QoS Aware Mixed Traffic Packet Scheduling in OFDMA-based LTE-Advanced Networks

Rehana Kausar, Yue Chen, Kok Keong Chai, Laurie Cuthbert and John Schormans
School of Electronic Engineering and Computer Science
Queen Mary University of London
London E1 4NS, UK
rehana.kausar,yue.chen,michael.chai,laurie.cuthbert,john.schormans@elec.qmul.ac.uk

***Abstract*— In this paper, a packet scheduling framework is proposed for LTE-Advanced downlink transmission. The proposed framework adds the new functionality of an adaptive TD scheduler with built-in congestion control to the existing conventional quality of service (QoS) aware packet scheduling algorithms. It optimizes multiuser diversity in both the time and frequency domains by jointly considering the channel condition, queue status and the QoS feedback. The framework aims to improve the system spectral efficiency by optimizing the use of available resources while maintaining QoS requirements of different service classes and a certain degree of fairness among users. The results show an improved QoS of Real Time traffic and a fair share of radio resources to Non Real Time traffic types.**

Keywords- Packet scheduling; OFDMA; QoS; LTE-A.

I. INTRODUCTION

Long Term Evolution Advanced (LTE-A) is an all-IP based future wireless communication network that is aiming to support a wide variety of applications and services with different quality of service (QoS) requirements. It is targeting superior performance in terms of spectral efficiency, system throughput, QoS and service satisfaction when compared with existing 3GPP wireless networks [1].

As one of the core functionalities in radio resource management, packet scheduling (PS) plays an important role in optimizing the network performance and it has been under extensive research in recent years. Different PS algorithms have been deployed aiming at utilizing the scare radio resource efficiently. The classic PS algorithms exploiting multiuser diversity are the MAX C/I and Proportional Fairness (PF) algorithms. MAX C/I algorithm allocates a physical resource block (PRB) to a user with the highest channel gain on that PRB, and can maximize the system throughput [2]. The PF algorithm takes fairness among users into consideration and allocates resources to users based on the ratio of their instantaneous throughput and its acquired time averaged throughput [3]. However these algorithms aim only at improving resource utilization based on channel conditions of users; QoS requirements "e.g." delay requirements of real time (RT) service or minimum throughput requirements of non-real time (NRT) service are not considered at all. In the next generation networks, apart from system throughput and user fairness, the crucial point is to fulfill users' QoS requirements in a multi-service mixed traffic environment. This is because different service types are competing for radio resources to fulfill their QoS requirements. To allocate radio resources efficiently and intelligently in such complex environments is challenging. Various methods have been proposed aiming to use radio resources efficiently to fulfill QoS requirements of different traffic types [4][16][17].

In [4], a service differentiation scheme is used which classify mixed traffic into different service classes and grants different scheduling priorities to them. Two types, VoIP and BE are considered and the results show an improvement in RT QoS at the cost of system spectral efficiency, when the RT queue is granted the highest priority. In [5], an urgency factor is used to boost the priority of a particular service. When any packet from a service flow is about to exceed its upper bound of QoS requirement, its priority is increased by adding an urgency factor. Although most of the packets are sent when they are nearly ready to expire, a lower packet loss is achieved thus improving the performance of system by guaranteeing QoS requirements to different services. In mixed traffic scenarios, queue state information (QSI) becomes very important in addition to channel state information (CSI) [6] [15]. A time domain multiplexing (TDM) system based Modified Largest Waited Delay First (M-LWDF) is presented in [6] which takes into account both QSI and CSI. This algorithm serves a user with the maximum product of Head of Line (HOL) packet delay, channel condition and an arbitrary positive constant. This constant is used to control the packet delay distribution for different users. This algorithm is applied in a frequency domain multiplexing (FDM) system in [7] to optimize sub-carrier allocation in OFDMA based networks. It shows improved performance in terms of QoS but like M-LWDF updates the queues state each TTI rather than after each sub-carrier allocation. In [8], M-LWDF is modified by updating the queue status after every sub-carrier allocation. It takes into account RT and NRT traffic types and provides better QoS for both services. The results show an improvement in delay for RT and throughput for NRT service. However this idea can be extended to more intelligent scheduling framework by adding more traffic types and making resource allocation more adaptive based on QoS.

 ISBN: 978-1-61208-100-7

In a multi-service environment, the crucial point is to clearly define the QoS requirements of different services, their demands for radio resources and their channel conditions and queue status to support their demands. Combined consideration of this information can lead to a more efficient PS algorithm, which can be further optimized for network level congestion control by giving QoS feedback.

The work in this paper addresses the scheduling problem in a multi-service wireless environment where the competition to get radio resources is keen and there are strict QoS requirements. A novel PS framework is proposed with added functionalities, to achieve better QoS of different traffic types, a fair share of throughput among users and improved spectral efficiency. The proposed PS framework segregates different types of traffic and sorts users in the service specific queues based on different queue sorting algorithms. A built-in congestion control Adaptive Scheduler is introduced in the TD which makes the system more adaptive to meet QoS guarantees of RT traffic and prevent NRT traffic from starvation. Multiuser diversity in both time and frequency domain are exploited by frequent updating of queue state information and channel condition which leads to a balance prioritizing among users of different traffic types.

The rest of the paper is organized as follows. Section 2 gives a detailed description of the proposed PS framework. Section 3 presents system model and its performance metrics. The simulation model and results are described in Section 4, and finally, conclusions are presented in Section 5.

II. PROPOSED SCHEDULING FRAMEWORK

A schematic diagram of proposed scheduling framework is shown in Figure 1.

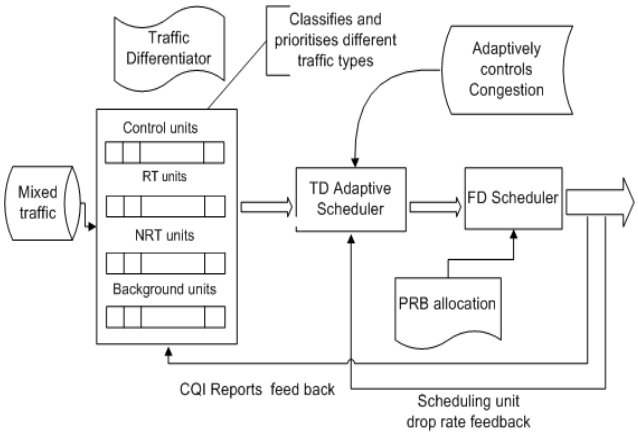

Figure 1. Proposed scheduling framework

The framework is composed of three main units: a traffic differentiator and prioritizing unit, a TD adaptive scheduler with built-in congestion control and a frequency domain (FD) scheduler where resources are mapped to users according to priority order selected in TD. Compared with other PS algorithms, the novelty of the proposed framework lies mainly in the TD adaptive scheduler. However in the traffic differentiator and prioritizing unit, delay-dependent queue-sorting algorithms make a difference compared with the schemes used in reference paper.

The detailed description of the functionality of each unit, the algorithms and policies used in each unit is presented below.

A. Traffic differentiator and prioritizing unit

The need for a differentiator arises when there are different traffic types demanding radio resources with different QoS requirements. In such an environment it becomes very important to classify traffic in service queues to enable queue specific prioritizing schemes to be applied flexibly. Service classification is in fact the first step towards optimizing utilization of available radio resources while dealing with mixed traffic. This is because with complete knowledge of QoS requirements of each class, just enough radio resources can be allocated to these classes. The QoS guarantees become more feasible when radio resources are allocated according to the well-defined demands of traffic types rather than by estimation.

In the proposed scheduling architecture mixed traffic is classified in four queues; Control (control information), RT conversational traffic (voice), NRT streaming (video file download) and background (email, SMS). These queues are chosen for the present study because they cover most of the common data types including low latency, high throughput and low priority. The Background traffic represents the best effort (BE) class of traffic and does not have any QoS requirements. The control traffic is the most important traffic type so it is put into a dedicated queue and served before other traffic types. In the present work control information for downlink (DL) scheduling is considered only as this study is for downlink transmission of LTE-A networks.

In the proposed PS framework, one user is assumed to have one service type and one scheduling unit (SU) carries the information about user, service type and buffer status. The queues in the differentiator are prioritized from top to bottom that is Control, RT, NRT and Background respectively. After differentiation, SUs are sorted within the queues using different queue sorting algorithms. The Control queue SUs are sorted by Round Robin (RR) algorithm because all control information has to be equally important, meanwhile RT, NRT and Background queue SUs are sorted by using queue specific priority metrics.

RT Traffic

The QoS requirement for RT traffic is defined as $d_k < DB_{RT}$ where d_k is delay of user k, DB^{RT} is the delay budget for RT traffic. The delay budget for RT traffic is 40ms [8] [17] in OFDMA-based networks. If this condition is not met then the SUs will be dropped from the queue. A

delay dependent queue sorting algorithm is used for RT users and the priority metric is formed by the product of normalized Head of Line (HOL) delay and the complex channel gain of the users. The Normalized HOL delay is a ratio of user's waiting time and the delay budget for RT traffic. The waiting time of a user is equal to number of transmission time intervals (TTIs) during which the user has not been allocated. The priority of user *k* at time *t*, $p_k(t)$ is

$$p_k(t) = F_k^{RT}(t) \times H_k^{RT}(t) \quad (1)$$

where $H_{k\in K}^{RT}$ is the channel gain of user *k* and $F_{k\in K}^{RT}$ is normalized waiting time of user *k* at time *t* given by.

$$F_{k\in K}^{RT}(t) = \frac{T_{Waiting}}{DB^{RT}} \quad (2)$$

where $T_{waiting}$ is the waiting time, DB^{RT} is upper bound of delay for RT traffic.

In each TTI, the user with the highest priority value is sorted at the front of the queue followed by users with priority value in descending order.

NRT Traffic

The priority metric for NRT streaming video traffic is the product of normalized HOL delay of each user and the ratio of its instantaneous throughput and the average throughput over a given time interval. In this queue the throughput ratio is used instead of channel gain to provide a balance between throughput and fairness. SUs are arranged according to the highest value of this priority metric thus not only satisfying their QoS requirements but also exploiting multiuser diversity in TD. The priority of user *k* at time *t*, $p_k(t)$ is

$$P_k(t) = F_{k\in K}^{NRT}(t) \times \frac{r_k}{R_k} \quad (3)$$

where $F_{k\in K}^{RT}$ is normalized waiting time of user *k* at time *t* and is given by

$$F_{k\in K}^{NRT}(t) = \frac{T_{Waiting}}{DB^{NRT}} \quad (4)$$

where $T_{waiting}$ is the waiting time, DB^{NRT} is upper bound of delay for NRT streaming video traffic, r_k is instantaneous throughput and R_k is average throughput of user k.

The time average throughput of user k is updated by the moving average as below as used in [9] and many other papers,

$$R_k(t+1) = \left(1 - \frac{1}{t_c}\right) R_k(t) + \frac{1}{t_c} \sum_{m=1}^{M} r_{k,m}^{/}(t) \quad (5)$$

where t_c is the length of time window to calculate the average data rate, $1/t_c$ is called the attenuation co-efficient with the widely used value 0.001, $r_{k,m}^{/}(t)$ is the acquired data rate of user k at PRB m if m is allocated to k else it is zero.

Background Traffic

Background traffic has no QoS requirements so priority is given to BE users based only on channel conditions. However to maintain some fairness between users, the proportional fairness (PF) algorithm is used as the queue sorting algorithm for Background queue. The priority of user *k* at time *t*, $p_k(t)$ is

$$P_k(t) = \frac{r_k}{R_k} \quad (6)$$

where r_k is instantaneous throughput, R_k is average throughput of user k as defined previously.

After prioritizing users in the queues the TD adaptive scheduler picks specific proportion of users from the queues.

B. Time Domain adaptive scheduler

This unit aims at guaranteeing the QoS of RT traffic and at the same time ensuring fairness for NRT traffic. It allocates just enough resources to meet the QoS requirements of RT and remaining resources are allocated to NRT services based on the requirements of service types. This scheduler unit enhances the adaptability of the whole framework by collecting the QoS feedback, such as SU drop rate, as its input to make decisions on the TD adaptive scheduling policy selection. The system is said to be in congestion when the QoS of the RT service is not met and due to system load RT SUs are dropping frequently. The TD adaptive scheduling unit is integrated with a built-in policy based congestion control that controls congestion of the system in the network.

The TD adaptive scheduling algorithm works as follows.

Let the total number of available PRBs be denoted by C. If λ denotes the proportion of PRB assigned to RT users and $(1-\lambda)C$ is assigned to NRT users then λ can be adaptively adjusted according to the practical user distribution or QoS of RT traffic. The proportion of capacity given to the NRT traffic for this paper is further divided in different types of the NRT traffic (control, NRT streaming video and Background) such that first the control queue is allocated enough PRBs to deliver control information of all users and then rest of the PRBs are allocated to the NRT and the Background queue. In this way control queue is at the top and is always allocated enough PRBs.

 ISBN: 978-1-61208-100-7

In this paper, three built-in congestion control policies are chosen to exemplify the adaptive capability of TD adaptive scheduler in which the value of λ is changed according to network conditions. The value of λ is changed based on a threshold χ which is set using the drop rate of SUs of RT traffic. When the number of SUs dropped exceeds the threshold χ, the built-in congestion control policy changes accordingly to reduce SUs drop rate. The distribution of the NRT capacity is adjusted according to the buffer status and requirements of NRT service types such as streaming traffic is more important and more frequently requested service than Background traffic and control information is always less than actual data to be sent.

In this paper, the PS algorithm in [4] with fair TD scheduling is considered as reference algorithm. The TD scheduler in [4] uses conventional channel dependent queue sorting algorithms and gives priority to different queues from top to bottom based on fair scheduling or by strict priority. In fair scheduling one user is picked from each queue at a time, starting from top queue and in strict priority queues are emptied completely one by one. In FD, resources are mapped in priority order to the users selected in TD.

C. *Frquency Domain scheduler*

Resources are actually mapped to SUs in the FD scheduler according to the priority selected in TD. Multiuser diversity is exploited by using channel dependent proportional fair (PF) algorithm in FD. For each SU, the best PRB (with highest throughput) is selected out of available PRBs and is allocated to this SU.

III. SYSTEM MODEL AND PERFORMANCE METRICS

In this work, an OFDMA system with minimum allocation unit as 1 PRB containing 12 sub-carriers in each TTI is considered. The DL channel is a fading channel within each scheduling drop. The received symbol $\mathrm{X}_{k,m}(t)$ at the mobile user k on sub channel m is the sum of White Gaussian Noise and the product of actual data and channel gain as shown below,

$$\mathrm{X}_{k,m}(t) = \mathrm{H}_{k,m}(t)\mathrm{I}_{k,m} + Z_{k,m}(t) \qquad (7)$$

where, $\mathrm{H}_{k,m}(t)$ is the complex channel gain of sub channel m for user k, $\mathrm{I}_{k,m}(t)$ is data symbol from eNB to user k at sub channel m and $Z_{k,m}(t)$ is complex White Gaussian Noise [8]. It is assumed as in [4], [5], [8] and [14] that the power allocation is same, $P_m(t) = P/M$ on all sub channels. Where, P is the total transmit power, $P_m(t)$ is the power allocated at sub channel m and M is total number of sub channels. At the start of each scheduling drop, the channel state information $\mathrm{H}_{k,m}(t)$ is known by the eNodeB.

The channel capacity of user k on sub channel m can be calculated by using Eq. (8) as used in [5][8],

$$C_{k,m}(t) = \mathrm{B}\log_2\left(1 + \frac{|H_{k,m}(t)|^2}{\sigma^2\Gamma} P_m(t)\right) \qquad (8)$$

where B is the bandwidth of each PRB, σ^2 is the noise power density and $\Gamma = -\ln(5\mathrm{BE}R)/1.5$ is the SNR gap determined by bit error rate BER.

In the proposed framework, users are served by one of the differentiated queues depending on their QoS requirements. For example RT users must not exceed their delay bounds, NRT users must achieve their minimum data rate and there should be fairness among Background users. At a given time *t*, PRBs are allocated to users by the following algorithm.

Step 1: Initialize queues for all traffic types and the number of PRBs.
Step 2: Sort users in these queues according to queue sorting algorithms given in equations 1, 3 and 6 for different traffic types.
Step 3: Select a number of users from these queues according to built-in policies in TD adaptive scheduler.
Step 4: Allocate PRB to the user with the highest priority.
Step 5: Remove the allocated PRB from the PRB list and the allocated user from the user list.
Step 6: Go to step 4 if the PRB list is not empty else go to next TTI.

Resource allocation is completed when all PRBs are allocated. The proposed PS framework is analyzed under performance metrics of system throughput, user fairness and QoS of different traffic types.

The system average throughput is the sum of average throughput across all users. To measure the fairness among users, Raj Jain fairness index is adopted that is given as below as used in [10][11],

$$Fairness = \frac{\left[\sum_{k=1}^{K} \tilde{R}_k\right]^2}{K\sum_{k=1}^{K}\left(\tilde{R}_k\right)^2} \qquad (9)$$

The value of fairness index is 1 for the highest fairness when all users have same throughput. In Equation (9), K is the total number of users and $\tilde{R}_k$ is the time average throughput of user k.

The value of SU drop rate and the average delay of RT traffic are used to evaluate QoS of RT traffic. SU drop rate is calculated by the ratio of number of RT SUs dropped to total number of RT SUs. In addition, the average delay for

all NRT traffic is also calculated to prevent NRT traffic from starvation.

IV. SIMULATION RESULTS AND DISCUSSION

Simulation model, results and analysis will be presented in this section.

A. Simulation model

A single cell OFDMA system with total system bandwidth of 10 MHz and PRB size of 180 kHz has been considered. Total system bandwidth is divided into 55 PRBs. The simulation parameters used for system level simulation are based on [12] and these are typical values used in many papers. The wireless environment is typical Urban Non Line of Sight (NLOS) and the LTE system works with a carrier frequency of 2GHz. The most suitable path loss model in this case is the COST 231Walfisch-Ikegami (WI) [13] as used in many other papers on LTE.

Users are assumed to have a uniform distribution and the total number of RT users is assumed to be equal to total number of NRT users as in [8]. Each TTI is 1 ms and the delay upper bound for RT traffic is taken 40 ms which is equivalent to 40 time slots. Total eNB transmission power is 46dBm (40w) and BER is 10^{-4} for all users.

B. Simulation results

The performance of the proposed framework is evaluated by comparing it with the stand alone PF and QoS aware PS algorithm in [4] referred to as the reference algorithm hereafter. All simulations are done in Mat lab.
Figure 2 shows the average delay of RT users with different adaptive TD scheduler policies for 80 active users.

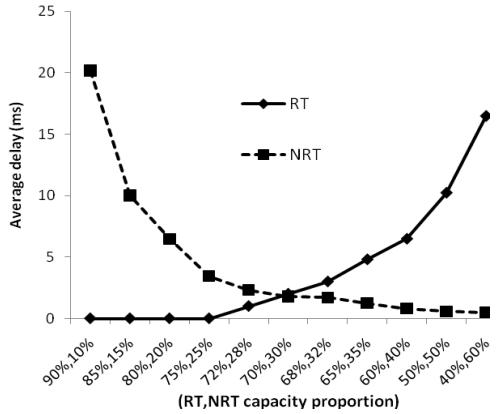

Figure 2. Comparison of TD adaptive scheduler policies

The average delay for RT traffic decreases as RT capacity proportion is increased and increases as RT capacity proportion is decreased. This change in average delay of RT traffic is shown by solid line in Figure 2. On the other hand, by increasing RT capacity, the average delay of NRT traffic does get very high. The average delay of RT and NRT traffic is analyzed under a number of TD adaptive scheduling policies to find a good trade-off so that RT traffic may not exceed its delay upper bound and at the same time the QoS of NRT may be satisfied. For this particular user distribution, the policy (70%, 30%) shows a balance point where both RT and NRT can get reasonable capacity proportion and it is adopted as the default policy in the next results. The proposed algorithm will start with (70%, 30%) policy and will be able to switch to other policies depending on network conditions.

Figure 3 shows the average delay of RT traffic under different system load when reference and the proposed algorithms are used. The standalone PF algorithm has no functionality for QoS of RT traffic that is why it is not included in this analysis.

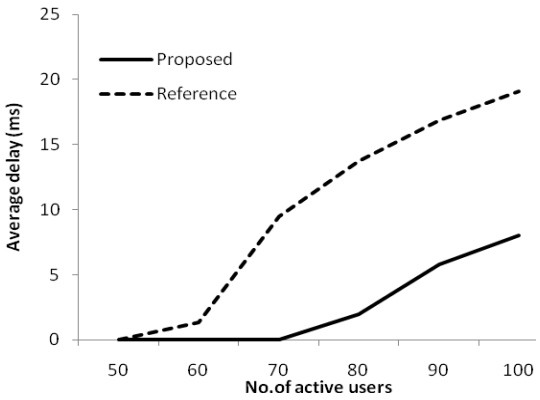

Figure 3. Delay under different system loads

The RT delay increases with system load for both reference and the proposed algorithm. However delay with proposed algorithm remains lower than the reference algorithm as shown. This is because the adaptive TD scheduler in the proposed algorithm adaptively controls the delay of RT traffic. In Figure 4 SUs drop rate for RT traffic under different system load is shown for the proposed and the reference algorithms.

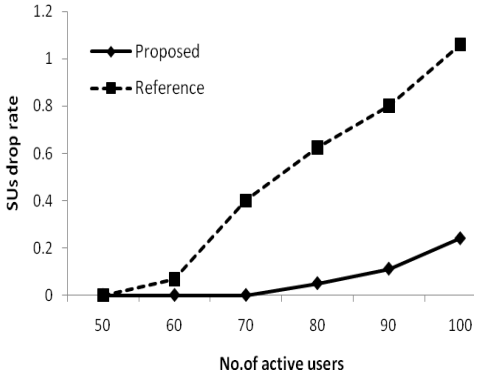

Figure 4. Sceduling unit drop rate Vs system load

There is no SU drop up to a load of 70 active users with proposed algorithm; however after that SU drop rate increases at a tolerable rate. The SUs drop rate for reference algorithm is zero when total number of users is 50 which is lower than the available number of PRBs (55). However with the increase in system load, SUs drop rate for reference algorithm increases significantly as shown.

Figure 5 shows throughput and fairness comparison of reference, proposed and PF algorithm. These simulations are done under the same system load of 110 users.

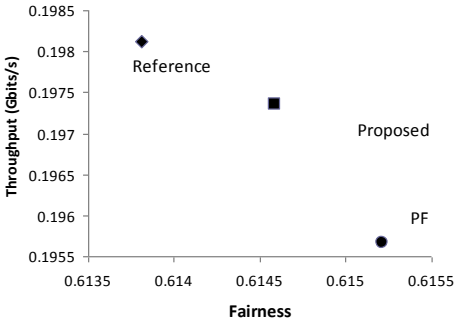

Figure 5. Trade-off between fairness and throughput

The system overall throughput for the proposed algorithm is lower than the reference algorithm by only 0.4%. This is because in the proposed algorithm, a delay-dependent queue-sorting algorithm is used and users with relatively low channel conditions but more waiting time are scheduled to guarantee QoS of RT traffic. This lowers the system overall throughput by a slight amount compared to the reference algorithm but more than PF algorithm.The fairness of proposed algorithm is improved as compared to the reference algorithm and is slightly less than PF algorithm as shown.In the three algorithms fairness of the PF algorithm is the highest with value 0.615213 as PF being an algorithm designed for user fairness and is taken as a reference for fairness analysis. The fairness index with the proposed algorithm is 0.61452 and with the reference algorithm fairness index is 0.61357.

In this way, the proposed algorithm sacrifices a little throughput (compared with reference algorithm) and fairness (compared with PF algorithm) but presents an better trade-off between throughput and fairness (compared with both reference and PF algorithms) as shown.

V. CONCLUSION

In this paper, we have presented a QoS aware PS framework that is composed of three main units for the resource allocation in DL transmission for OFDMA-based networks. These units use different queue sorting, TD adaptive scheduling and FD scheduling algorithms to guarantee better QoS to different traffic types. It is able to improve system spectral efficiency by optimizing the use of given radio resources and maintains a certain degree of fairness among users at the same time. This is achieved by adaptively providing just enough resources to RT traffic and distributing remaining resources efficiently to NRT services. The results show an improved QoS of RT traffic and a better trade-off between user fairness and system overall throughput.

REFERENCES

[1] Harri H. and Antti T.,"LTE for UMTS OFDMA and SC-FDMA Based Radio Access," John Wiley and sons ltd 2009, pp. 181-190.

[2] M. Sauter.(2008,April 23). Wireless Moves, 3GPP Moves on: LTE-Advaced.Website: http://mobilesociety.typepad.com/mobile_life/2008/04/3gpp-moves-on-html 29 .05.2010.

[3] Stefania S., Issam T., and Matthew B., "The UMTS Long Term Evolution Forum Theory to Practice," 2009 John Willey & Sons Ltd.ISBN: 978-0-470-69716-0.

[4] Jani P., Niko K., tero H., Martti M. and Mika R., "Mixed Traffic Packet Scheduling in UTRAN Long Term Evaluation Downlink," IEEE 2008, pp. 978-982.

[5] Gutierrez, F.Bader, and J.L.Pijoan, "Prioritization function for packet scheduling in OFDMA systems," Wireless internet conference 2008, Nov. 08, Maui, USA. http://dx.doi.org/ICST.WICON2008.5002.

[6] M.Andrews et al., "Providing quality of service over a shared wireless link," Communication magazine, IEEE, vol.39, 2001, pp. 150-154.

[7] P.Parag, S.Bhashyam, and R.Aravind, "A subcarrier allocation algorithm for OFDMA using buffer and channel state information," Vehicular Technology Conference. VTC-2005-Fall.2005 IEEE , pp. 622-625.

[8] Jun S., Na Yi, An Liu and Haige X., "Opportunistic scheduling for eterogeneous services in downlink OFDMA system," School of EECS, Peking University Beijing, P.R.China, IEEE computer Society 2009, pp. 260-264.

[9] G.Song et al, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," Communication Magazine,IEEE, vol.39,2001, pp. 150-154.

[10] B. Chisung, and C. Dong, "Fairness-Aware Adaptive Resource Allocation Scheme in Multihop OFDMA System," Communication letters IEEE, vol.11, pp. 134-136, Feb. 2007.

[11] Lin X., Laurie C. "Improving fairness in relay-based access networks," in ACM MSWIM, Nov. 2008, pp. 18-22.

[12] page 3GPP TSG-RAN, "TR25.814: Physical Layer Aspects for Evolved Utra," Version 7.0.0, June, 2006.

[13] IEEE 802.16j-06/013r3: "Multi-hope Rehay System Evaluation Methodology (Channel Model and Performance Metric)," IEEE 802.16 Broadband Wireless Access Working Group, 2007-02-19.

[14] Jiho J. and Kwang Bok L., "Tranmit power adaptation for multiuser OFDM systems," Selected Areas in Communication, IEEE Journal on vol.21, 2003, pp. 171-178.

[15] Suleiman Y.Yerima and Khalid Al-Begain, "Dynamic buffer management for multimedia QoS beyond 3G wireless networks," IAENG International Journal of computer science, 36:4, IJCS_36_4_14, Nov.2009.

[16] Won-Hyoung P., Sunghyun C., and Saewoong B., "Scheduling design for multiple traffic classes in OFDMA networks," IEEE 2006, pp. 790-795.

[17] T. Janevski, " *Traffic Analysis and Design of Wireless IP Networks"*, Artech House, Norwood, MA, 2003.

 ISBN: 978-1-61208-100-7