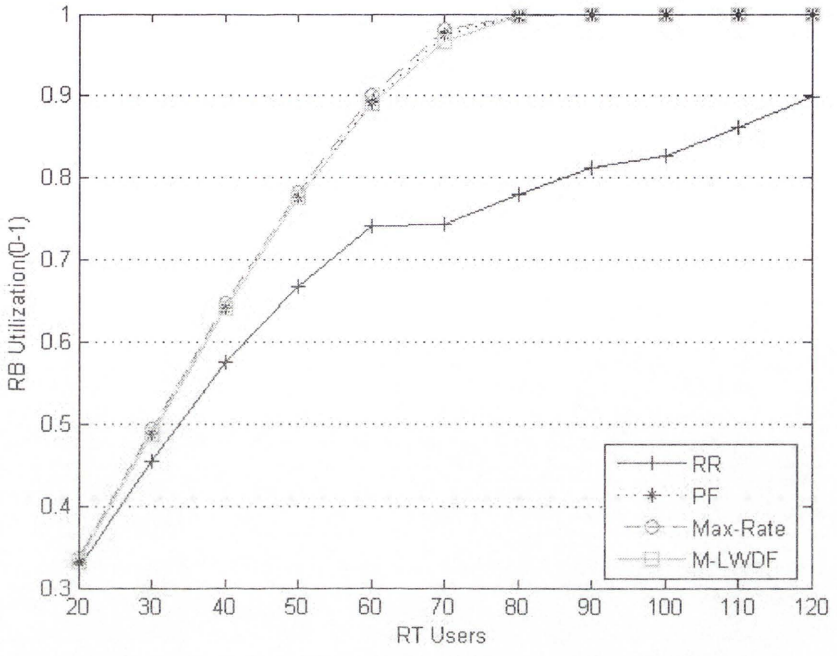

Figure 3-4: RB Utilization vs. Number of RT Users

We can conclude that in the 100% RT scenario, M-LWDF outperforms RR, PF and Max-Rate. PF and Max-Rate achieve the similar performance while RR performs the worst.

100% NRT Scenario

System with up to 300 NRT users is chosen in this sub-section. The evaluated packet scheduling algorithms are RR, PF and Max-Rate. The simulation results are given as follows.

Figure 3-5 shows the system throughput of packet scheduling algorithms. Max-Rate achieves highest system throughput as it always allocates RBs to the user with the highest instantaneous achievable data rate. PF has a good system throughput as well because it also takes the channel condition into consideration. On the other hand, RR has the worse throughput performance, as it allocates RBs to users in a circular order and doesn't consider the channel condition.

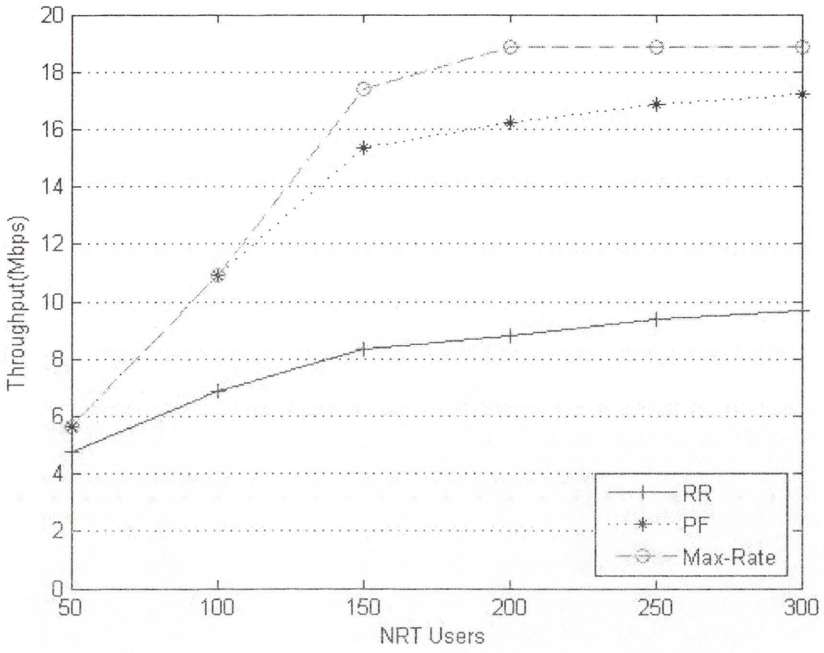

Figure 3-5: System Throughput vs. Number of NRT Users

The fairness and RB utilization is given in Figure 3-6 and Figure 3-7, respectively. RR has the best fairness but the worst RB utilization performance. As PF considers the average throughput for each user while making the scheduling decision, it achieves a slightly better fairness and RB utilization than Max-Rate.

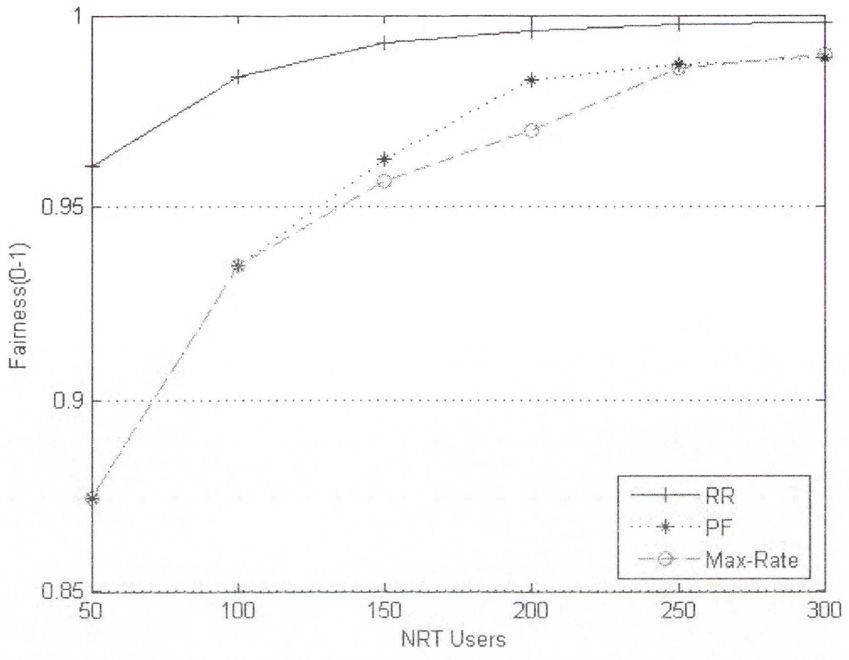

Figure 3-6: Fairness vs. Number of NRT Users

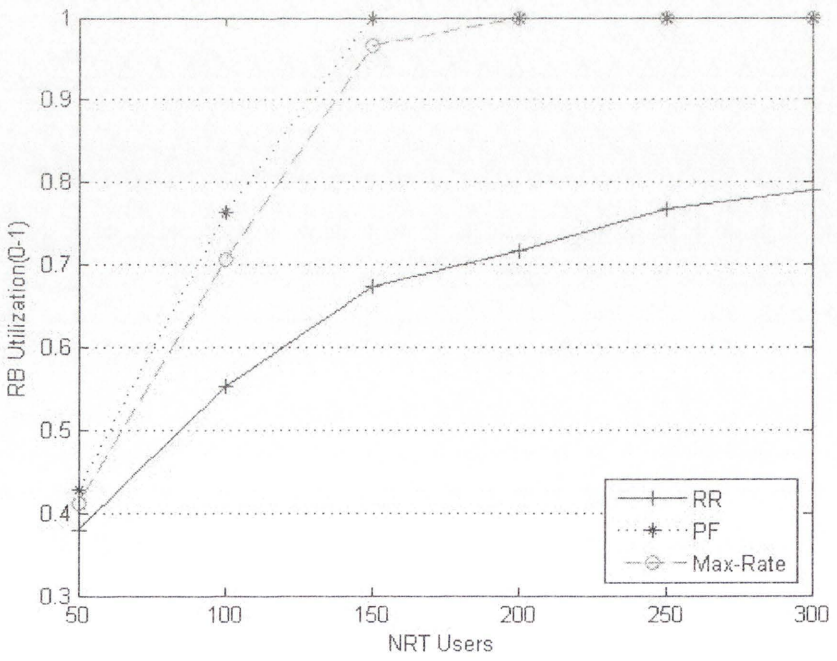

Figure 3-7: RB Utilization vs. Number of NRT Users

50% RT and 50% NRT Scenario

This sub-section considers system supporting equal number of RT and NRT service users. The evaluated packet scheduling algorithms are RR, PF, Max-Rate, M-LWDF and EXP/PF.

As shown in Figure 3-8, M-LWDF has the highest system throughput, followed by the Max-Rate and PF. EXP/PF and RR have a relatively lower system throughput when compared with other algorithms.

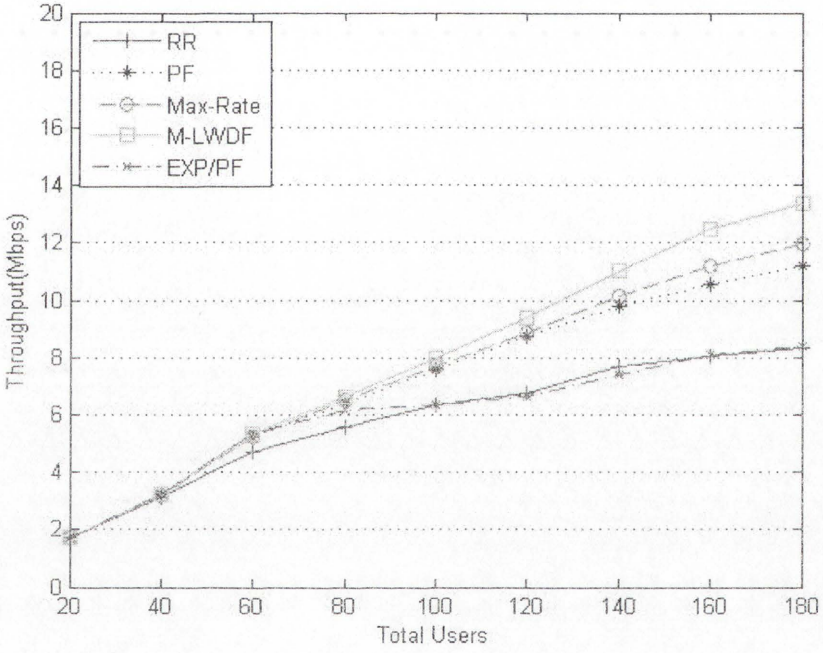

Figure 3-8: System Throughput vs. Number of Users

Figure 3-9 and Figure 3-10 display the delay and PLR performance for RT users. It can be observed that EXP/PF can support the largest number of RT users (70 RT users). The second best is M-LWDF with 60 RT users. RR achieves the worst delay and PLR performance. Max-Rate and PF performs slightly better than RR. Max-Rate, PF and RR can only support up to 20 RT users.

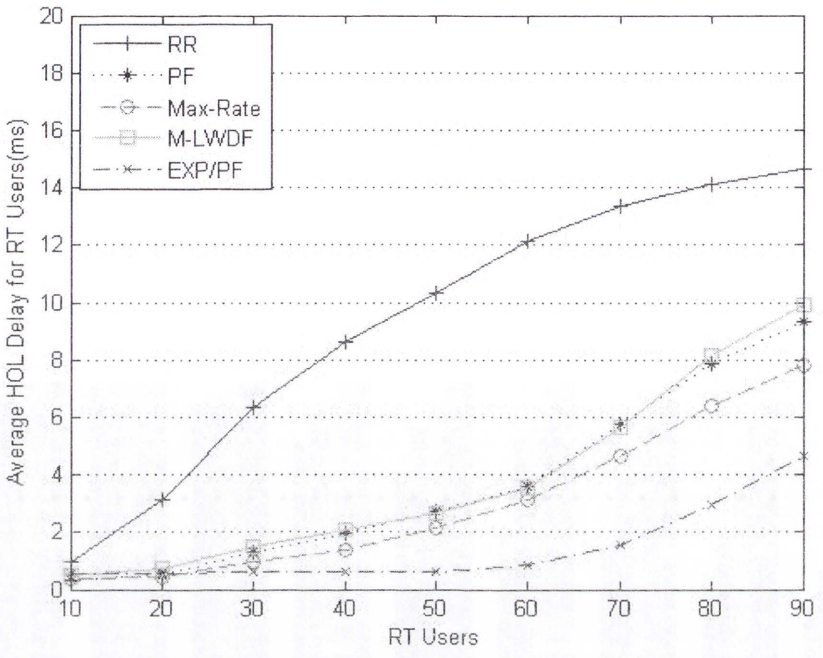

Figure 3-9: Average System HOL Delay for RT Users vs. Number of RT Users

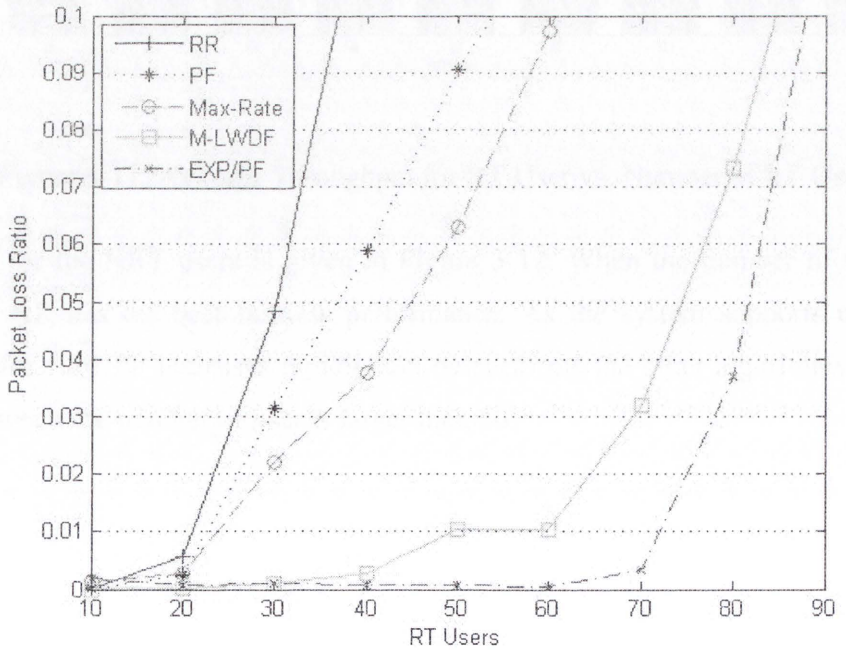

Figure 3-10: PLR for RT Users vs. Number of RT Users

Figure 3-11 gives the average throughput for each RT user. As discussed earlier, the RT users require an average throughput of no less than 100 kbps in order to avoid the buffer running dry. As shown in the figure, the required RT throughput can be maintained for up to 20 RT users (RR), 30 RT users (PF), 40 RT users (Max-Rate), 70 RT users (M-LWDF) and 80 RT users (EXP/PF), respectively.

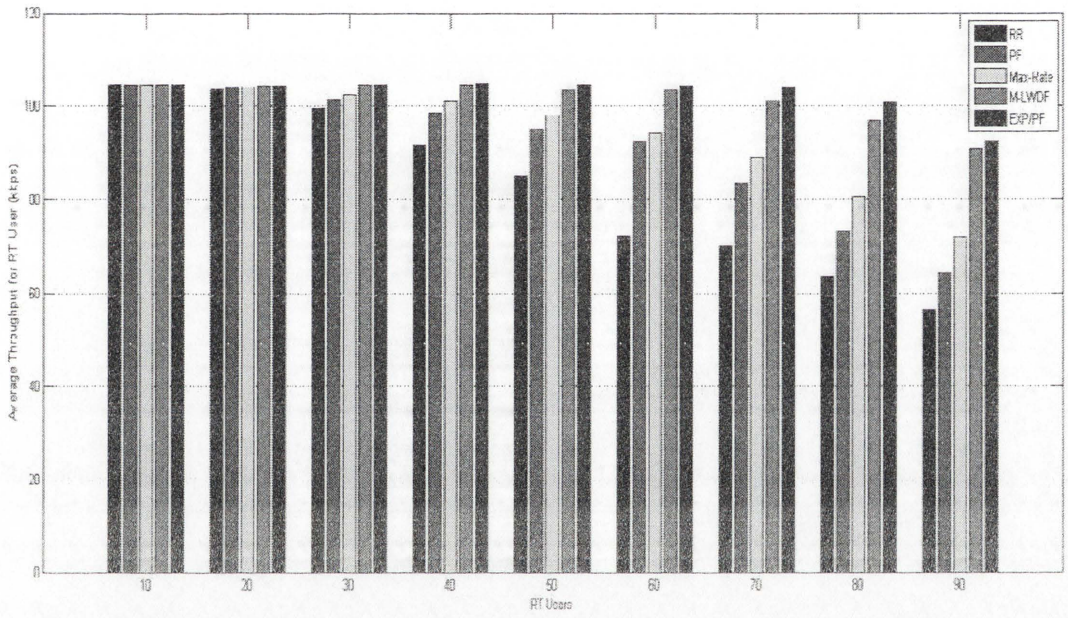

Figure 3-11: Average Throughput for RT User vs. Number of RT Users

Fairness for the NRT users is given in Figure 3-12. When the number of user is less than 30, RR has the best fairness performance. As the system supports more users, fairness for EXP/PF increases rapidly and outperforms the other algorithms in term of fairness when the number of user is larger then 50.

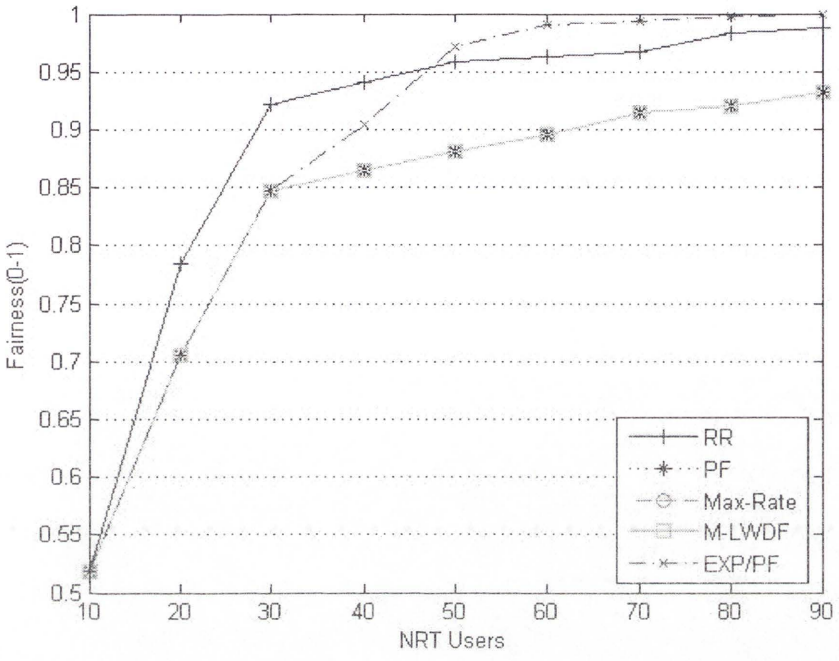

Figure 3-12: Fairness for NRT Users vs. Number of NRT Users

As shown in Figure 3-13, RB utilization of RR is much lower than that of other algorithms.

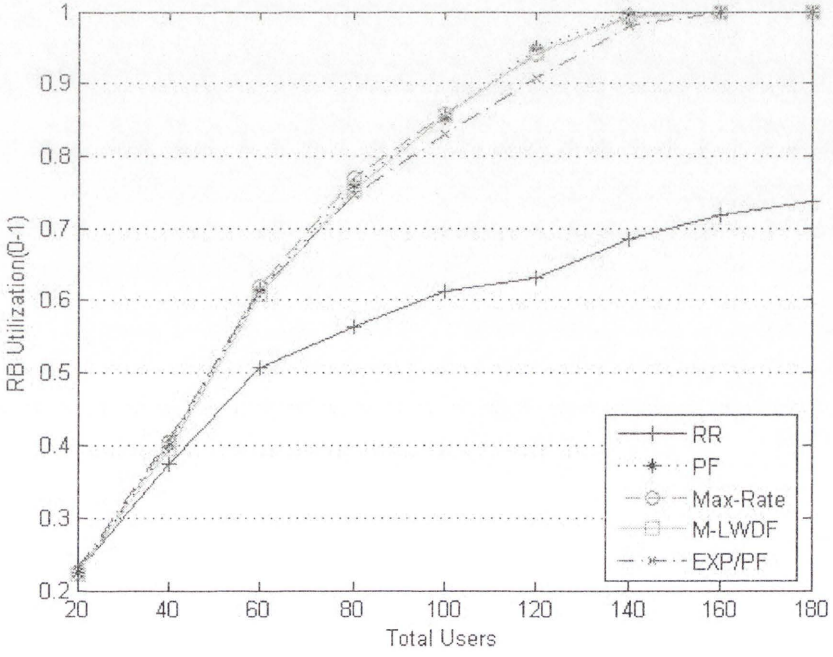

Figure 3-13: RB Utilization vs. Number of Users

Performance comparison of well-known packet scheduling algorithms in 50% RT and 50% NRT scenario is given in Table 3-4.

Table 3-4. Performance Comparison of Packet Scheduling Algorithms

	RR	PF	Max-Rate	M-LWDF	EXP/PF
Throughput	Adequate	Good	Good	Good	Adequate
HOL Delay	Bad	Adequate	Adequate	Adequate	Good
PLR	Bad	Bad	Bad	Good	Good
Fairness	Good	Adequate	Adequate	Adequate	Good
RB Utilization	Bad	Good	Good	Good	Good

To sum up, jointly considering all the requirements of RT users, EXP/PF is able to support the highest number of users (140 users), followed by M-LWDF (120 users). Max-Rate, PF and RR can only support up to 40 users. But this advantage of EXP/PF is achieved at the expense of sacrificing the system throughput. EXP/PF and RR have the worst throughput performance while M-LWDF has the highest system throughput, followed by Max-Rate and PF.

3.3.2 Performance Comparison of Recently Proposed Packet Scheduling Algorithms

Jeongsik Park's Algorithm and Sun Qiaoyun's Algorithm, which have been discussed in Section 3.2.7 and Section 3.2.8, respectively, will be evaluated in this subsection. In order to identify the suitability of these two recently proposed packet scheduling algorithms in the downlink LTE system, the evaluation results of these two algorithms will be compared with that of RR and M-LWDF.

The same channel and traffic model is deployed as used in Section 3.3.1. As both algorithms are proposed to support RT services, only RT users will be chosen in our simulation. The related parameters are given in Table 3-1 and Table 3-2.

The simulation results are given in the following figures.

Figure 3-14 gives system throughput performance of the evaluated packet scheduling algorithms. From the figure, it can be observed that Jeongsik Park's Algorithm and Sun Qiaoyun's Algorithm have the same system throughput when the system has less than 70 users. As the number of users increases above 70, Sun Qiaoyun's Algorithm obtains a higher system throughput when compared with Jeongsik Park's Algorithm. Both algorithms achieve a slightly worse throughput performance when compared with M-LWDF but much better than RR.

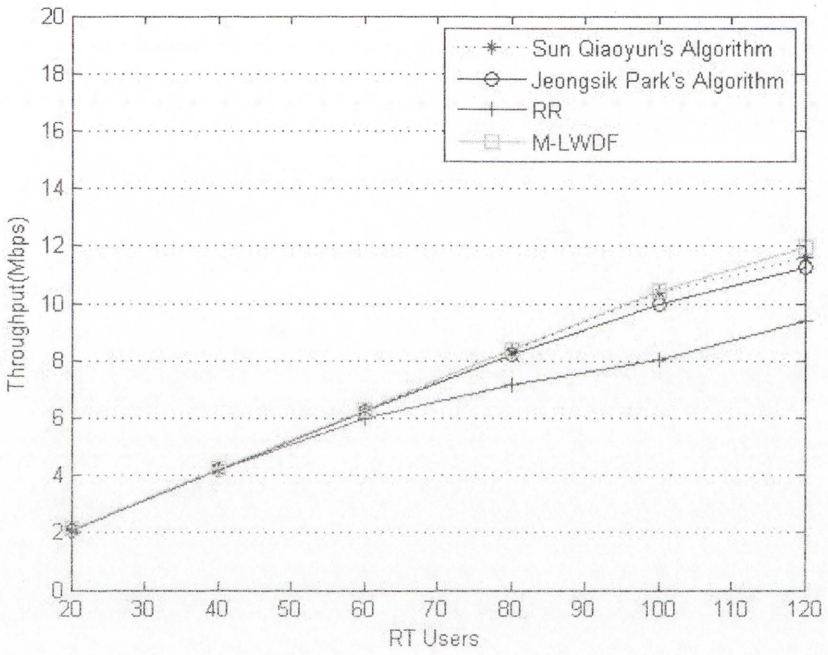

Figure 3-14: System Throughput vs. System Load

Figure 3-15 shows the average system HOL delay of the Jeongsik Park's Algorithm, Sun Qiaoyun's Algorithm, RR and M-LWDF with increasing number of users. From the figure, we can see that Sun Qiaoyun's Algorithm is able to maintain a lower average system HOL delay with increasing number of users when compared with Jeongsik Park's Algorithm. In the four algorithms, users in M-LWDF have the shortest waiting time before being given the opportunity to transmit their packets. Sun Qiaoyun's Algorithm performs the second best in term of the average system HOL delay while Jeongsik Park's Algorithm has the worst performance.

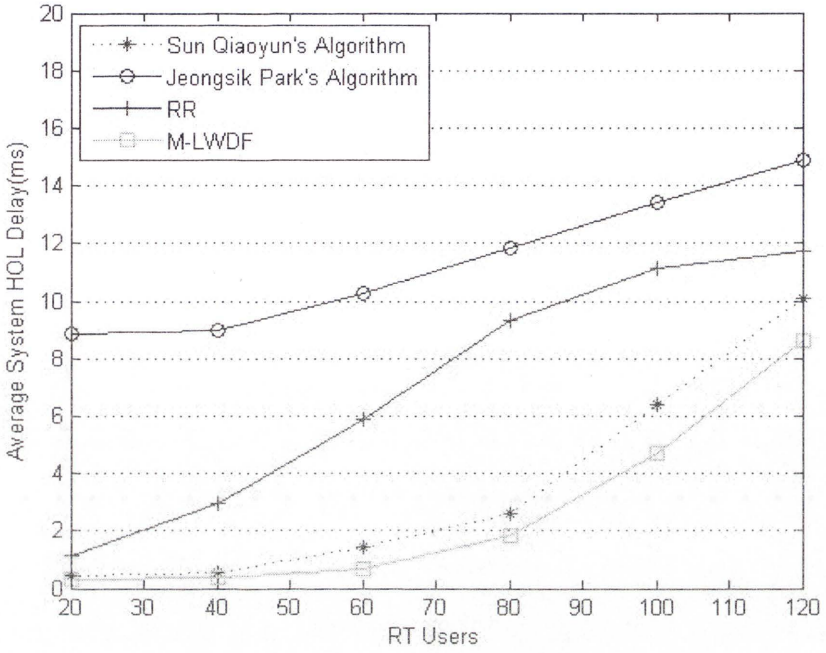

Figure 3-15: Average System HOL Delay vs. System Load

The packet loss ratio performance of the evaluated algorithms is given in Figure 3-16. We can see that to meet the 1% PLR requirement, Sun Qiaoyun's Algorithm can support more users compared to Jeongsik Park's Algorithm which is 100 and 60 users, respectively. Both algorithms have better PLR performance than RR, which can only support 40 users. Sun Qiaoyun's Algorithm has the similar performance with M-LWDF, with up to 100 users. The PLR performance comparison of the four algorithms is given in Table 3-5.

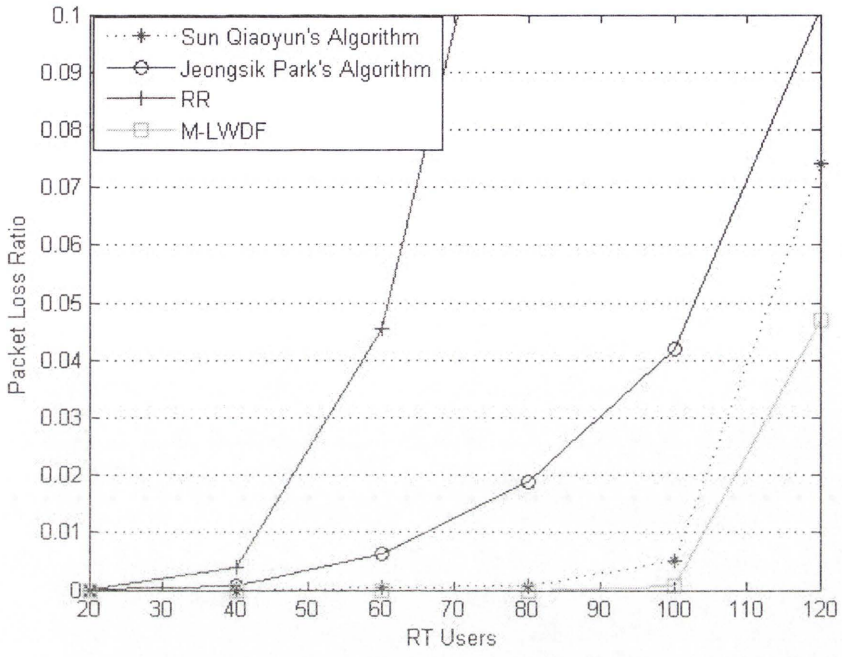

Figure 3-16: Packet Loss Ratio vs. System Load

Table 3-5. PLR Performance Comparison

Algorithm	RR	Jeongsik Park	Sun Qiaoyun	M-LWDF
Maximum Number of Users with PLR<1%	40	60	100	100

Figure 3-17 shows the RB utilization comparison of the selected algorithms. In the figure, it can be seen that Sun Qiaoyun's Algorithm and M-LWDF have an outstanding RB utilization performance, especially when the number of active users is above 80. Jeongsik Park's Algorithm and RR have the comparatively worse RB utilization performance.

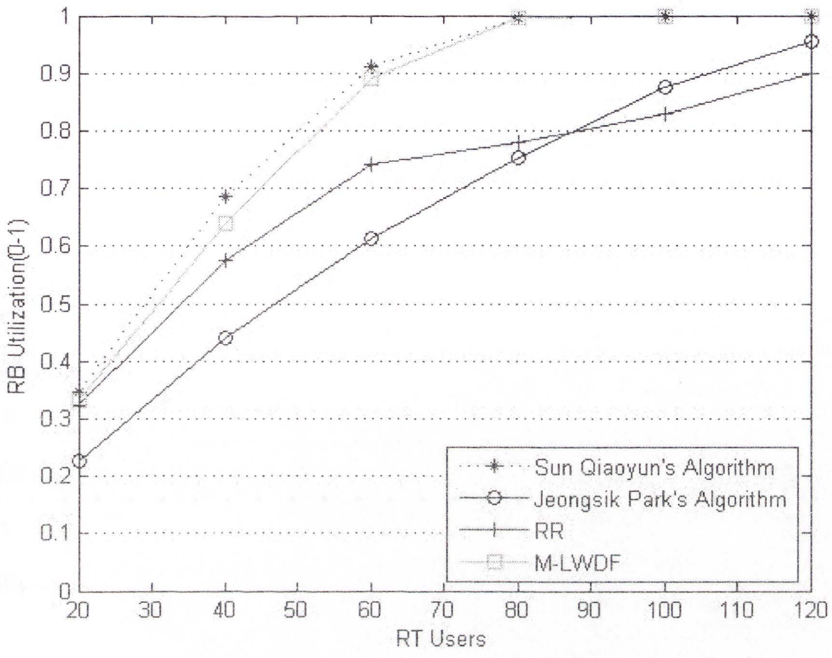

Figure 3-17: RB Utilization vs. System Load

Simulation results show that the two evaluated recently proposed algorithms achieve the similar throughput performance. On the other hand, when compared with Jeongsik Park's Algorithm, Sun Qiaoyun's Algorithm achieves a lower packet delay, a lower packet loss ratio and a better RB utilization with increasing number of users. Therefore, Sun Qiaoyun's Algorithm outperforms Jeongsik Park's Algorithm.

According to the simulation results, the two evaluated recently proposed algorithms outperform RR but are not as good as M-LWDF in terms of throughput and PLR. When compared with RR, Sun Qiaoyun's Algorithm improves the delay and RB utilization performance while Jeongsik Park's Algorithm has a comparatively higher average system HOL delay and lower RB utilization.

Therefore, we conclude that both of the evaluated recently proposed algorithms can be considered as PS candidates. Sun Qiaoyun's Algorithm is more appropriate than Jeongsik Park's Algorithm for the downlink 3GPP LTE system supporting the real-time traffic environment.

3.4 Summary

Five performance metrics of packet scheduling algorithms are discussed in this chapter. After that, six well-known packet scheduling algorithms and two recently proposed packet scheduling algorithms are reviewed in detail. The performance of these algorithms is compared in three different scenarios under a MATLAB simulation environment. The simulation results for well-known algorithms show that M-LWDF outperforms other algorithms in the 100% RT scenario, while EXP/PF is comparatively more suitable for the 50% RT and 50% NRT scenario. In the 100% NRT scenario, PF and Max-Rate achieve a good throughput and RB utilization performance while RR has the best fairness performance. For the recently proposed algorithms, Sun Qiaoyun's Algorithm is more appropriate than Jeongsik Park's Algorithm for the downlink LTE system supporting RT traffic.

Chapter 4

THEORETICAL DELAY ANALYSIS FOR OFDMA SYSTEM

This chapter discusses the theoretical delay analysis of OFDMA system. A downlink mobile network based on the OFDMA technology with Voice-over-IP (VoIP) traffic is considered. The Hybrid-Automatic Repeat Request (HARQ) is deployed to improve system performance. A brief introduction of VoIP and HARQ is provided, followed by the detailed discussion of the analytical models of delay.

4.1 Voice-over-IP (VoIP)

VoIP [43, 44] is a technology used for the delivery of voice traffic over the packet-switched Internet Protocol (IP) networks.

VoIP can be considered as an alternative to the traditional telephone network and has achieved great success in the last decade. Compared with the traditional telephone network, VoIP achieves a higher bandwidth efficiency and facilitates a better cooperation of the voice communication with multimedia applications [45].

The voice traffic is represented as talk spurts which can contain a group of packets. Figure 4-1 illustrates the packet stream for an actual voice traffic. All packets are generated with a fixed time interval. Depending on whether the actual speech power is above a threshold energy level, either an empty packet or a non-empty packet is generated at each time period.

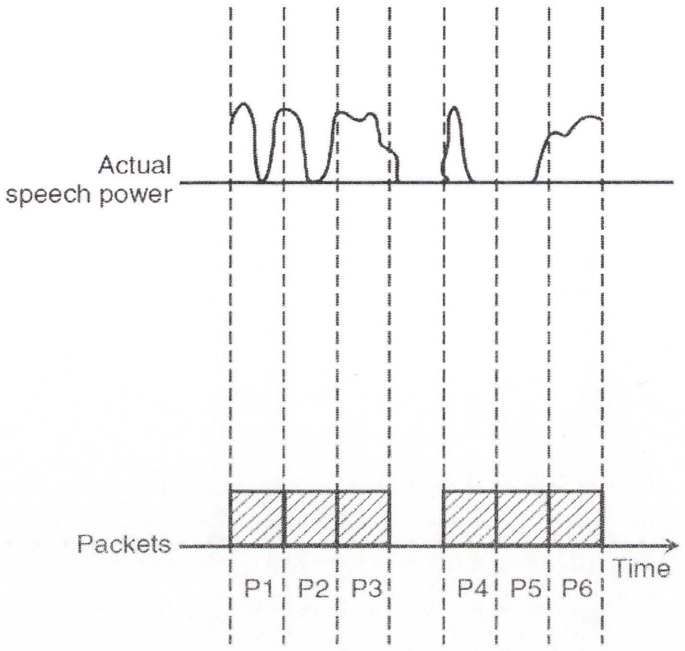

Figure 4-1: Packet Stream for an Actual Voice Traffic [43]

According to the traffic models conducted by P.T. Brady [46, 47], voice connections can be modelled as the ON-OFF pattern. The ON-OFF model is given in Figure 4-2. The ON period represents the talk spurt which is the voice traffic over IP; while the OFF period represents a period of silence. We assume that during the ON period, voice packets are generated at full-rate with a fixed inter-arrival time *T*. Within the silent period, a speaker generates empty packets.

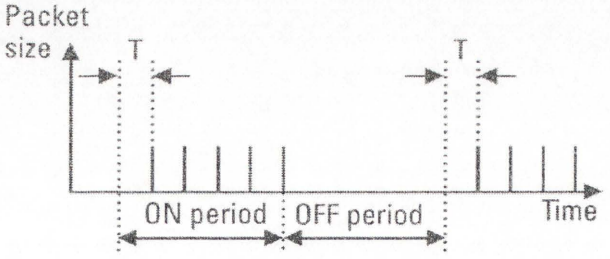

Figure 4-2: Characteristics of Voice Connections [40]

4.2 Hybrid-Automatic Repeat Request (HARQ)

Automatic Repeat-reQuest (ARQ) [48] is an error control mechanism used to guarantee the reliability of data transmission. It uses the acknowledgements and timeouts to

achieve the reliable transmission. An acknowledgement is the message sent by the receiver to the transmitter indicating whether a data packet has been correctly received or not. The timeout is the allowed waiting time for the reception of an acknowledgement.

The simplest ARQ scheme is the Stop-And-Wait (SAW) ARQ. An example of SAW ARQ protocol is illustrated in Figure 4-3. The transmitter sends one data packet at a time. After sending each packet, the transmitter does not send any new packets until it receives the acknowledgement from the receiver. If the data packet has been successfully received, the receiver sends acknowledge (ACK) signal to the transmitter and the new data packet will be transmitted after the ACK signal is received by the transmitter. If the received data packet can not be successfully decoded, negative acknowledgement (NACK) signal is sent and the transmitter will retransmit the missing data packet after receiving the NACK signal. If the acknowledgement signal does not reach the transmitter after a specific timeout, the same data packet will be retransmitted.

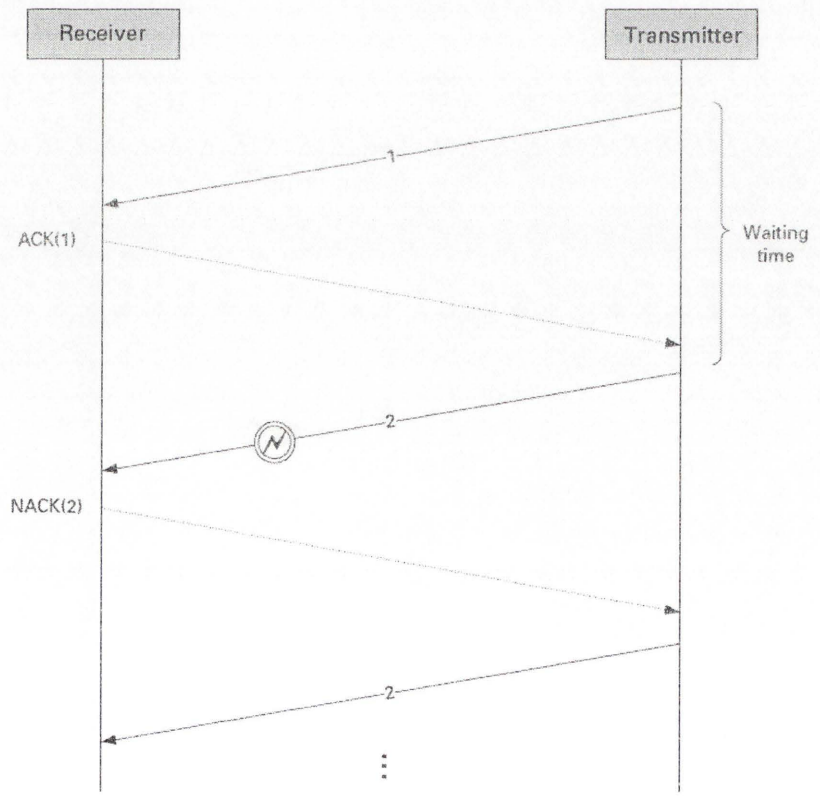

Figure 4-3: Stop-and-Wait (SAW) ARQ Protocol [48]

The *N*-interlace (or *N*-channel) stop-and-wait protocol [48] is a variant of the SAW protocol for the multi-channel system. As shown in Figure 4-4, each channel operates following the simple SAW protocol. Only one channel transmits the data packet at a time and the data packets are transmitted on the *N* channels in sequence order. The transmitter can start transmission on one channel when it is still waiting for the acknowledgement on the previous channels. In order to provide the continuous transmission, the number of channels *N* should be equal to the round-trip time (RTT), which is defined as the duration from the time the transmitter sends the packet to the time ACK or NACK signal reaches the transmitter [48].

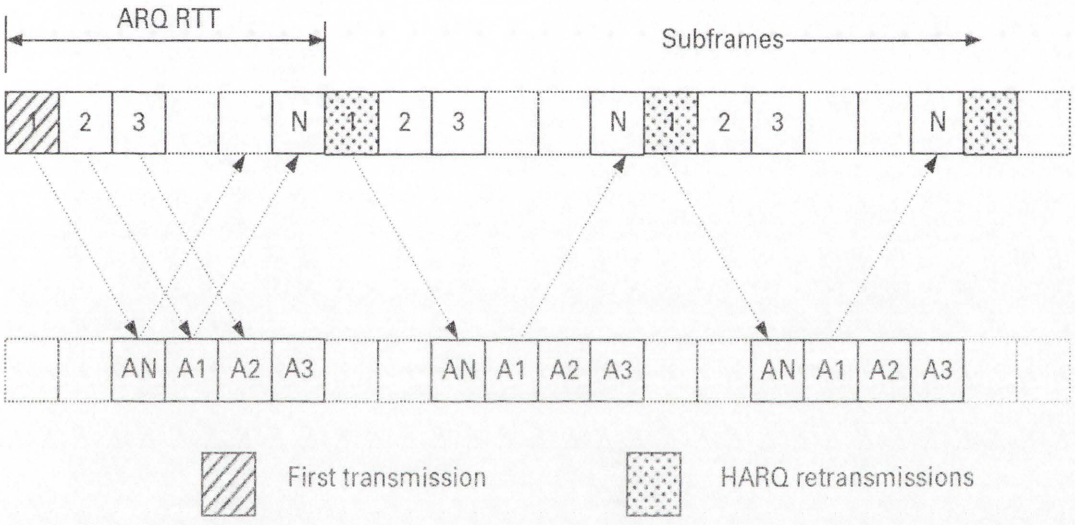

Figure 4-4: *N*-Interlace Stop-and-Wait (SAW) Protocol [48]

The *N*-interlace SAW protocol is chosen as the hybrid ARQ (HARQ) protocol for the LTE system, due to its desirable features such as simplicity, low ACK/NACK feedback overhead, low buffering requirement and so on [48].

4.3 Analytical Model of Delay for OFDMA System with VoIP Traffic

This section discusses the theoretical delay analysis of a downlink mobile network based on the OFDMA technology with VoIP traffic. The HARQ is deployed to improve system performance. As discussed in Section 4.2, the *N*-interlace SAW protocol is

adopted as the HARQ protocol. We denote the smallest scheduling resource unit as a 'tile'. Then the tile-interlace resources are assigned by the schedulers.

As discussed in Section 4.1, the VoIP traffic model consists of talk spurt level and voice packet level. The analytical models for latency based on both of the two levels are discussed in the following subsections [49].

4.3.1 Analytical Model for Talk Spurt Latency

The talk spurt latency involves both assignment and signalling latency. To initiate a new talk spurt transmission, the scheduler's assignment decision is sent to user through the signalling channel, followed by the transmission of voice packets through the assigned traffic channel. The queuing model used for talk spurt latency analysis is illustrated in Figure 4-5, in which λ represents the average talk spurt arrival rate, μ_1 and μ_2 are the average service rates of the signalling channel and traffic channel, respectively, and m_1 and m_2 denote the numbers of available tile-interlace resources within one interlace period for signalling transmission and traffic transmission, respectively. The interlace period equals to the RTT as discussed in Section 4.2,

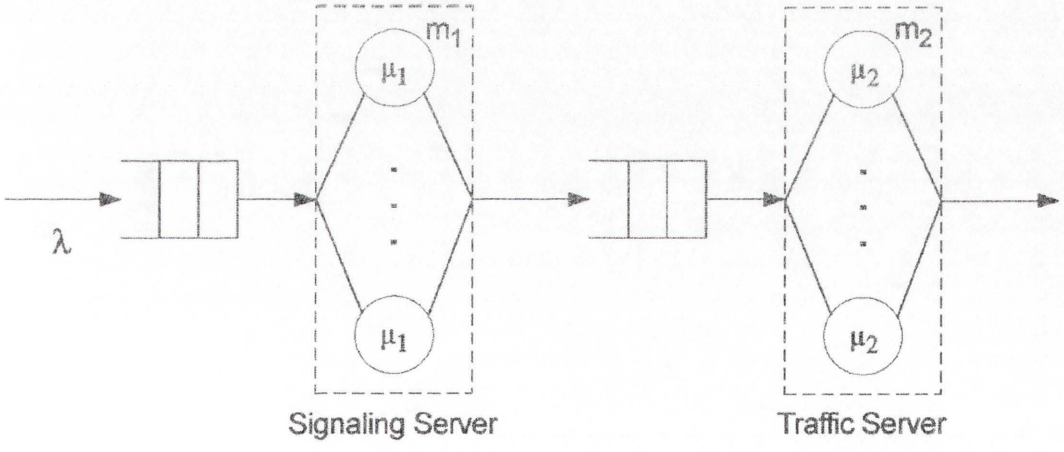

Figure 4-5: Queuing Model Used for Talk Spurt Resource Assignment Latency Analysis [49]

The system is modelled as M/M/m queues, which is equivalent to the case that each queue independently follows the Poisson arrival process. The probabilities of the jth queue having n talk spurts are given by [50]

$$P_j(n) = P_j(0)\frac{m_j^{\min(m_j,n)}\rho_j^n}{\min(n,m_j)!}, \quad for\ j = 1\ or\ 2, \tag{4-1}$$

where

$$\rho_j = \frac{\lambda}{m_j\mu_j}, \tag{4-2}$$

and

$$P_j(0) = \left[1 + \sum_{k=1}^{m_j-1}\frac{(m_j\rho_j)^k}{k!} + \sum_{k=m_j}^{\infty}\frac{(m_j\rho_j)^k}{m_j!m_j^{k-m_j}}\right] \tag{4-3}$$

in which j represents both the index of the queue and the index for describing m, μ and ρ.

If a talk spurt is waiting for resource assignment, the service time s_j of the servers follows an exponential distribution as

$$f_{s_j}(t) = m_j\mu_j e^{-m_j\mu_j t}. \tag{4-4}$$

If a talk spurt has already been assigned resources, then the service time $\tilde{s}_j$ of the servers follows an exponential distribution as

$$f_{\tilde{s}_j}(t) = \mu_j e^{-\mu_j t}. \tag{4-5}$$

If there is not larger than m_j talk spurts waiting in the jth queue, each talk spurt will be allocated one tile-interlace resource and all the talk spurts can be served immediately; otherwise, the server can not serve all the talk spurts at one time and the talk spurts have to wait in the queue for transmission.

The waiting times in the signalling server and traffic server are denoted as w_{b1} and w_{b2}, respectively. These waiting times can be updated by

$$w_{b_j} = \sum_{n=0}^{\infty} P_j(n+m_j)\left(\underbrace{s_j + \cdots + s_j}_{n+1}\right). \tag{4-6}$$

According to the probability theory, the probability density function of independent random variables is the convolution of the probability density functions of each variable [51]. Thus the distribution of the waiting times in the signalling server and traffic server can be computed from

$$f_{w_{b_j}}(t) = \sum_{n=0}^{\infty} P_j(n+m_j)\left(\underbrace{f_{s_j}(t) \otimes \cdots \otimes f_{s_j}(t)}_{n+1}\right). \tag{4-7}$$

where $\otimes$ denotes convolution.

The total queuing delay w_b seen by a new arrived talk spurt is given as

$$w_b = w_{b1} + \tilde{s} + w_{b2} \tag{4-8}$$

Thus, the distribution of talk spurt resource assignment latency can be obtained as

$$f_{w_b}(t) = f_{w_{b_1}}(t) \otimes f_{\tilde{s}_1}(t) \otimes f_{w_{b_2}}(t) \tag{4-9}$$

4.3.2 Analytical Model for Voice Packet Latency

Assume that the enhanced variable rate coder (EVRC) [52] is employed. All voice packets within a talk spurt are generated in regular intervals T_0 at full rate.

The talk spurt resource allocation and HARQ timeline is illustrated in Figure 4-6. The talk spurt will be allocated one tile-interlace resource in each HARQ cycle. Packets within the talk spurt will be transmitted in sequence order. The packet will be transmitted only if all the previous packets have been successfully received. If the packet fails to reach the receiver successfully, the packet will be retransmitted. Because interval T_0 can be much larger than the HARQ cycle, the talk spurt might go empty at some time. In this case, the allocated tile-interlace resource will be unused.

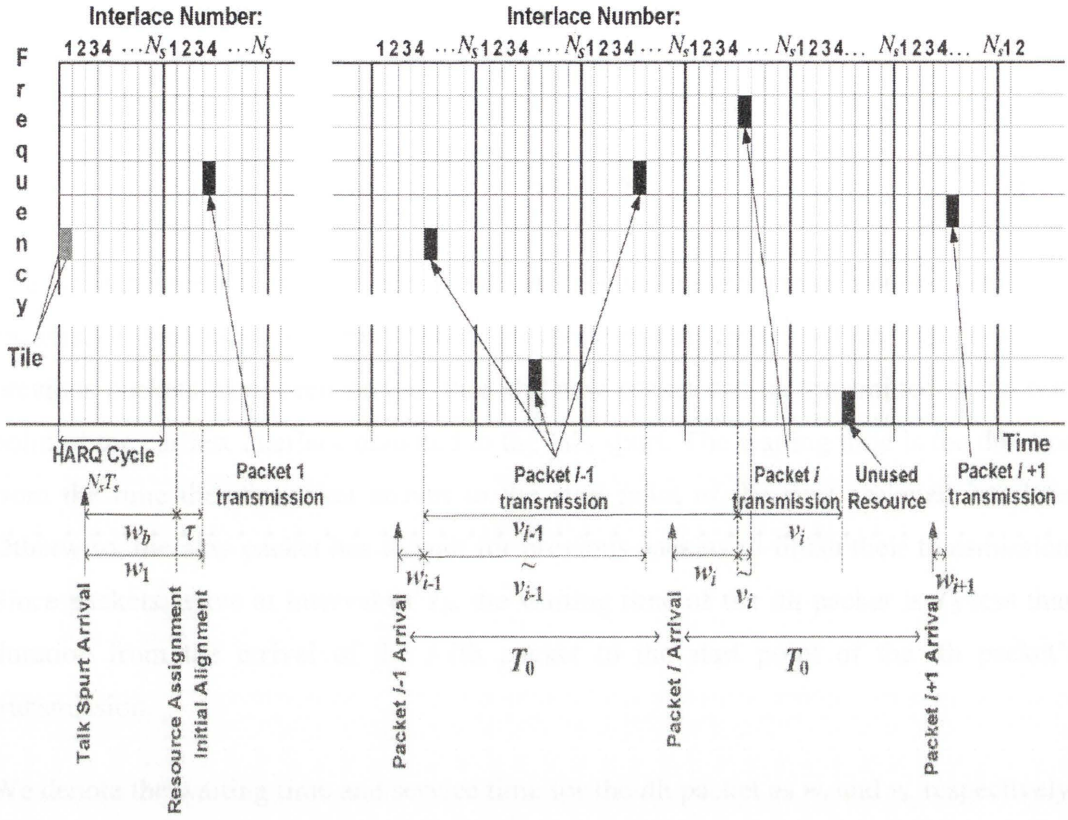

Figure 4-6: Talk Spurt Resource Allocation and HARQ Timeline [49]

For the first packet in the talk spurt, the waiting time is given by

$$w_1 = w_b + \tau, \tag{4-10}$$

in which τ is the delay due to the time taken for assignment of an interlace selected for transmission. This assignment latency follows the uniform distribution within $[0, NsTs)$, where Ts is the duration of each time slot.

The corresponding probability density function is given as

$$f_{w_1}(t) = f_{w_b}(t) \otimes f_{\tau}(t), \tag{4-11}$$

We assume that the talk spurt duration follows the exponential distribution and the cumulative distribution function is given as

$$D(k) = 1 - e^{-k\mu_2 T_0}, \qquad for\ k \geq 1. \tag{4-12}$$

Thus, the distribution of the number of packets in a spurt k can be approximated as

$$p_K(k) = D(k) - D(k-1) = e^{-(k-1)\mu_2 T_0} - e^{-k\mu_2 T_0}, \qquad for\ k \geq 1. \tag{4-13}$$

The waiting time of a newly arrived packet depends on the transmissions of previous packets. If the queue is empty when the new packet arrives which means all the previous packets have been served, then the new packet can be transmitted at the start point of the earliest interlace assigned to the talk spurt. The waiting time is the duration from the time the ith packet arrives to the start point of the next assigned interlace. Otherwise, the new packet has to wait for previous packets to finish their transmission. Since packets arrive at interval of T_0, the waiting time of the ith packet is T_0 less than duration from the arrival of the i-1th packet to the start point of the ith packet's transmission.

We denote the waiting time and service time for the ith packet as w_i and v_i, respectively. The waiting time of the ith packet can be updated by

$$w_i = \begin{cases} w_{i-1} + v_{i-1} - T_0, & w_{i-1} + v_{i-1} \geq T_0 \\ N_s T_s - \mathrm{mod}(T_0 - w_{i-1} - v_{i-1}, N_s T_s), & otherwise \end{cases}, \quad for\ i > 1. \tag{4-14}$$

The service time v_i seen by other packets waiting in the queue is related to the probability of the HARQ early packet termination at nth transmission, e.g.

$$f_v(t) = \sum_{n=1}^{N_{max}} h_n \delta(t - v_n), \qquad v_n = nN_s T_s, \tag{4-15}$$

where N_{max} represents the maximum allowed number of HARQ retransmission, and $\delta(t)$ is the continuous Dirac delta function.

According to (4-14), the distribution of waiting time of the ith packet can recursively be computed by

$$f_{w_i}(t) = \begin{cases} f_{d_{i-1}}(t+T_0), & t \geq N_s T_s \\ \sum_{j=0}^{C(t)} f_{d_{i-1}}(t+T_0 - jN_s T_s), & 0 \leq t < N_s T_s \end{cases}, \tag{4-16}$$

in which $f_{d_{i-1}}(t) = f_{w_{i-1}}(t) \otimes f_v(t)$ is the probability density function of *i-1*th packet's total delay seen by *i*th packet; $C(t) = \lfloor (t+T_0)/N_s T_s \rfloor$ is the number of HARQ cycle from the *i-1*th packet's arrival to the start of *i*th packet's transmission, and $\lfloor x \rfloor$ means the floor value of x.

Then, we can obtain the mathematical expression for the distribution of the total packet waiting time by jointly considering (4-13) and (4-16), which is given as

$$f_w(t) = \frac{1}{\overline{K}} \sum_{i=1}^{\infty} f_{w_i}(t) \sum_{k=i}^{\infty} p_K(k), \tag{4-17}$$

in which $\overline{K} = \sum_{j=1}^{\infty} j p_K(j)$ calculates the average number of packets within the talk spurt.

The total packet delay is defined as

$$\tilde{d} = w + \tilde{v}, \tag{4-18}$$

where $\tilde{v}$ denotes the transmission time seen by the transmitted packet itself and only considers the time between the beginning of the first to the end of the last HARQ transmission and the related probability density function can be obtained by modifying (4-15), e.g.

$$f_{\tilde{v}}(t) = \sum_{n=1}^{N_{max}} h_n \delta(t - \tilde{v}_n), \qquad \tilde{v}_n = (1 + (n-1)N_s)T_s, \tag{4-19}$$

Hence, the distribution of the total packet delay can be mathematically expressed as

$$f_{\tilde{d}}(t) = f_w(t) \otimes f_{\tilde{v}}(t) \tag{4-20}$$

We now have the mathematical expression for the total packet delay in the downlink OFDMA system with VoIP traffic.

4.4 Simulation Result

This subsection gives the visualisation result of the distribution of the talk spurt assignment latency w_b. The talk spurt resource assignment latency is evaluated with arrival rate of 430 bps, 450 bps and 470 bps, respectively. The relevant parameter values are set to m_1=1, m_2=24, T_s=1ms and T_b=50ms.

As given in Figure 4-7, the talk spurt with lower arrival rate achieves a better assignment latency performance, as there will be less arrival packets competing for the transmission.

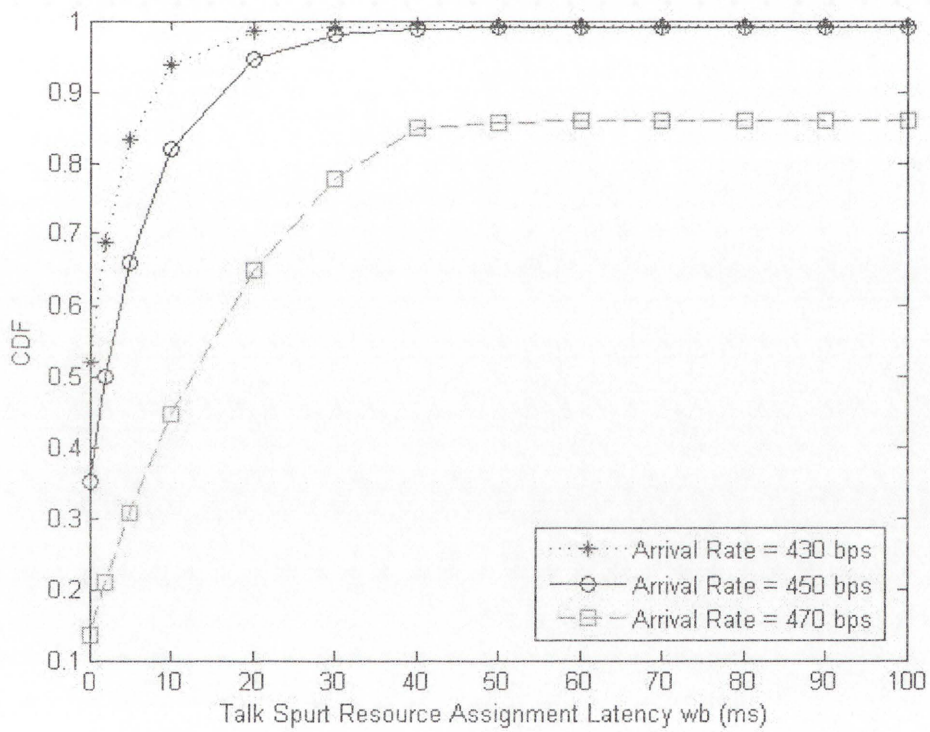

Figure 4-7 Talk Spurt Assignment Latency vs. Average Talk Spurt Arrival Rate

4.5 Summary

This chapter discusses the theoretical delay analysis model for the downlink OFDMA system with VoIP traffic. The HARQ is applied to provide the guaranteed service. A brief introduction of VoIP and HARQ is given. The delay analysis models based on both the talk spurt level and voice packet level are discussed in detail and the

mathematical expression for the total packet delay is obtained after the step-by-step derivations. The simulation result of talk spurt assignment latency distribution $F_{wb}(t)$ is provided.

Chapter 5

THEORETICAL THROUGHPUT ANALYSIS OF PACKET SCHEDULING ALGORITHMS

This chapter discusses the theoretical throughput analysis of packet scheduling algorithms. Based on the step-by-step derivations, we obtain the mathematical expressions of the expected throughput for PF algorithm and M-LWDF algorithm. The visualisation results for the throughput analysis of both algorithms are provided.

5.1 Theoretical Throughput Analysis of PF Algorithm

5.1.1 Throughput Analysis of PF Algorithm

Consider a scenario in which K users are competing for the data transmission from one base station (BS) over Rayleigh fading channel. The proportional fair (PF) algorithm, as described in Section 3.2.4, is adopted by the BS. The theoretical throughput analysis of this system has been discussed in [53-56].

The instantaneous achievable data rate of user i at time $t+1$ is denoted by $r_i(t+1)$. The k-point moving average throughput of user i up to time t is given by $R_i(t)$, which is defined as the average throughput of user i in the last k time slots. The moving average throughput of user i up to time $t+1$ can be updated by

$$R_i(t+1) = (1-\frac{1}{k}) * R_i(t) + \frac{1}{k} * r_i(t), \tag{5-1}$$

in which $I_i(t+1)$ is defined as the indicator function specifying whether user i is scheduled for transmission at time slot $t+1$.

$$I_i(t+1)=\begin{cases}1, & user\ i\ scheduled\ in\ slot\ t+1\\ 0, & else\end{cases}. \tag{5-2}$$

There is a relationship between the Signal to Interference-plus-Noise Ratio (*SINR*) and the instantaneous achievable data rate *r(t)*. P. J. Smith [57] states that in a Rayleigh fading environment, the achievable data rate *r* could be approximated by a Gaussian distribution. For Single-Input-Single-Output (SISO) case, it reduces to

$$E[r]=\int_0^\infty \log(1+SINR\times\lambda)\times e^{-\lambda}d\lambda, \tag{5-3}$$

and

$$\begin{aligned}\sigma_r^2 &= \int_0^\infty \log(1+SINR\times\lambda)^2\times e^{-\lambda}d\lambda\\ &-\left(\int_0^\infty \log(1+SINR\times\lambda)\times e^{-\lambda}d\lambda\right)^2\end{aligned}, \tag{5-4}$$

where *E[r]* and σ_r are the mean value and the standard deviation of *r(t)*.

From (5-1), assuming wide-sense stationary $R_i(t)$, the expected value of the average throughput of user *i* up to time *t+1* is given as

$$\begin{aligned}&E[R_i(t+1)]=E[R_i(t)]\\ &=E[(1-\frac{1}{k})R_i(t)+I_i(t+1)\times\frac{r_i(t)}{k}]\\ &=(1-\frac{1}{k})E[R_i(t)]+\frac{1}{k}E[I_i(t+1)\times r_i(t)]\end{aligned}. \tag{5-5}$$

Hence,

$$E[R_i(t)]=E[I_i(t+1)\times r_i(t)]. \tag{5-6}$$

On substitution (5-2) to (5-6), we can obtain

$$\begin{aligned} E[R_i(t)] &= E[I_i(t+1)\times r_i(t)] \\ &= E\left[1\times r_i(t+1)\middle|I_i(t+1)=1\right]\times \Pr(I_i(t+1)=1) \\ &+ E\left[0\times r_i(t+1)\middle|I_i(t+1)=0\right]\times \Pr(I_i(t+1)=0) \\ &= E\left[1\times r_i(t+1)\middle|I_i(t+1)=1\right]\times \Pr(I_i(t+1)=1) \end{aligned}, \quad (5\text{-}7)$$

where $Pr(I_i(t+1)=1)$ is the probability that user i will be chosen for transmission at time $t+1$.

Applying Bayes's theorem, which is $P(a|b)\times P(b) = P(b|a)\times P(a)$, (5-7) can be written as

$$\begin{aligned} E[R_i(t)] &= \Pr(I_i(t+1)=1)\times \int_0^\infty x f_{r_i}(x|I_i(t+1)=1)dx \\ &= \int_0^\infty x f_{r_i}(x)\Pr(I_i(t+1)=1|r_i(t+1)=x)dx \end{aligned}, \quad (5\text{-}8)$$

where $Pr(I_i(t+1)=1|\ r_i(t+1)=x)$ is the conditional probability that user i will be scheduled to transmit at time $t+1$, if the instantaneous achievable date rate of user i at time $t+1$ is assigned with the value x and $f_{r_i}(\cdot)$ denotes the probability density function of r_i.

According to the scheduling criterion of PF algorithm given in (3-10), user i will be selected for transmission only if any other user j , $j\neq i$, has smaller value of the scheduling criterion than user i, which is $\frac{r_j(t+1)}{R_j(t+1)} < \frac{r_i(t+1)}{R_i(t+1)}$. It holds for large t, k that

$$\begin{aligned} \Pr(I_i(t+1)=1|r_i(t+1)=x) &= \Pr(\forall j\neq i, \frac{r_j(t+1)}{R_j(t+1)} < \frac{r_i(t+1)}{R_i(t+1)}|r_i(t+1)=x) \\ &= \Pr(\forall j\neq i, \frac{r_j(t+1)}{R_j(t+1)} < \frac{x}{R_i(t+1)}) \\ &= \Pr(\forall j\neq i, r_j(t+1) < R_j(t+1)\frac{x}{R_i(t+1)}) \\ &= \prod_{j=1,j\neq i}^{K} F_{r_j}(R_j(t+1)\frac{x}{R_i(t+1)}) \approx \prod_{j=1,j\neq i}^{K} F_{r_j}(\frac{E[R_j]}{E[R_i]}x) \end{aligned}, \quad (5\text{-}9)$$

in which $F_{r_i}(\cdot)$ is the accumulated distribution function of r_i.

For Gaussian distribution r_i as given in (5-3) and (5-4), applying (5-9) to (5-8) yields

$$\begin{aligned} E[R_i(t)] &\approx \int_0^\infty x f_{r_i}(x) \prod_{j=1, j\neq i}^{K} F_{r_j}\left(\frac{E[R_j]}{E[R_i]}x\right)dx \\ &= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \\ &\times \prod_{j=1, j\neq i}^{K} F_{r_j}\left(\frac{E[R_j]}{E[R_i]}(y\sigma_{r_i} + E[r_i])\right) d(y\sigma_{r_i} + E[r_i]) \\ &= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \\ &\times \prod_{j=1, j\neq i}^{K} F_{r_j}\left(\frac{E[R_j]}{E[R_i]}(y\sigma_{r_i} + E[r_i])\right) dy \end{aligned} \tag{5-10}$$

For the instantaneous achievable data rate as described in (5-3) and (5-4), one can verify that

$$\begin{cases} E[r_i] > E[r_j] \text{ and } \dfrac{E[r_i]}{\sigma_{r_i}} > \dfrac{E[r_j]}{\sigma_{r_j}}, & \text{if } \sigma_{r_i} > \sigma_{r_j} \\ E[r_i] > E[r_j], & \text{if } \sigma_{r_i} = \sigma_{r_j} \end{cases} \tag{5-11}$$

Using (5-11), we can prove [54]

$$\frac{E[R_i]E[r_j] - E[R_j]E[r_i]}{E[R_j]\sigma_{r_i} - E[R_i]\sigma_{r_j}} < 0, \quad \text{for} \quad \sigma_{r_i} \neq \sigma_{r_j} \tag{5-12}$$

Case 1

When all σ_{ri} $(i=1,2,\ldots,K)$ are equal, according to (5-11) all users have the same expected value of instantaneous data rate $E[r_i]$ $(i=1,2,\ldots,K)$.

Since $F_{r_i}(x) = F_{(0,1)}((x - E[r_i])/\sigma_{r_i})$ for Gaussian r_i, where $F_{(0,1)}(.)$ denotes the standard normal distribution function with zero mean and unit variance, we have

$$E[R_i(t)] = \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}$$

$$\times \prod_{j=1, j\neq i}^{K} F_{r_j}(\frac{E[R_j]}{E[R_i]}(y\sigma_{r_i} + E[r_i]))dy \qquad (Guess \frac{E[R_j]}{E[R_i]} = \frac{E[r_j]}{E[r_i]} = 1). \quad (5\text{-}13)$$

$$= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} \times (F_{r_j}(y\sigma_{r_i} + E[r_i]))^{K-1} dy$$

Using the assumption that $\sigma_{r_i} = \sigma_{r_j}$ and $E[r_i] = E[r_j]$, we rewrite (5-13) as

$$E[R_i(t)] = \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} \times (F_{r_j}(y\sigma_{r_i} + E[r_i]))^{K-1} dy$$

$$= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} \times (F_{r_j}(y\sigma_{r_i} + E[r_j]))^{K-1} dy \quad . \quad (5\text{-}14)$$

$$= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) f_{(0,1)}(y) \times (F_{(0,1)}(y))^{K-1} dy$$

Case 2

When not all σ_{ri} $(i=1,2,\ldots,K)$ are equal, denote $Z = \arg\max_j \frac{E[R_i]E[r_j] - E[R_j]E[r_i]}{E[R_j]\sigma_{r_i} - E[R_i]\sigma_{r_j}}$.

Then it is can be proved that $Z \geq \arg\max_j \frac{E[R_i]E[r_j] - E[R_j]E[r_i]}{E[R_j]\sigma_{r_i} - E[R_i]\sigma_{r_j}} \geq -\frac{E[r_i]}{\sigma_{r_i}}$

and $Z \leq -\max_j[-\frac{E[r_j]}{\sigma_{r_j}}]$. So (5-10) can be written as

$$E[R_i(t)] = \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$
$$\times \prod_{j=1, j\neq i}^{K} F_{r_j}\left(\frac{E[R_j]}{E[R_i]}(y\sigma_{r_i} + E[r_i])\right) dy$$
$$+ \int_{Z}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$
$$\times \prod_{j=1, j\neq i}^{K} F_{r_j}\left(\frac{E[R_j]}{E[R_i]}(y\sigma_{r_i} + E[r_i])\right) dy \quad . \tag{5-15}$$

Since the first integral in the right hand side of (5-15) is not less than 0, we obtain

$$E[R_i(t)] \geq \int_{Z}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$
$$\times \prod_{j=1, j\neq i}^{K} F_{r_j}\left(\frac{E[R_j]}{E[R_i]}(y\sigma_{r_i} + E[r_i])\right) dy \quad . \tag{5-16}$$

Using (5-12), we can obtain the following equation:

$$\frac{E[R_j]}{E[R_i]}(y\sigma_{r_i} + E[r_i]) > y\sigma_{r_j} + E[r_j] \, . \tag{5-17}$$

Applying (5-16) to (5-17), we then have

$$E[R_i(t)] \geq \int_{Z}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$
$$\times \prod_{j=1, j\neq i}^{K} F_{r_j}\left(\frac{E[R_j]}{E[R_i]}(y\sigma_{r_i} + E[r_i])\right) dy$$
$$\geq \int_{Z}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \quad , \tag{5-18}$$
$$\times \prod_{j=1, j\neq i}^{K} F_{r_j}(y\sigma_{r_j} + E[r_j]) dy$$
$$= \int_{Z}^{\infty} (y\sigma_{r_i} + E[r_i]) f_{(0,1)}(y) \times (F_{(0,1)}(y))^{K-1} dy$$
$$\geq \int_{M}^{\infty} (y\sigma_{r_i} + E[r_i]) f_{(0,1)}(y) \times (F_{(0,1)}(y))^{K-1} dy$$

where $M = -argmax_j[E[r_j]/\sigma_{rj}]$ $(j=1,2,\ldots,K)$.

We express (5-13) and (5-18) by the same equation,

$$
\begin{aligned}
E[R_i(t)] &\geq \int_M^{\infty} (y\sigma_{r_i} + E[r_i]) f_{(0,1)}(y) \times (F_{(0,1)}(y))^{K-1} dy \\
&= \frac{E[r_i]}{K} + \int_M^{\infty} y\sigma_{r_i} f_{(0,1)}(y) \times (F_{(0,1)}(y))^{K-1} dy \\
&= \frac{E[r_i]}{K} + \int_M^{\infty} y\sigma_{r_i} \times \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} \times \left(\frac{1}{2} \times \left(1 + erf(\frac{y}{\sqrt{2}}) \right) \right)^{K-1} dy \\
&= \frac{E[r_i]}{K} + \int_M^{\infty} y\sigma_{r_i} \times \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} \times \left(\frac{1}{2} \times \left(1 + \frac{2}{\sqrt{\pi}} \int_0^{\frac{y}{\sqrt{2}}} e^{-t^2} dt \right) \right)^{K-1} dy
\end{aligned}
\qquad (5\text{-}19)
$$

We now have the mathematical expression for the users' mean throughput when the PF scheduling algorithm is used.

5.1.2 Simulation Result for PF Algorithm

The visualisation result of throughput analysis for PF algorithm, as discussed in Section 5.1.1, will be given in this subsection.

To validate the analytical result (5-19) under different load and SINR, we test the system with the number of users from 1 to 50 and with the fixed SINR value for User 1 at 0.8 dB, 10.8 dB and 20.8 dB, respectively. For any other users, the instantaneous data rate is randomly generated with the mean and standard deviation given by (5-3) and (5-4).

Figure 5-1 describes the theoretical analysis results on the single user's throughput for PF algorithm. With the increasing values of SINR, the user's normalized throughput performance is enhanced. The single user's throughput decreases while the system supports more users.

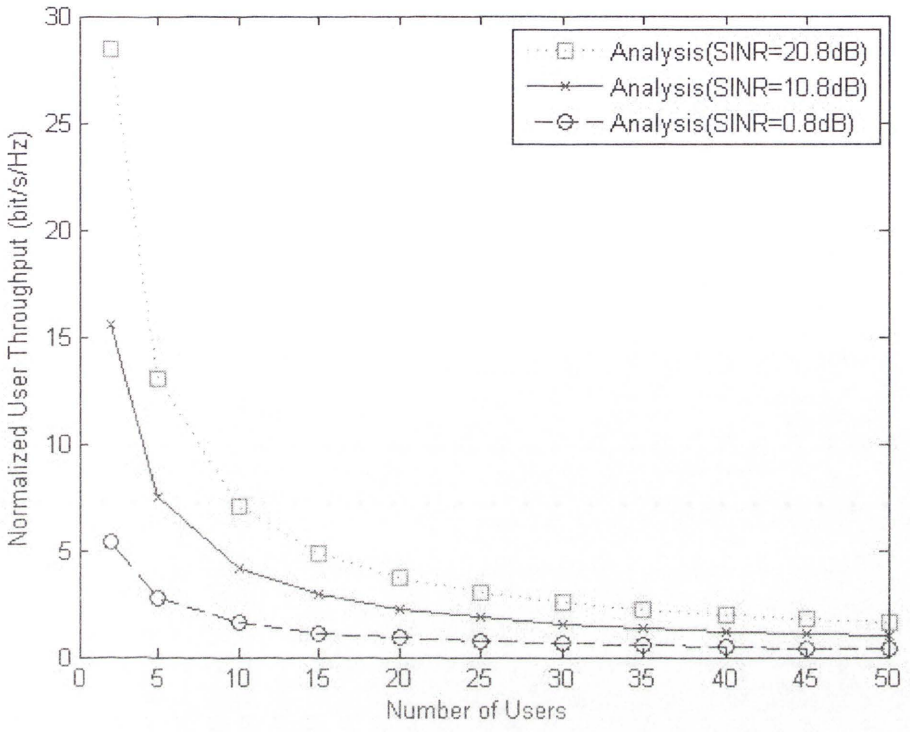

Figure 5-1: Normalized Single User's Throughput for PF Algorithm vs. System Load

If we assume all users have the same channel condition, we evaluate the system with SINR values of 0.8 dB, 10.8 dB and 20.8 dB respectively. The normalized system throughput performance for PF algorithm is given in Figure 5-2. The overall system throughput increases with the increasing system load as well as with increasing SINR values.

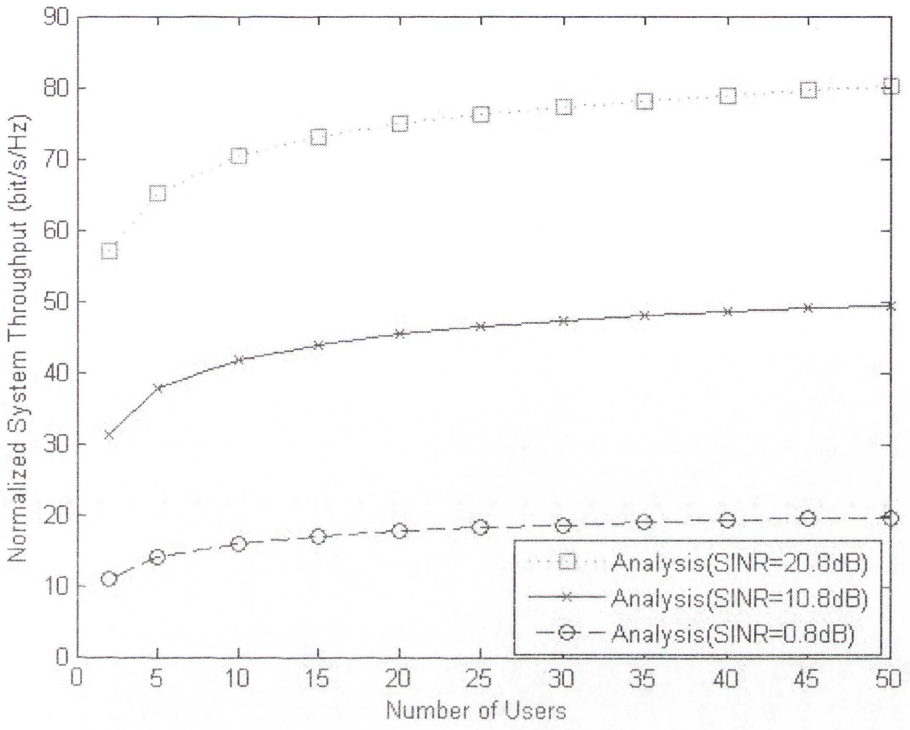

Figure 5-2: Normalized System Throughput for PF Algorithm vs. System Load

Figure 5-3 shows the limit of the normalized system throughput. As can been seen, with the increasing number of users, the normalized system throughput can goes up to approximately 24 bit/s/Hz (SINR=0.8dB), 58 bit/s/Hz (SINR=10.8dB) and 92 bit/s/Hz (SINR=20.8dB), respectively.

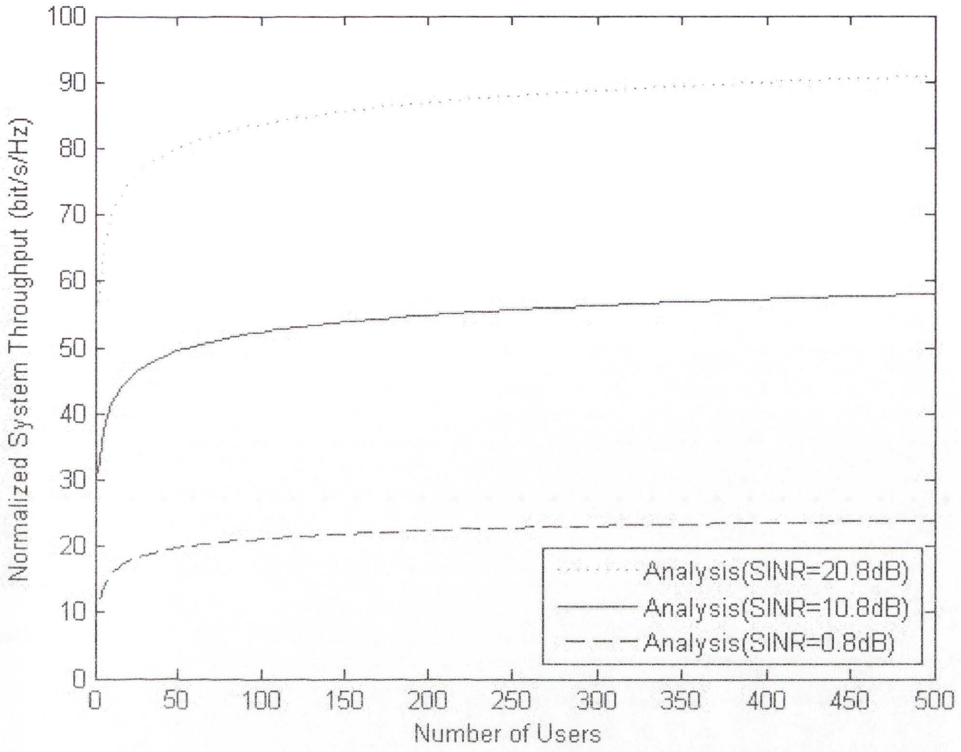

Figure 5-3: Limit of Normalized System Throughput for PF Algorithm

5.2 Theoretical Throughput Analysis of M-LWDF Algorithm

5.2.1 Throughput Analysis of M-LWDF Algorithm

This subsection gives the theoretical throughput analysis of M-LWDF algorithm in the downlink LTE system. The Rayleigh fading system with K users and N RBs is modeled.

Assume that all sub-bands in OFDMA system have independent identical fading characteristic for all users. Thus instantaneous capacities of different users on the same RBs are independent. Then, the average network throughput can be calculated by:

$$Average\ \ Network\ \ Throughput = K \times N \times E[R_{ij}(t)], \tag{5-20}$$

where $R_{ij}(t)$ denotes the average throughput of user i on RB j at time slot n and $E[R_{ij}(t)]$ is the expectation value of $R_{ij}(t)$.

As discussed in Section 5.1, the instantaneous achievable data rate *r(t)* is approximated by the Gaussian distribution and follows (5-3) and (5-4) for the single user case.

If we assume $R_{ij}(t)$ to be wide-sense stationary, then (5-8) can be modified as

$$E[R_{ij}(t)] = \int_0^\infty x f_{r_{ij}}(x) \Pr(I_{ij}(t+1) = 1 \,|\, r_{ij}(t+1) = x) dx \quad , \tag{5-21}$$

where $f_{r_{ij}}(x)$ is the probability density function of r_{ij} and $\Pr(I_{ij}(t+1) = 1 \,|\, r_{ij}(t+1) = x)$ is the conditional probability that user i will be scheduled on RB j at time $t+1$, given that the instantaneous achievable rate of RB j at time $t+1$ is x.

Based on the scheduling criterion of M-LWDF algorithm which has been discussed in Section 3.2.5, for statistically independent r_{ij}, the probability of user i being selected for transmission on each RB at each TTI can be computed by

$$\begin{aligned} &\Pr(I_{ij}(t+1) = 1 \,|\, r_{ij}(t+1) = x) \\ &= \Pr(\forall m \neq i, \mathrm{a_m} W_m(t+1) \frac{r_{mj}(t+1)}{R_{mj}(t+1)} < \mathrm{a_i} W_i(t+1) \frac{r_{ij}(t+1)}{R_{ij}(t+1)}) , \\ &= \Pr(\forall m \neq i, r_{mj}(t+1) W_m(t+1) < \frac{\mathrm{a_i}}{\mathrm{a_m}} \frac{R_{mj}(t+1)}{R_{ij}(t+1)} x W_i(t+1)) \end{aligned} \tag{5-22}$$

in which $W_i(t)$ represents the HOL waiting time of user i at time t.

Further assuming that all users have the same delay requirements (e.g. $a_i = a_m, \forall m \neq i$), it holds for the large values of t and k that

$$
\begin{aligned}
&\Pr(I_{ij}(t+1)=1 \mid r_{ij}(t+1)=x) \\
&=\Pr(\forall m \neq i, r_{mj}(t+1)W_m(t+1) < \frac{a_i}{a_m}\frac{R_{mj}(t+1)}{R_{ij}(t+1)}xW_i(t+1)) \\
&\approx \Pr(\forall m \neq i, r_{mj}(t+1)W_m(t+1) < xW_i(t+1)) \\
&=\int_{w=0}^{\infty} f_{W_i}(w)\times \Pr(\forall m \neq i, r_{mj}(t+1)W_m(t+1) < xw)dw \\
&=\int_{w=0}^{\infty} f_{W_i}(w)\times \prod_{m=1,m\neq i}^{K} F_{r_{mj}W_m}(xw)dw
\end{aligned}
\quad , \qquad (5\text{-}23)
$$

in which f_{Wi} is the probability density function of W_i and $F_{r_{mj}W_m}$ is the product cumulative distribution function of $r_{mj}*W_m$.

On substitution of (5-23) to (5-21), we obtain

$$
\begin{aligned}
E[R_{ij}(t)] &= \int_0^{\infty} x f_{r_{ij}}(x)\int_{\tau=0}^{\infty} f_{W_i}(\tau)\times \prod_{m=1,m\neq i}^{K} F_{r_{mj}W_m}(x\tau)d\tau dx \\
&= \int_0^{\infty} x f_{r_{ij}}(x)\int_{\tau=0}^{\infty} f_{W_i}(\tau)\times \left[\prod_{m=1,m\neq i}^{K}\int_0^{x\tau} f_{r_{mj}W_m}(m)dm\right]d\tau dx
\end{aligned}
\quad , \qquad (5\text{-}24)
$$

where $f_{r_{mj}W_m}$ is the probability density function of $r_{mj}*W_m$.

According to [58], we can get

$$
f_{r_{mj}W_{mi}}(m) = \int_{-\infty}^{\infty}\left|\frac{1}{w}\right| f_{r_{mj},W_m}\left(w,\frac{m}{w}\right)dw . \qquad (5\text{-}25)
$$

Since r_{mj} and W_m are independent and the waiting time is no less than zero, we can rewrite (5-25) as

$$
\begin{aligned}
f_{r_{mj}W_{mi}}(m) &= \int_{-\infty}^{\infty} \left|\frac{1}{w}\right| f_{r_{mj}}(w) f_{W_m}\left(\frac{m}{w}\right) dw \\
&= \int_{0}^{\infty} \frac{1}{w} f_{r_{mj}}(w) f_{W_m}\left(\frac{m}{w}\right) dw
\end{aligned}
\tag{5-26}
$$

On substitution of (5-26) to (5-24), we obtain

$$
\begin{aligned}
E[R_{ij}(t)] &= \int_0^{\infty} x f_{r_{ij}}(x) \int_{\tau=0}^{\infty} f_{W_i}(\tau) \times \left[\prod_{m=1,m\neq i}^{K} \int_0^{x\tau} f_{r_{mj}W_m}(m) dm \right] d\tau dx \\
&= \int_0^{\infty} x f_{r_{ij}}(x) \int_{\tau=0}^{\infty} f_{W_i}(\tau) \times \left[\prod_{m=1,m\neq i}^{K} \int_0^{x\tau} \int_0^{\infty} \frac{1}{w} f_{r_{mj}}(w) f_{W_m}\left(\frac{m}{w}\right) dw dm \right] d\tau dx \\
&= \int_0^{\infty} x f_{r_{ij}}(x) \int_{\tau=0}^{\infty} f_{W_i}(\tau) \times \left[\prod_{m=1,m\neq i}^{K} \int_0^{\infty} \int_0^{x\tau} \frac{1}{w} f_{r_{mj}}(w) f_{W_m}\left(\frac{m}{w}\right) dm dw \right] d\tau dx, \\
&= \int_0^{\infty} x f_{r_{ij}}(x) \int_{\tau=0}^{\infty} f_{W_i}(\tau) \times \left[\prod_{m=1,m\neq i}^{K} \int_0^{\infty} f_{r_{mj}}(w) \int_0^{x\tau} \frac{1}{w} f_{W_m}\left(\frac{m}{w}\right) dm dw \right] d\tau dx \\
&= \int_0^{\infty} x f_{r_{ij}}(x) \int_{\tau=0}^{\infty} f_{W_i}(\tau) \times \left[\prod_{m=1,m\neq i}^{K} \int_0^{\infty} f_{r_{mj}}(w) F_{W_m}\left(\frac{x\tau}{w}\right) dw \right] d\tau dx
\end{aligned}
\tag{5-27}
$$

where F_{Wm} is the cumulative distribution function of W_m.

According to [49], we assume that the HOL waiting time of user i follows an exponential distribution as

$$
f_{W_i}(t) = E[r_i] e^{-E[r_i]t}, \tag{5-28}
$$

and

$$
F_{W_i}(t) = 1 - e^{-E[r_i]t}. \tag{5-29}
$$

Then, (5-27) can be rewritten as

$$E[R_{ij}(t)] = \int_0^\infty x f_{r_{ij}}(x) \int_{\tau=0}^\infty f_{W_i}(\tau) \times \left[\prod_{m=1,m\neq i}^{K} \int_0^\infty f_{r_{mj}}(w) F_{W_m}\left(\frac{x\tau}{w}\right) dw \right] d\tau dx$$

$$= \int_0^\infty x f_{r_{ij}}(x) \int_{\tau=0}^\infty f_{W_i}(\tau) \times \left[\prod_{m=1,m\neq i}^{K} \left[1 - \int_0^\infty f_{r_{mj}}(w) e^{-E[r_m]\frac{x\tau}{w}} dw \right] \right] d\tau dx$$

$$= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^\infty (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^\infty E[r_i] e^{-E[r_i]\tau}$$

$$\times \left[\prod_{m=1,m\neq i}^{K} \left[1 - \int_0^\infty f_{r_{mj}}(w) e^{-E[r_m]\frac{(y\sigma_{r_i}+E[r_i])\tau}{w}} dw \right] \right] d\tau dy$$

$$= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^\infty (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^\infty E[r_i] e^{-E[r_i]\tau}$$

$$\times \left[\prod_{m=1,m\neq i}^{K} \left[1 - \int_{\frac{E[r_m]}{\sigma_{r_m}}}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{m^2}{2}} e^{-E[r_m]\frac{(y\sigma_{r_i}+E[r_i])\tau}{m\sigma_{r_m}+E[r_m]}} \sigma_{r_m} dm \right] \right] d\tau dy$$

$$= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^\infty (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^\infty E[r_i] e^{-E[r_i]\tau}$$

$$\times \left[1 - \int_{\frac{E[r_m]}{\sigma_{r_m}}}^\infty \frac{\sigma_{r_m}}{\sqrt{2\pi}} e^{-\frac{m^2}{2}} e^{-E[r_m]\frac{(y\sigma_{r_i}+E[r_i])\tau}{m\sigma_{r_m}+E[r_m]}} dm \right]^{K-1} d\tau dy$$

$$\leq \int_{\frac{E[r_i]}{\sigma_{r_i}}}^\infty (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^\infty E[r_i] e^{-E[r_i]\tau}$$

$$\times \left[1 - \int_0^\infty \frac{\sigma_{r_m}}{\sqrt{2\pi}} e^{-\frac{m^2}{2}} e^{-E[r_m]\frac{(y\sigma_{r_i}+E[r_i])\tau}{m\sigma_{r_m}}} dm \right]^{K-1} d\tau dy$$

$$= \int_{\frac{E[r_i]}{\sigma_{r_i}}}^\infty (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^\infty E[r_i] e^{-E[r_i]\tau}$$

$$\times \left[1 - \frac{\sigma_{r_m}}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{m^2}{2}} e^{-E[r_m]\frac{(y\sigma_{r_i}+E[r_i])\tau}{m\sigma_{r_m}}} dm \right]^{K-1} d\tau dy \quad . (5\text{-}30)$$

It can be proven that

$$\int_0^\infty e^{-\frac{m^2}{a}} e^{-\frac{b}{m}} dm = \frac{bG_{0,3}^{3,0}\left(\frac{ab^2}{4}\middle| -\frac{1}{2},0,0\right)}{4\sqrt{\pi}}, \ if\ a>0\ and\ b>0, \tag{5-31}$$

where the G-function is the Meijer G-function which is defined as

$$G_{p,q}^{m,n}\left(z\middle|\begin{matrix}a_1,\ldots,a_p\\ b_1,\ldots,b_q\end{matrix}\right) = \frac{1}{2\pi i}\int_L \frac{\prod_{j=1}^{m}\Gamma(b_j-s)\prod_{j=1}^{n}\Gamma(1-a_j+s)}{\prod_{j=m+1}^{q}\Gamma(1-b_j+s)\prod_{j=n+1}^{p}\Gamma(a_j-s)} z^s ds. \tag{5-32}$$

On substitution of (5-31) to (5-30), we can have

$$\begin{aligned}
E[R_{ij}(t)] \le & \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^{\infty} E[r_i] e^{-E[r_i]\tau} \\
& \times \left[1 - \frac{\sigma_{r_m}}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{m^2}{2}} e^{-E[r_m]\frac{(y\sigma_{r_i}+E[r_i])\tau}{m\sigma_{r_m}}} dm \right]^{K-1} d\tau dy \\
= & \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^{\infty} E[r_i] e^{-E[r_i]\tau} \times \left[1 - \frac{\sigma_{r_m}}{\sqrt{2\pi}} \times \frac{1}{4\sqrt{\pi}} \frac{E[r_m]}{\sigma_{r_m}} \right. \\
& \left. \times (y\sigma_{r_i} + E[r_i])\tau \times G_{0,3}^{3,0}\left(\frac{1}{4} \times 2 \times \left(\frac{E[r_m]}{\sigma_{r_m}} (y\sigma_{r_i} + E[r_i])\tau \right)^2 \middle| -\frac{1}{2},0,0 \right) \right]^{K-1} d\tau dy \\
= & \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^{\infty} E[r_i] e^{-E[r_i]\tau} \times \left[1 - \frac{E[r_m]}{4\sqrt{2\pi}} (y\sigma_{r_i} + E[r_i])\tau \right. \\
& \left. \times G_{0,3}^{3,0}\left(\frac{1}{2} \times \left(\frac{E[r_m]}{\sigma_{r_m}} (y\sigma_{r_i} + E[r_i])\tau \right)^2 \middle| -\frac{1}{2},0,0 \right) \right]^{K-1} d\tau dy
\end{aligned} \tag{5-33}$$

Finally, the theoretical average network throughput for M-LWDF algorithm can be expressed as:

Average Network Throughput

$$\leq K \times N \times \int_{\frac{E[r_i]}{\sigma_{r_i}}}^{\infty} (y\sigma_{r_i} + E[r_i]) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_{\tau=0}^{\infty} E[r_i] e^{-E[r_i]\tau} \times \left[1 - \frac{E[r_m]}{4\sqrt{2\pi}} \times \right. \quad (5\text{-}34)$$

$$\left. (y\sigma_{r_i} + E[r_i])\tau \times G_{0,3}^{3,0}\left(\frac{1}{2} \times \left(\frac{E[r_m]}{\sigma_{r_m}} (y\sigma_{r_i} + E[r_i])\tau\right)^2 \middle| -\frac{1}{2}, 0, 0\right)\right]^{K-1} d\tau dy$$

5.2.2 Simulation Result for M-LWDF Algorithm

The visualisation result of the theoretical throughput analysis for M-LWDF, as discussed in Section 5.2.1, will be given in this subsection. The analytical result (5-34) is evaluated with the number of users from 1 to 50 and with the fixed SINR value for User 1 at 0.8 dB, 10.8 dB and 20.8 dB, respectively. For any other users, the instantaneous data rate is randomly generated with the mean and standard deviation given by (5-3) and (5-4).

The normalized throughput of User 1 for M-LWDF algorithm is illustrated in Figure 5-4. The user achieves a higher throughput with the larger SINR value. When there are more users competing for transmission, the single user is allocated less radio resource and has a less throughput.

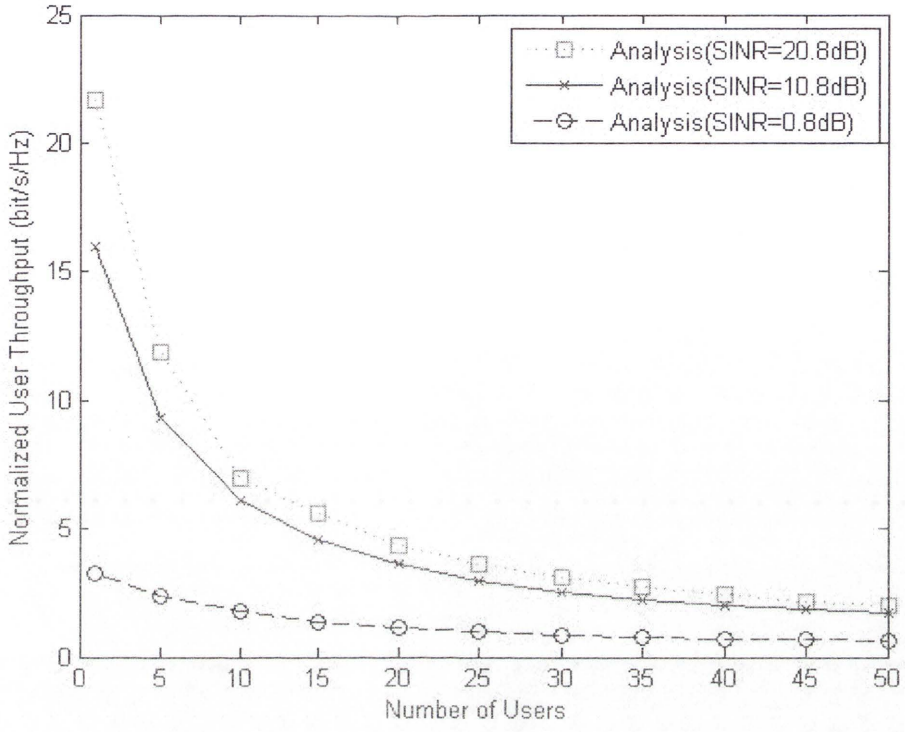

Figure 5-4: Normalized Single User's Throughput for M-LWDF Algorithm vs. System Load

Figure 5-5 gives the normalized system throughput analysis result for M-LWDF algorithm. The normalized system throughput goes up with the increasing number of users. The system with the higher SINR value has the larger normalized system throughput.

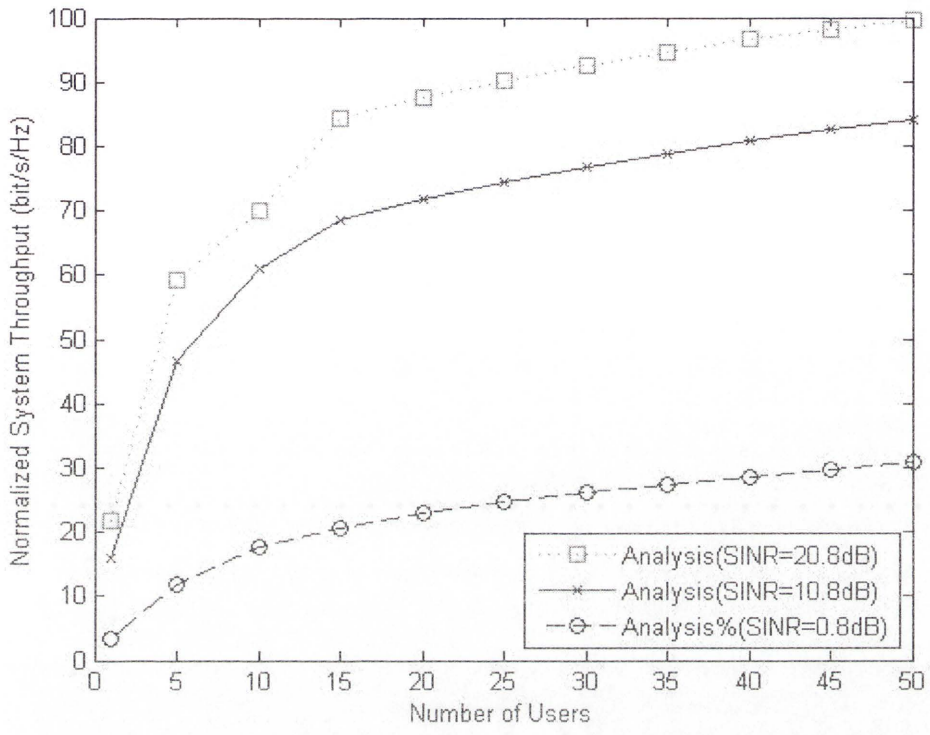

Figure 5-5: Normalized System Throughput for M-LWDF Algorithm vs. System Load

5.3 Summary

This chapter provides the detailed throughput analysis of PF algorithm and M-LWDF algorithm. We obtain the mathematical expressions of the expected throughput for both algorithms. The visualisation results of the theoretical analysis for both algorithms are provided.

Chapter 6

CONCLUSIONS AND FUTURE RESEARCH WORK

This chapter summarizes the thesis and discusses the potential research directions for the future.

6.1 Conclusion

Packet scheduling is one of the most important RRM functions in the downlink LTE system to provide the intelligent allocation of radio resources for active users. Because of the diversity of the traffic types in wireless systems, active users may have different QoS requirement. In order to satisfy the various QoS requirements and efficiently utilize the radio resource, several packet scheduling algorithms have been proposed. Literature review of the existing packet scheduling algorithms is given in Section 3.2. In order to identify the appropriate candidate for the downlink LTE, it is of great importance to evaluate the performance of these packet scheduling algorithms.

The performance of packet scheduling algorithms is evaluated under the downlink LTE simulation environment in Section 3.3 in terms of the performance metrics given in Section 3.1. Packet scheduling algorithms are tested in three scenarios including 100% RT scenario, 100% NRT scenario and 50% RT and 50% NRT scenario. According to the simulation results for five well-known packet scheduling algorithms, M-LWDF has the best performance in the 100% RT scenario, while EXP/PF is comparatively more appropriate for the 50% RT and 50% NRT scenario. In the 100% NRT scenario, Max-Rate and PF have a good throughput and RB utilization performance while RR has the best fairness performance. In addition, two recently proposed algorithms, namely Sun Qiaoyun's algorithm and Jeongsik Park's algorithm, are evaluated in the 100% RT

scenario. The simulation results show that Sun Qiaoyun's algorithm is more suitable than Jeongsik Park's algorithm for the downlink LTE system with RT traffic.

Besides the simulation results, the theoretical performance analysis results of packet scheduling algorithms in the downlink LTE system are provided in this thesis. The analytical model for the delay in the OFDMA system with VoIP traffic is discussed in Chapter 4. The HARQ is deployed to provide the guaranteed service. The analytical model of the delay consists of the talk spurt level and voice packet level and the mathematical expression for the total packet delay is provided. In Chapter 5, the theoretical throughput analysis results of PF algorithm and M-LWDF algorithm are given in detail. Mathematical expressions of the expected system throughput for both algorithms are provided. Both algorithms achieve a higher system throughput with the increasing system load and with increasing SINR values.

6.2 Future Research Work

The potential future research directions are discussed as follows:

The simulation in this thesis is based on the single cell scenario with wrap-around LTE system. The simulation can be extended to the more realistic multiple-cell scenario. As handover is the essential RRM mechanism and can greatly improve the system performance when there is more than one cell [59], the performance of packet scheduling algorithm should be evaluated by jointly considering the handover mechanism.

All the theoretical performance analysis results in this thesis are based on the statistical analysis. The performance of packet scheduling algorithms is evaluated as either probability density function or the expected value. However, there are some other more realistic analytical models that can be used in the performance analysis, e.g. queueing theory, Markov chain, etc. By applying these analytical models [60-64], the theoretical performance bounds can be another way to evaluate the performance of packet scheduling algorithms in the downlink LTE system.

ABBREVIATIONS

3G	3rd Generation Wireless Network
3GPP	3rd Generation Partnership Project
ACK	Acknowledgement
BS	Base Station
CDMA	Code Division Multiple Access
CP	Cyclic Prefix
CQI	Channel Quality Indicator
CSI	Channel State Information
EPC	Evolved Packet Core
E-UTRAN	Evolved UTRAN
EXP/PF	Exponential/Proportional Fair
EVRC	Enhanced Variable Rate Coder
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FIFO	First-in-First-out
GSM	Global System for Mobile communications
HARQ	Hybrid-Automatic Repeat Request
HOL	Head of Line
HSDPA	High-Speed Downlink Packet Access
IMT-2000	International Mobile Telecommunications-2000
IP	Internet Protocol
LCR	Level Crossing Rate
LTE	Long Term Evolution
Max-Rate	Maximum-Rate
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
M-LWDF	Maximum-Largest Weighted Delay First

MME	Mobile Management Entity
NACK	Negative Acknowledgement
NRT	Non-Real Time
OFDM	Orthogonal Frequency Division Multiplex
OFDMA	Orthogonal Frequency Division Multiple Access
PDF	Probability Density Function
PDN	Packet Data Network
PF	Proportional Fair
P-GW	PDN Gateway
PLR	Packet Loss Ratio
PS	Packet Scheduling
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
QSI	Queue State Information
RB	Resource Block
RMS	Root Mean Square
RNC	Radio Network Controller
RNP	Radio Network Planning
RR	Round Robin
RRM	Radio Resource Management
RT	Real Time
RTT	Round-Trip Time
SAW	Stop-and-Wait
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDF	Sub-band Discrimination Factor
S-GW	Serving Gateway
SINR	Signal to Interference-plus-Noise Ratio
SISO	Single-Input-Single-Output
SNR	Signal-to-Noise-Ratio
TDD	Time Division Duplex
TDMA	Time Division Multiple Access

TFT	Time For Transmission
TTI	Transmit Time Interval
UDF	User Discrimination Factor
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Terrestrial Radio Access Network
VoIP	Voice over IP
WCDMA	Wideband Code Division Multiple Access

SYMBOLS

E	:	A constant
λ	:	Average talk spurt arrival rate
δ_i	:	The maximum probability for HOL packet delay of user i to exceed the delay threshold of user i
τ	:	Alignment delay for the first packet in the talk spurt
τ_i	:	The delay threshold of user i
τ_{max}	:	The maximum delay constraint out of RT service users
$\theta_{i,n}$	:	Discrete Doppler phase
$\xi(t)$	:	Rayleigh process
$\xi_i(t)$	:	The shadow fading of user i at time t
$\eta(t)$	:	Suzuki process
$\zeta(t)$	:	Log-normal process
$\delta(t)$	:	Continuous Dirac delta function
$\gamma_{i,j}(t)$	:	The SNR level of user i on sub-carrier j at time t
$\mu(t)$	:	Complex Gaussian random process
$\mu_i(t)$	:	Priority metric of user i at time t in (3-36) White Gaussian noise process in (2-6)
μ_1	:	Average service rate of the signalling channel
μ_2	:	Average service rate of the traffic channel
$\tilde{\mu}_i(t)$	:	Approximated Gaussian process
ρ	:	Normalized threshold level
$\rho_i(t)$	:	The shadow fading autocorrelation function of user i at time t
$\gamma_i(t)$	:	The number of urgent packets for user i at time t
a	:	A slope of a NRT service
$Bcurr_i(t)$	:	The buffer size of user i at time slot t
$Bcurr\_avg_i$	:	The average value of $Bcurr_i(t)$

c	:	The speed of light
$c_{i,n}$	:	Doppler coefficient
$C(t)$	:	The number of HARQ cycles between the i-1th packet's arrival with the start of ith packet's transmission
$C_{avg}(t)$	:	The average value of $C_{ij}(t)$
$C_{max}(t)$	:	The maximum value of $C_{ij}(t)$
$C_{ij}(t)$	:	The MCS level of user i for sub-band j at time slot t
$\tilde{d}$	:	Total packet delay
$data\_rate(t)$	:	The achievable data rate of a user time t
f_c	:	The carrier frequency
$f_{i,n}$	:	Discrete Doppler frequency
f_{max}	:	The maximum Doppler frequency
$f_{w\|r}(t)$	:	The conditional probability density function of HOL packet delay of user i, given the instantaneous achievable rate of RB j at time $n+1$
$Gain_{i,j}(t)$	:	The channel gain of user i on RB j at time t
$H(t)$	:	Channel matrix
h_n	:	HARQ early termination probability
H_r	:	Height of mobile or receiver in meters
h_t	:	Height of base station or transmitter in meters
I	:	Inter-cell Interference
$I_i(t)$	:	Indicator function of the event that user i is scheduled to transmit at time t
$I(t)$	:	Index matrix
k	:	A constant
$k(t)$	:	The index number of the user who is selected for transmission at time t
K	:	The total number of users
$mpath_{i,j}(t)$	:	The multi-path gain of user i on sub-carrier j at time t
m_1	:	Number of available tile-interlace resources for signalling transmission within one interlace period
m_2	:	Number of available tile-interlace resources for traffic

		transmission within one interlace period
$M(t)$	:	The average number of RT packets waiting at e-Node B buffer at time t
N	:	The total number of available RBs
N_i	:	The number of sinusoids of the ith Gaussian process
N_s	:	The period of HARQ retransmission in slots
N_0	:	The noise power spectral density
$nbits_{i,j}(t)/symbol$	:	The number of bits per symbol of user i at time t on a sub-carrier within RB j
$nsymbols/slot$	:	The number of symbols per slot
$nslot/TTI$	:	The number of slots per TTI
nsc/RB	:	The number of sub-carriers per RB
$p_j(n)$	:	The probability that n talk spurt are present in the jth queue
$PLR_i(t)$	:	Packet loss ratio of user i at time t
$PLR_{req,i}$	:	PLR threshold of user i
P_{total}	:	Total eNodeB downlink transit power
$pdiscard_i(t)$	:	The size of discarded packets of user i at time t
$pdrop_i(t)$	:	The size of dropped packets of user i at time t
$pl_i(t)$	:	The path loss of user i at time t
$psize_i(t)$	:	The size of all packets that have arrived into eNodeB buffer of user i at time t
$ptotaltransmit_{max}$	:	The total size of the transmitted packets of the most served user
$ptotaltransmit_{min}$	:	The total size of the transmitted packets of the least served user
$ptransmit_i(t)$	:	The size of transmitted packets of user i at time t
r	:	Instantaneous achievable data rate
$r_i(t)$	:	The achievable data rate of user i at time t
$R_i(t)$	:	The average data rate of user i at time t
R_{thr}	:	The threshold level
$SDF_i(t)$	:	SDF of user i at time t
$shd_i(t)$	:	The shadow fading gain of user i at time t
s_j	:	The multi-server service time seen by a talk spurt waiting for resource assignment

$\tilde{s}_j$	:	The service time experienced by a talk spurt with assigned resources
T	:	The total simulation time
T_s	:	Time slot duration
T_0	:	Regular Interval of the vocoder
t_c	:	The size of an update window
$total_{RB\_used}(t)$	:	Total number of RBs that have been used for transmission at time t
$UDF_j(t)$	:	UDF of sub-band j at time t
v	:	The user's velocity
$\tilde{v}$	:	Transmission time experienced by the transmitted packet itself
v_i	:	Transmission time of ith packet in the talk spurt seen by other packets waiting in the queue
w	:	Overall packet waiting time
w_i	:	Waiting time of ith packet in the talk spurt
w_b	:	Total queue delay experienced by a new talk spurt
w_{b1}	:	Waiting time in the signalling server
w_{b2}	:	Waiting time in the traffic server
W_{max}	:	The maximum HOL packet delay of all RT service users
$W_{max,i}$	:	Maximum allowable delay of user i
$W_i(t)$	:	The HOL packet delay for user i at time t

REFERENCES

[1] 3GPP, "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN) (Release 7)," 3GPP TR25.913, March 2006.

[2] H. Fattah and H. Alnuweiri, "A cross-layer design for dynamic resource block allocation in 3G Long Term Evolution system," in *Mobile Adhoc and Sensor Systems, 2009. MASS '09. IEEE 6th International Conference on*, 2009, pp. 929-934.

[3] H. Holma and A. Toskala, *LTE for UMTS - OFDMA and SC-FDMA Based Radio Access*, First ed.: John Wiley & Sons Ltd, 2009.

[4] Ericsson, "Ericsson LTE and SAE - The next step in mobile broadband," 2009.

[5] A. Krishnarajah and K. Sandrasegaran, "Lecture Notes: Mobile Communication Systems - Long Term Evolution," University of Technology, Sydney, 2008.

[6] Agilent, "3GPP Long Term Evolution: System Overview, Product Development, and Test Challenges," 2008.

[7] A. Krishnarajah and K. Sandrasegaran, "Lecture Notes: Mobile Communication Systems - HSPA and LTE," University of Technology, Sydney, 2010.

[8] H. Holma and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE*, Fourth ed.: John Wiley & Sons Ltd., 2007.

[9] H. A. M. Ramli, "Packet Scheduling Algorithm for the Future Wireless Systems," University of Technology Sydney, Sydney2009.

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

[11] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution*: John Wiley & Sons, Ltd, 2009.

[12] P. Lescuyer and T. Lucidarme, *Evolved Packet System (EPS)*: John Wiley & Sons, Ltd, 2008.

[13] J. Zyren, "Overview of 3GPP Long Term Evolution Physical Layer," W. McCoy, Dr, Ed.: Freescale Semiconductor, 2007.

[14] E. Dahlman, "3G Long Term Evolution," Ericsson Research2007.