Journal of Computer Science

Original Research Paper

Utility-based Scheduling Frameworks for Efficient Quality-of-Service Differentiation in a Mixture of Real-time and Non-real-time Traffics

[1]Olalekan Bello, [1]Hushairi Zen, [1]AL-Khalid Othman and [2]Khairuddin Abdul Hamid

[1]Department of Electrical and Electronics Engineering, University of Malaysia, Sarawak, Malaysia
[2]University Malaysia of Computer Science and Engineering, Putrajaya, Malaysia

Article history
Received: 28-07-2015
Revised: 17-09-2015
Accepted: 23-09-2015

Corresponding Author:
Olalekan Bello
Department of Electrical and Electronics Engineering, University of Malaysia, Sarawak, Malaysia
Email: lekkibel@yahoo.com

Abstract: This paper proposes a utility-based scheduling framework for efficient differentiation of users' Quality-of-Service (QoS) in a broadband wireless access system involving heterogeneous mixed traffic flows. The utility-based scheduling framework, called Maximum QoS Satisfaction (MQS), is based on three novel Radio Resource Allocation (RRA) techniques; delay-based scheduling policy for Real-Time (RT), minimum-rate-based scheduling policy for Non-Real-Time (NRT) and a throughput-based scheduling policy for Best-Effort (BE) services. Simulation study shows that MQS achieves superior performances in terms of average system throughput and user satisfaction both in single and heterogeneous mixed traffic scenarios, when compared to some existing ones.

Keywords: BWASs, QoS, Utility, Throughput, User Satisfaction

Introduction

Broadband Wireless Access Systems (BWASs) have experienced incredible development in recent decades. Some of the popular networks in this category include the High-Speed Downlink Packet Access (HSDPA) (Forkel *et al.,* 2005), Worldwide Interoperability for Microwave Access (WiMAX) (IEEE Computer Society and Society, 2005) and Long-Term Evolution (LTE) (Dahlman *et al*., 2007) which are all based on Third Generation (3G) and Fourth Generation (4G) technologies.

Although these 3G and 4G technologies are attractive and efficient, they present some challenging issues; one, the wireless channel is characterized by fast-fading due to user mobility, two it must support a wide range of multimedia applications with diverse Quality of Service (QoS) requirements. To overcome the issue of channel fading, orthogonal frequency division multiplexing (OFDM) and orthogonal frequency division multiple access (OFDMA) have been adopted by network standards as the physical layer technology of choice (Rodrigues and Casadevall, 2009).

The need for supporting various traffics in end-to-end transmission makes it inevitable for networks to guarantee the satisfactory provision of the quality of services in wireless links. In resource allocation problems, it is widely accepted that the higher the data arrival rate of a traffic flow, invariably the higher is its average throughput. But in a mixture of diverse traffics, the amount of sharable resources that the users get depends not only on their data arrival rates but also on their QoS constraints. However, when arrival rates are same or slightly different the traffic with the higher priority QoS requirement must be satisfied more. Therefore, a heterogeneous traffic scheduling algorithm must consider the specific QoS requirement of each application in allocating the common network resources. In this study, we propose a Maximum QoS Satisfaction (MQS) utility-based scheduling to allocate common network resources to a mixture of RT and NRT data traffics simultaneously. To achieve this, we designed MQS comprising of three different novel utility functions: a sigmoidal-type utility function for RT, and a diminishing marginal utility function for NRT and BE traffic flows. Each utility function incorporates an appropriate QoS metric (e.g., delay, throughput...) and QoS requirement (e.g., maximum delay, minimum throughput…) in order to ensure each user's QoS satisfaction. The rest of this paper is organized as follows. In section 2, we present some related works. In section 3 the system model and assumptions are discussed. Section 4 proposes a novel scheduling framework based on sigmoid-type utility functions. In section 5, we present the system performances such as average system throughput and user satisfaction. In section 6, we summarize our work.

Related Works

Wireless standards leave open, design of scheduling algorithms which will take advantage of multiuser and frequency diversities provided by OFDMA to improve overall system efficiency. As a result, the Maximal-Sum-Rate (MSR) is proposed in (Knopp and Humblet, 1995). The MSR adopted by HSDPA aims only at maximizing the system capacity (average system throughput) by scheduling the user with the best channel condition during each scheduling epoch, thus sacrificing fairness in resource allocation.

Therefore, the Proportional-Fairness (PF) scheduling algorithm was proposed in (Kelly, 1997; Kelly *et al*., 1998; Jalali *et al*., 2000) and implemented for 1xEVDO systems to provide a trade-off between throughput and resource allocation fairness. Fairness in this sense means equal allocation of network resources or scheduling opportunities. Even though PF scheduling provides multiuser diversity with fair resource allocation, it is still unable to achieve optimal fairness: it schedules more users in favorable channel conditions than users with poor channel conditions. In addition, none of these two algorithms can support QoS differentiation as required in BWASs.

To provide service differentiation and achieve satisfactory performance for a wireless broadband system, a dynamic bandwidth allocation algorithm which takes into accounts both the channel and queue conditions, and QoS requirement of each user is required in the Base Station (BS) of a wireless network. To this end, several QoS-based scheduling policies such as the Exponential (EXP) (Shakkottai and Stolyar, 2001), Maximum Delay Utility (MDU) (Song, 2005), Modified-Largest Weighted Delay First (M-LWDF) (Andrews *et al*., 2000; 2001) were proposed. However, these algorithms use only packet delay as the QoS metric to schedule both the RT and NRT traffics irrespective of whether such traffics demand delay or not; therefore any QoS provisioning provided by them is only relative.

Wang and Jia (2010), using strict priority implemented all the QoS services for WiMAX standards in the proposed algorithm. The algorithm uses strict priority discipline which allows the higher priority connections to starve the lower priority connection of bandwidth. However, a higher priority does not even ensure any absolute performance but only provides relatively better performance (Lee *et al*., 2009). Lee *et al*. (2009) also noted that traffic prioritization raises the issue of fairness as it already determines the order of access. In (Balakrishnan and Canberk, 2014) a traffic-aware QoS provisioning scheduling algorithm was proposed for constant-bit rate, video streaming and BE. However, the utility functions are based on average waiting time and traffic priority; no application-specific QoS requirements are included. Hence, efficient QoS provisioning will be difficult to achieve. The utility function designed to allocate resources to heterogeneous traffics must consider the QoS parameters appropriate for each traffic flow; utility function can measure the amount of utility (satisfaction) a user derives in terms of resources allocated to him. Therefore, Rodrigues and Casadevall (2009) proposed adaptive Delay-Based Fairness (ADF) and adaptive Throughput-Based Fairness (ATF) to schedule RT and NRT traffics, respectively. Although, each of the utility functions contains the appropriate QoS metric, they did not specify any QoS requirement hence QoS provisioning is not possible.

The scheduling algorithm in (Song and Li, 2005) uses two utility functions, one for delay-constrained traffic and the other for BE traffic. However, average waiting time (average packet delay) is used to schedule the two different services. The scheduler in (Wang *et al*., 2007) considers support for three major classes of service, i.e., BE traffic with no QoS requirement, NRT traffic with data rate requirement and real-time traffic with delay requirements. However, the algorithms do not aim at scheduling the different traffics at the same time.

Al-Manthari *et al*. (2009) proposed an algorithm that incorporates three utility functions, one for delay-based traffic, one for minimum rate-based traffic and the other for maximum-rate-based traffic. The algorithm specifies three different types of constant parameters for each type of traffic for the purpose of differentiating their QoS requirements. The main problem with this approach is the complexity of the algorithm and the difficulty of choosing an optimal value for each set of the parameters. The Urgency and Efficiency Based Packet Scheduling (UEPS) algorithm in (Ryu *et al*., 2005) is proposed to schedule RT and NRT traffic flows. Although, it uses different utility functions to schedule RT and NRT traffics, head-of-line (HOL) packet delay is still used to allocate resources to both RT and NRT traffics, thereby making efficient provisioning of their QoSs difficult to achieve.

Supporting QoS implies an adaptation to various applications (Lee *et al*., 2009). Therefore, for efficient resource allocation and guarantee for QoS satisfaction, it is important that common network resources are allocated to different traffics using only the specific requirements of their applications. To satisfy this important requirement, Lima *et al*. (2014) proposed two utility-based RRA policies, the Throughput-based Satisfaction Maximization (TSM) policy and the Delay-based Satisfaction Maximization (DSM) policy based on sigmoid utility function and both aimed at maximizing the number of satisfied users in the system. However, it uses a similar

(bell-shaped) utility curve in both the TSM and DSM scheduling policies, thus making a guarantee of a higher user satisfaction index to a higher priority traffic flow highly improbable, especially when they have similar traffic model; because using the same utility curve for scheduling different traffic types may potentially achieve the same result as applying same QoS parameter to schedule different classes of service.

System Model and Assumptions

We consider the downlink of a time-slotted OFDMA-based wireless packet access network. We assume a total bandwidth of B which is divided into K independent subcarriers and shared by M users who are randomly located at various distances and angles from their serving base station (BS) within a cell; therefore each user experiences a path loss. The base station (BS) can transmit a total power of P which is uniformly allocated (uniform power allocation) among the total number subcarriers K, and it is assumed to be equipped with single transmit antenna to provide service to M active users, each equipped, without loss of generality, with a single receive antenna. Transmission between the BS and active users or mobile stations (MSs) takes place in time slots of a fixed duration, T_s , which is assumed to be less than the channel coherence time τ_c. Thus, the channel gain, H_m, is constant during each time slot and is independent of the channel for other time slots, i.e., a quasi-static fading channel is assumed. In the BS, the incoming packets of each user arrive from some upper layers and then buffered in its first-come-first-out (FIFO) queue with some given space for F packets waiting to be scheduled. We assume that each user (or subscriber) only has one traffic flow which can be chosen from video streaming and File Transfer Protocol (FTP) and each traffic flow m is assigned a queue Q_m.

Proposed Scheduling Frameworks

The idea behind the proposed scheduling framework is to develop separate utility functions for each traffic flow so as to be able to incorporate the QoS metric that is specific to each application. Therefore, for a delay-based traffic, we formulate a sigmoidal-type (positive and increasing) utility function in terms of packet delay, D_m^{hol} and which can be expressed by:

$$U\left(D_m^{hol}\right)=\frac{exp\left\{\rho\left(D_m^{hol}-D_m^{\max}\right)\right\}}{exp\left\{\rho\left(D_m^{hol}-D_m^{\max}\right)\right\}+exp\left\{-\rho\left(D_m^{hol}-D_m^{\max}\right)\right\}} \quad (1)$$

where, $U\left(D_m^{hol}\right)$ is a step-shaped utility function, $D_m^{\max}$ is the maximum delay requirement of user m and the parameter, $\rho=\frac{k}{D_m^{\max}}$ is the normalizing parameter and $k \subset$ (1, 2, 3) is the constant that determines the shape of the utility curve. The RT users' utility derived from the network increases as the HOL packet delay, D_m^{hol}, increases; that is, the user's chances of being allocated resources increases as his HOL packet delay increases with respect to his maximum delay requirement, $D_m^{\max}$. Therefore, we have that the RT utility function is an increasing utility function. The recursive HOL packet delay is approximately computed by:

$$D_m^{hol}\left(t+1\right)=D_m^{hol}\left(t\right)+\frac{\left(\varpi_m T_s-R_m[t]T_s\right)}{\varpi_m} \quad (2)$$

where $R_m[t]$ is the achievable data rate for user m at time slot t, T_s is time slot duration and ϖ_m is the bit arrival rate for user m. The term $\frac{\left(\varpi_m T_s-R_m[t]T_s\right)}{\varpi_m}$ represents the instantaneous HOL packet delay. When $R_m[t]$ is zero, the HOL packet delay is incremented by T_s. When the arrival rate for user m equals his achievable data rate, i.e., $\varpi_m = R_m[t]$ the instantaneous packet delay is zero. However, in a heterogeneous mixed traffic involving NRT and RT services, the average throughput for NRT must be allowed to gradually decrease after it has achieved its required minimum rate so that the RT traffics will be able to satisfy their delay requirements. Therefore, a positive and decreasing utility function obeying the law of diminishing marginal utility will better capture this objective for NRT traffic, and can be mathematically modeled by:

$$U\left(\overline{R}_m\right)=\frac{exp\left\{-\rho\left(\overline{R}_m-R_m^{\min}\right)\right\}}{exp\left\{\rho\left(\overline{R}_m-R_m^{\min}\right)\right\}+exp\left\{-\rho\left(\overline{R}_m-R_m^{\min}\right)\right\}} \quad (3)$$

Similarly, the parameter $\rho=\frac{k}{R_m^{\min}}$ is the normalizing parameter, $R_m^{\min}$ is the minimum throughput requirement and $\overline{R}_m$ is the data rate of user m averaged over all time slots. BE traffic is generally considered an NRT service, but which requires no minimum rate guarantee. Therefore, the minimum rate requirement in Equation 3 can be substituted with zero to produce a utility function for BE traffic represented by:

$$U\left(\overline{R}_m\right)=\frac{exp\left\{-\rho\left(\overline{R}_m\right)\right\}}{exp\left\{\rho\left(\overline{R}_m\right)\right\}+exp\left\{-\rho\left(\overline{R}_m\right)\right\}} \quad (4)$$

where, the normalizing parameter, $\rho=\frac{k}{L}$. The packet length, L (in bits), is used for BE traffic to prevent its

data rate from increasing to infinity. The value of k=1 is adopted in the three utility functions. The optimization objective which maximizes the total utility of the network can be formulated as:

$$max_{\Re_m} \sum_{m=1}^{M} U(x_m).R_m[t] \quad (5)$$

Subject to

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

$$C2: \Re_n \bigcap \Re_m = \varnothing, n \neq m, \forall n, m \in \{1,2,.....M\}$$

where M is the total number of users in a cell, $\Re$ is the set of all subcarriers in the system, $\Re_m$ is the subset of subcarriers assigned to user m, $U(x_m)$ is a utility function based on a generic variable x_m that can represent a resource usage or QoS metric of user m and $R_m[t]$ is the instantaneous data rate of user m in time slot t. The optimization problem in Equation 5 is the maximization of the utility weighted sum rate. Constraints C1 and C2 state that the union of all subsets of subcarriers assigned to different users must be contained in the total set of subcarriers available in the system, and that these subsets must be disjoint, i.e., the same subcarrier cannot be shared by two or more users in the same time slot. Power constraints are not included because joint optimization of subcarrier and power is nonlinear and so it is complex to solve. Besides, optimal solutions are often difficult to be found. However, sub-optimal solutions that have been proposed in literature considered segregating the problem into two steps: first, dynamic resource assignment with fixed power allocation, second, adaptive power allocation with fixed resource assignment. We can simplify the optimization problem by skipping the second stage since the works in (Tung and Yao, 2002; Shen *et al*., 2003) have found that adaptive power allocation does not offer substantial gains over equal power allocation at high SNRs, Furthermore, Rhee and Cioffi, 2000 noted that equal power allocation offers a low complexity. When equal power allocation is applied, the problem in Equation 5 has a closed form solution as it's the objective function is now linear with respect to $R_m[t]$. As a result, the optimization objective function can be regarded as simply a dynamic resource allocation, whose weights are adaptively controlled by the utility function. In order to ensure that service bits for each user is less than or equal to available bits in his queue, a frugality constraint (Song *et al*, 2005) is imposed to avoid wastage of bandwidth. Therefore, for optimization objective function with equal power allocation among subcarriers and according to Song *et al*, 2005, we have that the user with index $m*$ is chosen to transmit on the subcarrier K at the time slot t if the condition below is satisfied:

$$m* = arg\ max_{\Re_m} \left\{ w_m[k,t].\min\left(r_m[k,t], \frac{Q_m[k,t]}{T_s} \right) \right\} \quad (6)$$

where $w_m[k,t]$ the scheduling is weight corresponding to the utility function for user m on subcarrier k in time slot t, $Q_m[k,t]$ is the queue for user m and $r_m[k,t]$ denotes the instantaneous achievable transmission rate of the subcarrier k with respect to user m during time slot t. The instantaneous achievable transmission rate is computed as $r_m[k,t] = \frac{B}{K}\log_2\left(1 + \frac{P_m[k,t]H_m[k,t]}{\delta\Gamma}\right)$, where $P_m[k,t]$ is the equal power on each subcarrier, $H_m[k,t]$ is the channel gain, δ is noise power and Γ is SNR gap. The SNR gap, Γ = -[ln(5.*BER*)/1.5], where BER=10^{-6} is assumed. Adaptive modulation and coding scheme (AMCS) of the data rate is computed as $r_m^*[k,t] = 2 \times round\left(\frac{r_m[k,t]}{2}\right)$. Therefore, the instantaneous achievable data rate in the optimization problem for user m is $R_m[t] = \min\left(r_m^*[k,t], \frac{Q_m[k,t]}{T_s}\right)$.

Simulation Results

In this section, we will highlight the distribution of average system throughput and user satisfaction among the traffic flows for the proposed scheduling algorithm, MQS, in comparison with EXP, UEPS, DSM/TSM, PF and MSR scheduling policies. The simulation considers a system bandwidth of 5MHz divided into 128 subcarriers and slot duration of 2.0571ms. We assume a total transmit power of 33.9897*dBm* and a total noise power of -151*dBm* at the receiver front-end. Apart from the Rayleigh flat fading which is based on the Stanford University Interim (SUI) channel model 4 (Erceg *et al*., 1999), the transmitted signal undergoes a distant-dependent path-loss given by:

$$PL(d)[dB] = 20\log\left(\frac{4\pi d_o}{\lambda}\right) + 10n\log\left(\frac{d}{d_o}\right) + \chi_\sigma \quad (7)$$

Where Path-loss exponent, n=3; reference distance, d_o=100*m*; shadowing, x_σ = 8*dB* and wavelength, λ = 120*mm* in a cell with radius, r = 1000*m*. A discrete-time system-level simulator was developed using the MATLAB simulation software package. The simulation took duration of 20s.

Traffic Model

The RT services are provided by video streaming. The FTP belongs to NRT class of service. For video streaming traffic, we consider that the packet with a size of 512 bits is exponentially generated at a packet inter-arrival time of 2ms with a maximum delay threshold of 100ms. The FTP traffic flow with a packet size of 1024 bits, requests minimum data rate of 9.6 kbps. Full buffer models in which the queue is infinite is adopted for the FTP (NRT) traffics. This assumption of infinite or full buffer model for simulation purposes is justifiable especially for next generation NRT traffics with very large amount of data to transmit. For RT traffic, the buffer size is computed as $Buffer_{RT} = \varpi \times D_m^{\max}$. This buffer management ensures that the system's delay budget for the RT traffic is satisfied.

Performance Metrics

The following performance metrics are compared for different scheduling and resource allocation algorithms in the downlink of an OFDMA system:

- Average throughput: The data rate of a user averaged over all time slots. It can also be defined as the average number of successfully delivered bits over the lifetime of the user's connection.
- User satisfaction: In RT scheduling, packets are allocated network resource before the expiration of its deadline otherwise the packet is useless and dropped from the base station queue. Therefore, an RT user is said to be satisfied if resources are allocated to him within the delay budget. Similarly, an NRT user is said to be satisfied if his average throughput during the session equals or exceeds his required minimum data rate. Session duration depends upon the number of time slots used in the simulation. The user satisfaction is, therefore, defined as the percentage of the ratio between the numbers of users who are satisfied in terms of their required QoSs to the total number of users in each service class. Mathematically, percentage of user satisfaction can be expressed by:

$$S_i = \frac{J_i^{satisfied}}{J_i} \times 100 \tag{8}$$

where $J_i^{satisfied}$ is the number satisfied users in service class i and J_i is the total number of users in each service class.

Resource Allocation with Real Time Services

In this section, we compare the average system throughput and user satisfaction for MQS, UEPS, EXP, DSM, PF and MSR scheduling algorithms. Packet dropping policy is implemented; as such the HOL packet is dropped from a user's buffer if it exceeds its delay budget (maximum tolerable delay). Figure 1 depicts the average system throughput as a function of the number of RT users.

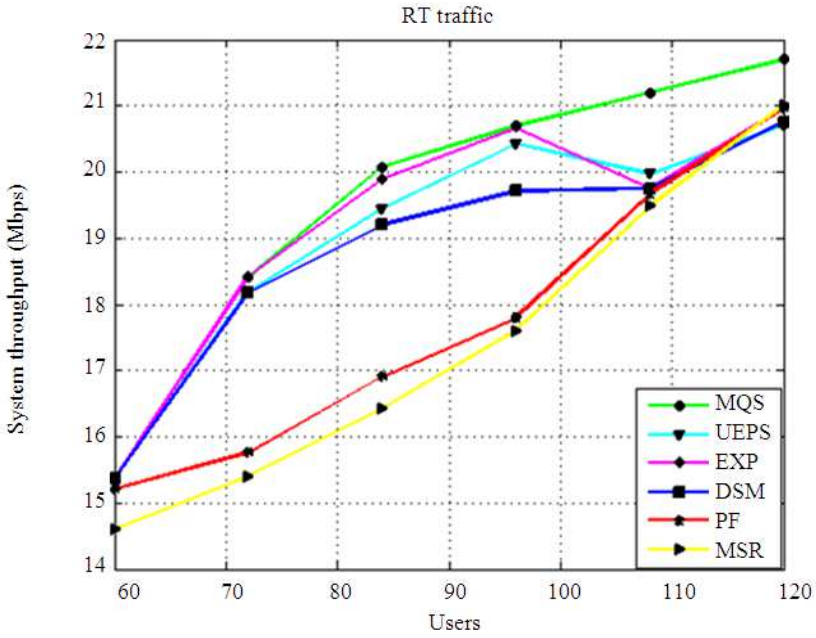

Fig. 1. Average system throughput for a video streaming service

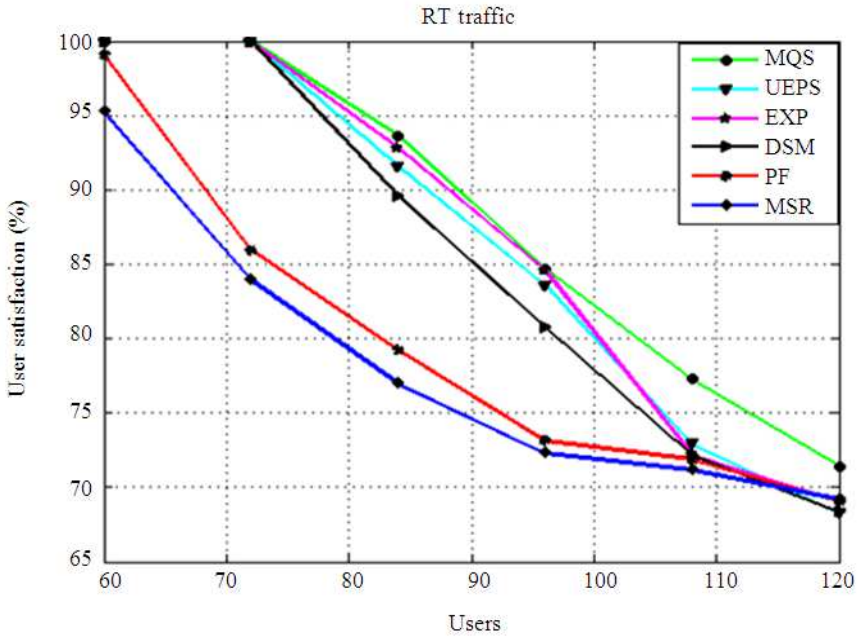

Fig. 2. User call satisfaction for a video streaming service

As it can be seen, the MQS achieves the best throughput performance followed by the other delay-aware algorithms such as UEPS, EXP and DSM. One would expect the opportunistic schedulers such as PF and MSR to present higher average system throughputs than the delay-aware schedulers. However, this was not so because MSR maximizes the system capacity by always choosing a few users and by nature of the RT traffic model used, the buffers of these few users do not have so much data to transmit hence the poorest throughput performance. Similarly, PF only slightly provides higher average system throughput than MSR, as it attempts to increase the number of selected users by using the user's relative channel condition to fairly distribute the available system resources.

Generally, the delay-aware schedulers perform better than the opportunistic ones because they are more adapted to avoiding excessive delays and minimizing packet loss.

Figure 2 shows the user call satisfaction for different scheduling schemes. It can be seen, for all the scheduling schemes, that when traffic load increases users become less satisfied because the available resources have to be shared by the increasing number of users. However, the MQS provides the highest ratio between the numbers of satisfied RT users to the total number of RT users in the network. For other algorithms, it can be observed that their achievable user satisfaction indexes tend to converge as the traffic load increases beyond the traffic load of 108 users while MQS still continues to maintain a superior performance.

Resource Allocation with Non-Real Time Services

This section compares MQS, TSM and PF scheduling policies for a scenario with NRT traffics that provide only FTP services. For NRT traffics, full-buffer traffic model is adopted. The average system throughput for various FTP traffic loads is depicted in Fig. 3. As expected, the MSR provides the highest average throughput by selecting fewer users who always have full buffers of data to transmit. The PF which sacrifices some of the system throughput for throughput fairness also performs better than MQS and TSM. The MQS achieves greater throughput compared to TSM. The reason for this is that TSM uses a bell-shaped utility function in which the utility for the user decreases rapidly after his required minimum throughput might have been achieved, whereas in the case of MQS the user's utility decreases slowly thus accumulating greater average throughput over the same scheduling interval.

Figure 4 shows that MQS achieves slightly higher user satisfaction performance compared to TSM. The PF performs better than MSR in terms of user satisfaction, although this achievement is lower compared to TSM and MQS. The extremely poor performance of MSR is expected, because the MSR always chooses fewer users to satisfy in terms of their required minimum rate requirement.

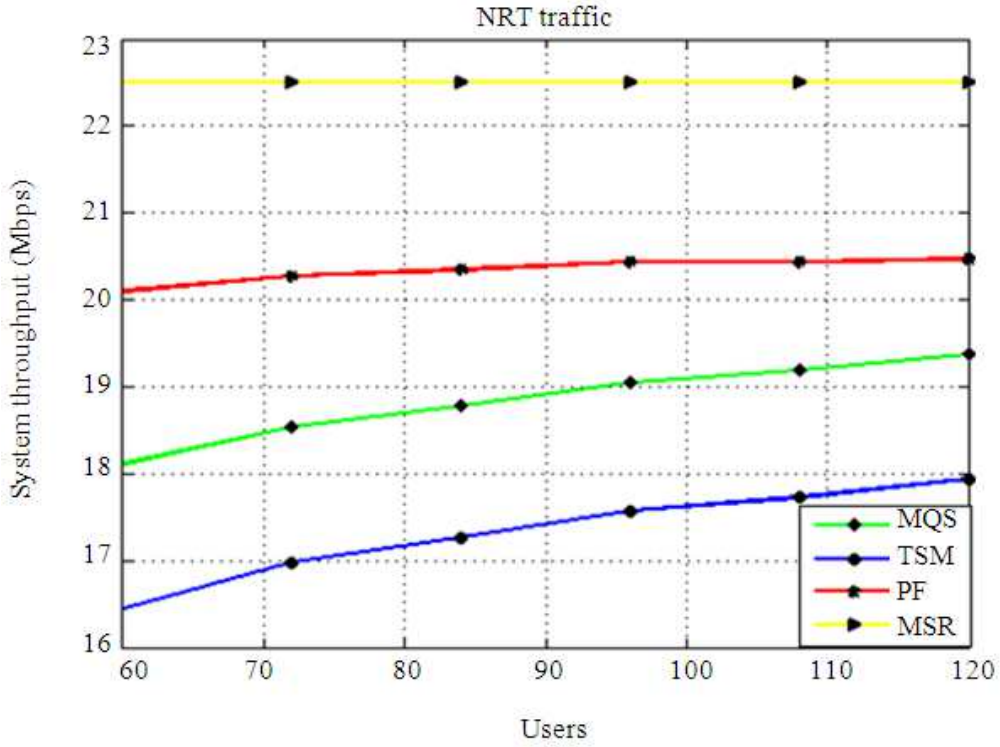

Fig. 3. Average system throughput for an FTP service

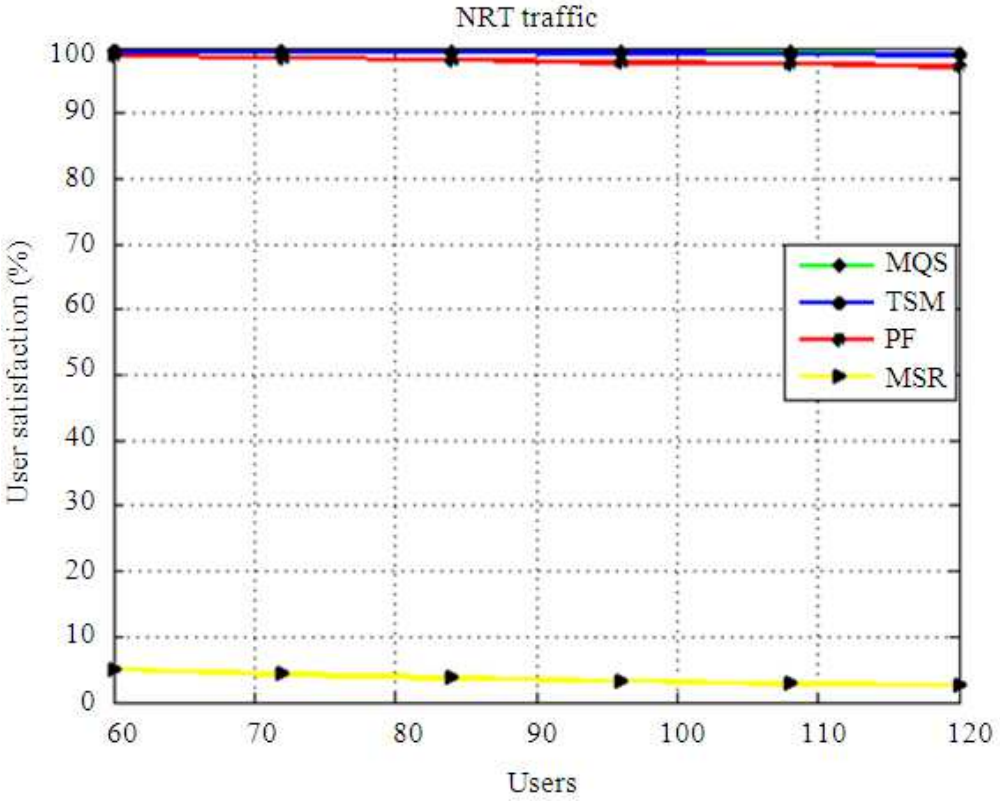

Fig. 4. User call satisfaction for an FTP service

Resource Allocation in a Mixture of Real Time and Non-Real Time (Heterogeneous) Services

In this section, we compare MQS, DSM/TSM and PF scheduling policies in a scenario with mixture of video streaming and FTP traffics with equal number of users. The MQS and DSM/TSM are both designed to schedule a mixture of delay-sensitive and minimum-rate-sensitive traffics at the same time; the PF is not.

Therefore, in Fig. 5, the PF is unable to allocate higher throughputs to higher-priority video streaming traffic. The MQS achieves similar throughput performance as DSM/TSM for video traffics. However, in the case of FTP traffics, it performs better than DSM/TSM which allocates almost zero throughputs when the number of users increases beyond 108. In terms of user satisfaction as depicted in Fig. 6, both MQS and DSM/TSM provide almost equal user satisfaction performance for video users. In the case of FTP, the MQS achieves almost a 100% user satisfaction while the user satisfaction achieved by DSM/TSM decreases as the traffic loads increase and reaching 0% at traffic load of 108 users. This is because DSM/TSM gives more priority to RT traffics. PF could not still allocate higher user satisfaction to video users than to FTP users. Unlike the PF, the DSM/TSM can satisfy video streaming users more than FTP ones; however, this is achieved at the expense of starving FTP in terms of the allocated resources.

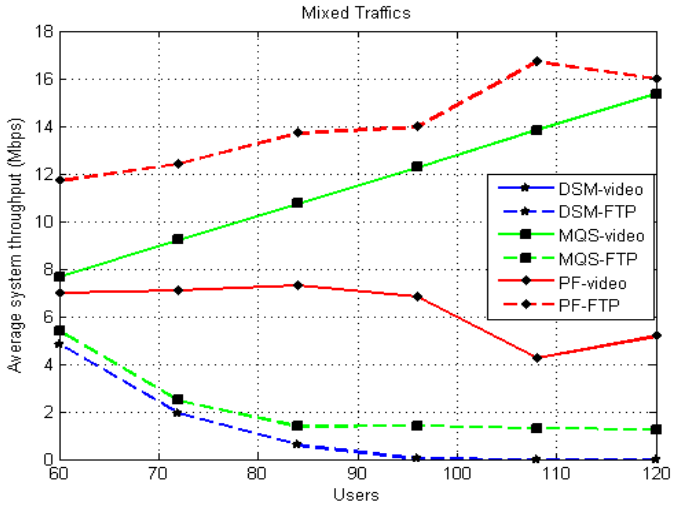

Fig. 5. Average system throughput in a mixture of video streaming and FTP traffics

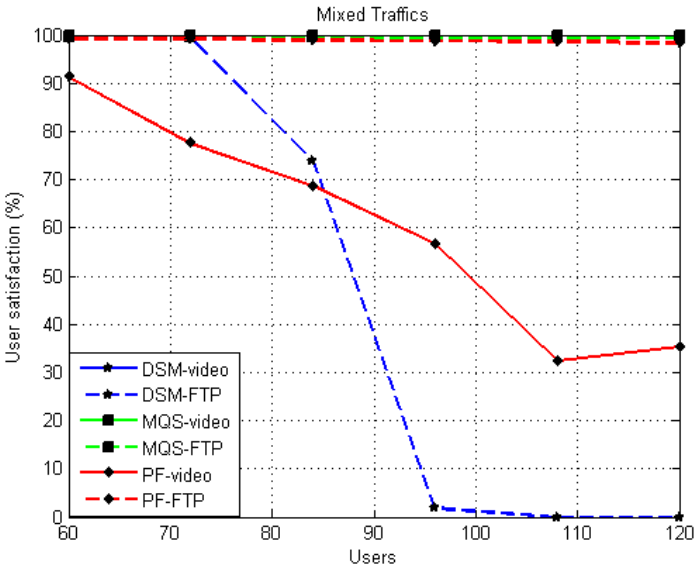

Fig. 6. User call satisfaction in a mixture of video streaming and FTP traffics

Conclusion

In this study, we studied the existing scheduling algorithms and found that most of them are designed to schedule both RT and NRT traffics using a single QoS metric (i.e. using traffic priority or packet delay for both RT and NRT). For efficient resource allocation and guarantee for QoS satisfaction, it is important that the common network resources are allocated to different traffics based on the specific QoS demanded. To achieve this, we developed a novel utility-based scheduling framework to allocate network resources to a mixture of RT, NRT and BE services using different utility functions as well as applying the specific QoS requirement for each service. The proposed scheduling framework, called Maximum QoS Satisfaction (MQS), is of low complexity because it requires only the QoS requirement of each admitted user to be supplied in the uplink transmission.

In the simulation, equal power allocation among frequency resources is assumed in order to further reduce the complexity of the optimization problem. The simulation study shows that MQS achieves the best average system throughput and user satisfaction performances in a scenario with RT services. In a scenario with NRT services, it was outperformed in terms of system throughput by MSR and PF as would be expected, owing to; one the nature of the full-buffer traffic model used for NRT in which each user has unlimited data in buffer and two the opportunistic nature of MSR and PF algorithms.

However, the MQS outperforms TSM, MSR and PF in terms of user satisfaction in NRT traffic scheduling. In a heterogeneous mixed traffic scenario, MQS, although allocates higher throughput to video streaming users than FTP ones, it does not discriminate much between the two in terms of users' QoS satisfaction. PF is not designed to handle QoS; hence it allocates higher throughput and higher user satisfaction to lower priority FTP traffic than to higher priority video traffic. Because DSM/TSM uses a similar utility curve to schedule both FTP and video traffics, it results in greatly starving the lower priority FTP traffic at the expense of higher priority video both in terms of throughput and user satisfaction. Hence, it has the potential of churning most of the FTP users out of the network. Finally, the MQS achieves its main objective of providing efficient QoS differentiation in a heterogeneous mixed traffic scenario; and thus enables network operators guarantee satisfactory provision of services to all in order to maintain a high number of subscribers, decrease churn and attract new subscribers.

Acknowledgement

This work is supported by DPI of the University of Malaysia, Sarawak.

Funding Information

The authors of this paper have no financial funding to report.

Author's Contributions

All the authors work in writing the manuscript in addition to the following:

Olalekan Bello: Studied the related works and formulated the proposed scheduling framework.

Hushairi Zen: Designed the discrete-level simulation algorithm and tested the model.

AL-Khalid Othman: Analyzed and interpreted the results.

Khairuddin Abdul Hamid: Designed the work plan and reviewed the presentation.

Ethics

This article is original and contains unpublished materials. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Al-Manthari, B., H. Hassanein, N.A. Ali and N. Nasser, 2009. Fair class-based downlink scheduling with revenue considerations in next generation broadband wireless access systems. IEEE Trans. Mobile Comput., 8: 721-734. DOI: 10.1109/TMC.2009.30

Andrews, M., K. Kumaran, K. Ramanan, A. Stolyar and P. Whiting *et al*., 2000. CDMA data QoS scheduling on the forward link with variable channel conditions. Bell Labaratories, Lucent Technologies.

Andrews, M., A. Stolyar, K. Kumaran, R. Vijayakumar and K. Ramanan *et al*., 2001. Providing quality of service over a shared wireless link. IEEE Commun. Magazine, 39: 150-154. DOI: 10.1109/35.900644

Balakrishnan, R. and B. Canberk, 2014. Traffic-aware QoS provisioning and admission control in OFDMA hybrid small cells. IEEE Trans. Vehicular Technol., 63: 802-810. DOI: 10.1109/TVT.2013.2280124

Dahlman, E,. S. Parkvall, J. Sköld and P. Beming, 2007. 3G Evolution: HSPA and LTE for Mobile Broadband. 1st Edn., Academic Press, Jordan Hill, Oxford, ISBN-10: 9780123725332, pp: 448.

Erceg, V., L.J. Greenstein, S.Y. Tjandra, S.R. Parkoff and A. Gupta *et al*., 1999. An empirically based path loss model for wireless channels in suburban environments. IEEE J. Selected Areas Commun., 17 (7): 1205-1211. DOI: 10.1109/49.778178

Forkel, I., H. Klenner and A. Kemper, 2005. High Speed Downlink Packet Access (HSDPA)-Enhanced Data Rates for UMTS Evolution. Computer Networks, Elsevier, 49: 325-340. DOI: 10.1016/j.comnet.2005.05.012

IEEE Computer Society, I.M.T. and T. Society, 2005. IEEE standards for local and metropolitan area networks-Part 16: Air interface for fixed and mobile broadband wireless access systems-amendment 3: Management PLANe procedure and services. I.M.T. IEEE Computer Society.

Jalali, A., R. Padovani and R. Pankaj, 2000. Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. Proceedings of the 51st IEEE Vehicular Technology Conference, May 15-18, IEEE Xplore Press, Tokyo, pp: 1854-1858. DOI: 10.1109/VETECS.2000.851593

Kelly, F.P., A.K. Maulloo and D.K.H. Tan, 1998. Rate control for communication networks: Shadow prices, proportional fairness and stability. J. Operat. Res. Society, 49: 237-252. DOI: 10.1057/palgrave.jors.2600523

Kelly, F., 1997. Charging and rate control for elastic traffic. Eur. Trans. Telecommun., 8: 33-37. DOI: 10.1002/ett.4460080106

Knopp, R. and P.A. Humblet, 1995. Information capacity and power control in single-cell multiuser communications. Proceedings of the IEEE International Conference on Communications, Jun. 18-22, IEEE Xplore Press, Seattle, WA., pp: 331-335. DOI: 10.1109/ICC.1995.525188

Lee, B.G., D. Park and H. Seo, 2009. Wireless Communications Resource Management. 1st Edn., Wiley-IEEE Press, ISBN-10: 978-0-470-82356-9, pp: 320.

Lima, F.R.M, E.B. Rodrigues, T.F. Maciel and M. Nordberg, 2014. Resource Allocation for Improved User Satisfaction with Applications to LTE. Resource Allocation and MIMO for 4G and Beyond, Porto R. (Ed.), Springer-Verlag New York, New York, ISBN-10: 978-1-4614-8057-0, pp: 63-104.

Rhee, W. and J.M. Cioffi, 2000. Increase in capacity of multiuser OFDM system using dynamic subchannel allocation. Proceeding of the 51st IEEE Vehicular Technology Conference, May 15-18, IEEE Xplore Press, Tokyo, pp: 1085-1089. DOI: 10.1109/VETECS.2000.851292

Rodrigues, E.B. and F. Casadevall, 2009. Adaptive radio resource allocation framework for multi-user OFDM. Proceedings of the 69th IEEE Vehicular Technology Conference, Apr. 26-29, IEEE Xplore Press, Barcelona, pp: 1-6. DOI: 10.1109/VETECS.2009.5073363

Ryu, S., B. Ryu, H. Seo and M. Shin, 2005. Urgency and efficiency based packet scheduling algorithm for OFDMA wireless system. IEEE International Conference on Communications, 4: 2779-2785. DOI: 10.1109/ICC.2005.1494854

Shakkottai, S. and A.L. Stolyar, 2001. Scheduling algorithms for a mixture of real-time and non-real-time data in HDR. Int. Teletraffic Sci. Eng., 4: 793-804. DOI: 10.1016/S1388-3437(01)80170-0

Shen, Z., J.G. Andrews and B.L. Evans, 2003. Optimal power allocation in multiuser OFDM systems. Proceedings of the IEEE Global Telecommunications Conference, Dec. 1-5, IEEE Xplore Press, pp: 337-341. DOI: 10.1109/GLOCOM.2003.1258257

Song, G. and Y. Li, 2005. Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks. IEEE Commun. Magazine, 43: 127-134. DOI: 10.1109/MCOM.2005.1561930

Song, G.C., 2005. Cross-layer resource allocation and scheduling in wireless multicarrier networks. PhD thesis, Georgia Institute of Technology, Georgia.

Tung, T.L. and K. Yao, 2002. Channel estimation and optimal power allocation for a multiple-antenna OFDM system. EURASIP J. Applied Signal Process., 2002: 330-339. DOI: 10.1155/S1110865702000689

Wang, H. and W. Jia, 2010. An optimized scheduling scheme in OFDMA WiMAX networks. Int. J. Commun. Syst., 23: 23-39. DOI: 10.1002/dac.1106

Wang, X., G.B. Giannakis and A.G. Marques, 2007. A unified approach to QoS-guaranteed scheduling for channel-adaptive wireless networks. Proc. IEEE, 95: 2410-2431. DOI: 10.1109/JPROC.2007.907120