EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

QoS-aware composite scheduling using fuzzy proactive and reactive controllers

Nabeel Khan[1*], Maria G Martini[1] and Dirk Staehle[2]

Abstract

We consider in this paper downlink scheduling for different traffic classes at the MAC layer of wireless systems based on orthogonal frequency division multiple access (OFDMA), such as the recent 3rd Generation Partnership Project (3GPP) long-term evolution (LTE)/LTE-A wireless standard. Our goal is to provide via the scheduling decisions quality of service (QoS), but also to guarantee fairness among the different users and traffic classes (including delay-sensitive and best-effort traffic). QoS-aware scheduling strategies, such as modified largest weighted delay first (M-LWDF), exponential (EXP), exponential proportional fair (EXP-PF), and the log-based scheduling rules, prioritize delay-sensitive traffic by considering rules based on the head-of-line (HoL) packet delay and the tolerated packet loss rate, whereas they serve best-effort traffic by considering the classical proportional fair (PF) rule. These scheduling rules do not prevent resource starvation for best-effort traffic. On the other side, if both traffic types are scheduled according to the PF rule, then delay-sensitive flows suffer from delay bound violations. In order to fairly distribute the resources among different service classes according to their QoS requirements and channel conditions, we employ the concept of fuzzy logic in our scheduling framework. By employing the fuzzy logic concept, we serve all the traffic classes with one priority rule. Simulation results show better intra-class and inter-class fairness than state-of-the-art scheduling rules. The proposed scheduling framework enables to appropriately balance delay requirements of traffic, system throughput, and fairness.

1 Introduction

3rd Generation Partnership Project (3GPP) Release 8, and its subsequent modifications, define the long-term evolution (LTE) standard [1] that will take the cellular technology in the 2020s. In wireless communication systems, radio resources are shared by multiple users; hence, medium access control (MAC) layer scheduling becomes extremely important in determining the overall performance of an LTE system. The efficiency of the link level, from the LTE base station (eNodeB) to the mobile terminal, largely depends on the design of the scheduler, whose task is to determine which users should be served and to assign resources.

An efficient scheduler must ensure a good trade-off between efficiency and fairness in the system (according to the service needs of each user) by fully utilizing the available radio resources. MAC layer scheduling strategies can be classified as channel-aware and channel-unaware, where channel aware scheduling algorithms take channel conditions into account and maximize the system throughput. Note, however, that the main target of mobile operators would be the end-user satisfaction, not merely the maximization of system throughput. Scheduling in the LTE standard is more challenging than in earlier standards mainly because earlier standards were based on single carrier systems, where resources were usually divided in terms of time slots or codes among the users, whereas LTE is a multicarrier system where system resources are shared among users in terms of time and frequency sub-bands.

Some approaches solve the problem of resource allocation using optimal solutions, and in other cases, resource allocation and resource assignments are performed in two separate steps; other approaches simply target at adapting schemes originally proposed for time division multiple access (TDMA) to orthogonal frequency division multiple access (OFDMA) systems. Thus, scheduling solutions available in the literature broadly fall into three classes.

*Correspondence: n.khan@kingston.ac.uk
[1]Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Rd, Kingston upon Thames, Surrey KT1 2EE,UK
Full list of author information is available at the end of the article

1. In [2-5], resource allocation is modeled as a convex optimization problem. The water-filling algorithm is used to solve the convex optimization problem by considering a continuous objective function. Linear integer programming is also widely used in solving the resource allocation problem by first transforming the nonlinear optimization problem into a linear problem. The main drawback of these strategies is the high computation complexity. Since the transmission time interval (TTI) in LTE is only 1 ms, these algorithms are not feasible from an implementation point of view.
2. In the second class of approaches, such as in [6-8], scheduling is performed in two steps. The first step consists of resource allocation, which determines the number of resources allocated to each user. The resource allocation step is followed by the resource assignment step, which determines which resources are assigned to each user. This class of scheduling algorithms are specifically designed for delay-sensitive applications and does not provide a priority differentiation between delay-sensitive and best-effort flows.
3. The third approach is the adaptation of TDMA strategies for OFDMA systems. Scheduling rules designed for single carrier systems such as the proportional fair (PF) [9], modified largest weighted delay first (M-LWDF) [10], and exponential proportional fair (EXP-PF) [11] are adapted for an OFDMA system by calculating these rules on each resource. This adaptation is referred to as an OFDMA/TDMA strategy. These scheduling rules are analyzed by [12] for delay-sensitive applications over an LTE system. According to [12], M-LWDF is the best scheduling rule for delay-sensitive applications in terms of fairness and efficiency. A very good survey on these scheduling strategies for LTE is provided in [13]. As each of these scheduling rules are based on the proportional fair rule, the calculation of these scheduling metrics on each physical resource block (PRB) allows the exploitation of multi-user time and frequency diversities. The complexity of the OFDMA/TDMA approach grows linearly with the number of users and resources. Thus, it can be implemented in real time. However, for delay-sensitive traffic, these scheduling rules cannot provide fairness for users with relatively low signal-to-interference noise ratio (SINR) [14].

In this work, we address the following issues of the third class of strategies:

- Intra-class fairness issues for delay-sensitive traffic: scheduling rules for delay-sensitive traffic consider the ratio of instantaneous channel quality and time-averaged throughput (proportional fair rule) along with either the linear [10], logarithmic [15], or exponential [11,15] function of the head-of-line (HoL) delay [16]. The HoL delay refers to the amount of time packets that reside in the buffer and is also known as the sojourn time. It is important to note that the video is delay-sensitive traffic; hence, packets arriving late are generally not useful at the receiver. Therefore, packets are associated with a predefined HoL delay bound and packets violating the delay bound are dropped from the queue. The utilization of HoL delay and the proportional fair rule in the scheduling decisions are not sufficient to avoid delay bound violation of flows having lower channel quality. Video traffic exhibits highly variable bit rate characteristics, i.e., the instantaneous peak rate is higher than the average rate. Lower channel quality video flows exhibiting peak instantaneous rate have high probability of delay bound violation mainly because of the proportional fair rule in the scheduling decisions. In other words, these scheduling rules achieve higher HoL delay for the packets of flows having higher average rate and lower channel quality. On the other hand, flows having good channel quality and lower average rate are scheduled way before their delay bound. The probability of delay bound violation of the flows exhibiting lower channel quality and higher average rate is very high which results in an unfair system.
- Inter-class fairness issues: in the literature [13], composite scheduling rules serve the best-effort traffic by using the classical proportional fair rule, i.e., ratio of instantaneous channel quality to the time-averaged throughput [9,17-19]. They prioritize delay-sensitive traffic by considering either the logarithmic, exponential, or linear function of the HoL delay. However, such composite scheduling strategies result in a higher priority difference between the delay-sensitive and best-effort traffic classes. In other words, the higher the difference among the scheduling priorities of traffic classes, the lower will be the multi-user channel diversity exploitation. In LTE, multi-user channel diversity has more significance since it is a multi-carrier system which allows multi-user diversity exploitation in the time and frequency domain.

By using the concept of fuzzy logic priority [20], we couple the flow's delay urgency (ratio of packet's HoL delay and delay bound) with the time-averaged channel quality. Instead of exploiting the time-averaged throughput and the linear, logarithmic, or exponential function of the HoL delay, we use a fuzzy function of the HoL

delay coupled with time-averaged channel quality as introduced in [21]. In [21], the HoL delay along with the time-averaged channel quality is processed by a fuzzy proactive controller. Further, whenever a flow suffers a delay bound violation, the scheduler reacts to this event and increases the priority of that flow. The delay bound violation input is processed by a fuzzy reactive controller. In this work, we propose a composite scheduling rule for delay-sensitive as well as the best-effort traffic. In the earlier work, the scheduling rule considers only the video traffic. In this work, the scheduling rule and scenarios are extended to handle more than one delay-sensitive traffic types. Furthermore, the main goal of the proposed composite scheduling rule is to balance the probabilities of quality of service (QoS) violation of the delay-sensitive as well as the best-effort traffic types.

A block diagram representing the proposed fuzzy composite scheduling (FCS) is given in Figure 1. The scheduling metric comprises a time-domain priority component based on reactive and proactive controllers and a frequency domain priority based on detailed information on instantaneous channel quality indicator (CQI) feedback per PRB. In order to dynamically adjust the priority level between best-effort and delay-sensitive flows, we utilize a fuzzy-based dynamic resource controller (DRC) (discussed in Section 3.3), as shown in Figure 1. Intra-class fairness (fairness in terms of achieved QoS among the flows within each of the traffic classes) is provided by the fuzzy proactive and reactive controllers whereas inter-class fairness (priority differentiation between the delay-sensitive and best-effort flows) is provided by the DRC. In fuzzy logic, the *output fuzzy set* is defined as the range of all possible output values that can be assigned to a fuzzy controller. The output of the controller lies within the output fuzzy set. The larger the output fuzzy set, the higher the priority of the controller. In the proposed scheduling framework, each traffic class has its own output fuzzy set. We assign a fixed output fuzzy set to the delay-sensitive traffic class and dynamically adjust the output fuzzy set of the best-effort traffic. The output fuzzy set of the best-effort traffic class is set by the DRC based on the latency (packet's HoL delay) and QoS violation of the delay-sensitive flows as shown in the figure. The higher the latency and QoS violations of the delay-sensitive flows, the lower the output fuzzy set of the

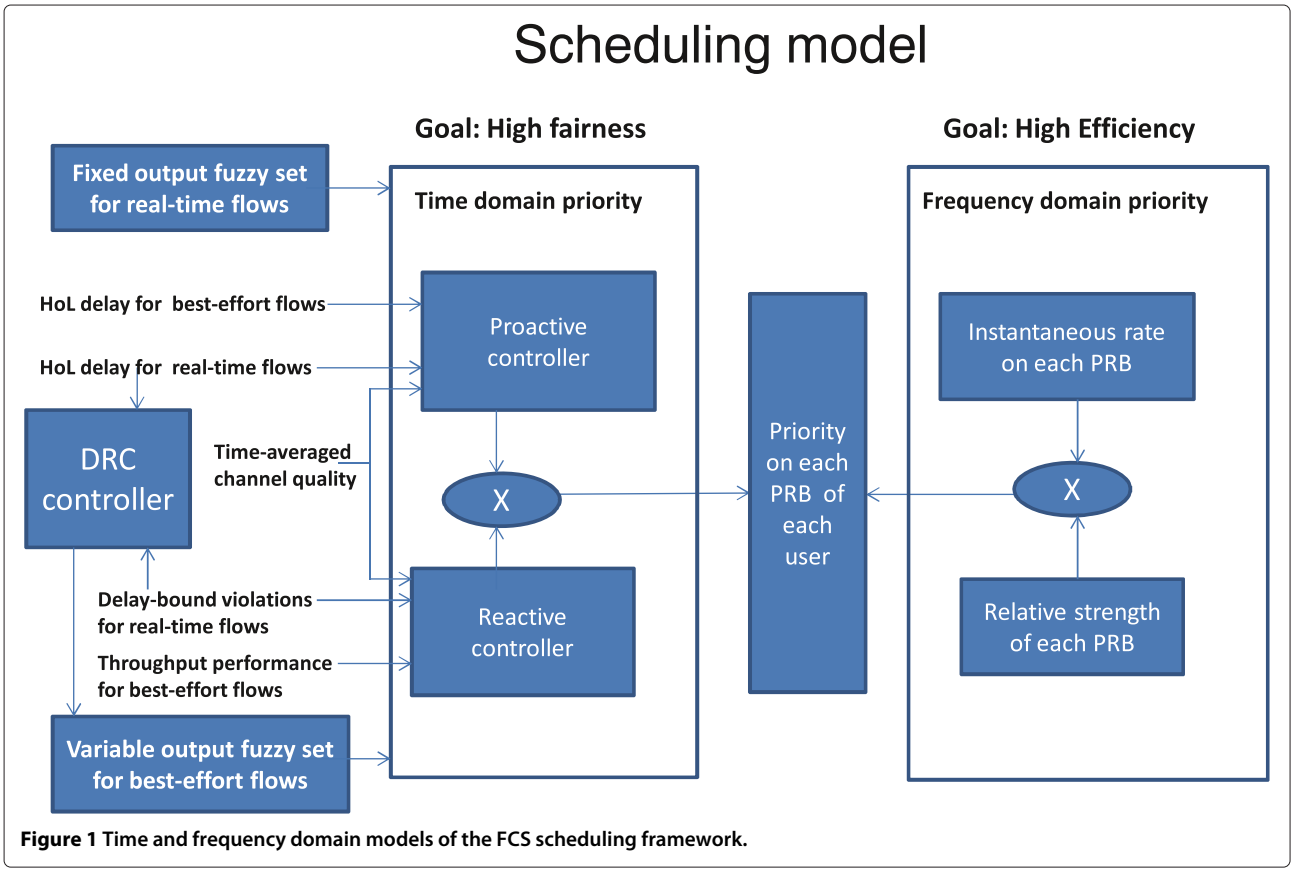

Figure 1 Time and frequency domain models of the FCS scheduling framework.

best-effort traffic. The final priority on each PRB is a function of the time and frequency domain priority metrics as shown in Figure 1.

The remainder of this paper is organized as follows. Section 2 presents the considered system model and the problem statement. Section 3 presents the details of our fuzzy logic-based scheduling strategy. Section 4 presents the performance evaluation of the proposed approach. In particular, the solutions considered as benchmark for the assessment of our scheduling algorithm are presented in Section 4.1, whereas the simulation setup is presented in Section 4.2; results are presented and discussed in Section 4.3. Conclusions are drawn in Section 5.

2 System model and problem statement

We consider a multiuser downlink single input single output (SISO) LTE/LTE-A system. The single-cell scenario comprises an eNodeB MAC scheduler responsible for allocating PRBs to different users in the cell. Each user i is assigned a buffer at the eNodeB, and packets of different traffic classes are streamed into the buffer of the eNodeB. For delay-sensitive traffic, we consider video and VoIP traffic (the scheduling framework can accommodate all LTE service classes), whereas for best-effort traffic, we consider constant bit rate (CBR) traffic. The packets of each traffic class entering the buffer are time stamped by the scheduler. Packets of delay-sensitive traffic are dropped from the buffer if the HoL packet delay is longer than the target HoL delay bound. The main QoS parameters for video and VoIP flows are the HoL packet delay and the packet loss rate (PLR), whereas throughput is the important QoS parameter for the flows of best-effort traffic. We consider the HoL delay for best-effort traffic, and we assign a target delay for the flows of this traffic class. However, since we can assume best-effort traffic is delay tolerant, therefore, packets violating the target HoL delay are not dropped from the buffer. We use a CQI feedback mechanism, where each user feedbacks information about the channel quality on each PRB. Due to the adoption of adaptive modulation and coding (AMC) in LTE, each CQI value corresponds to a specific value of spectral efficiency for each PRB.

At scheduling epoch n, we define the normalized time-averaged wideband spectral efficiency, $\overline{\chi}_i^{(n)}$, of user i over the moving average window of size n_c as:

$$\overline{\chi}_i^{(n)} = \frac{1}{\chi_{\max}} \left[\frac{1}{n_c} \sum_{k=n-n_c}^{n} \chi_i^{(k)} \right] \tag{1}$$

where

$$\chi_i^{(n)} = \frac{1}{M_{\mathrm{PRB}}} \sum_{\varphi=1}^{M_{\mathrm{PRB}}} \chi_{i,\varphi}^{(n)} \tag{2}$$

is the average PRB spectral efficiency of user i at scheduling instant n and $\chi_{i,\varphi}^{(n)}$ is the instantaneous subband spectral efficiency of user i at PRB φ. $\chi_{\max}$ is a constant, i.e., the spectral efficiency (5.5547 bits/s/Hz) corresponding to the maximum CQI feedback, and M_{PRB} is the number of PRBs available for allocation at each scheduling epoch.

Given the available information about:

- the HoL packet delay for each flow $H_i^{(n)}$,
- the channel quality of each flow on each PRB, hence the resulting spectral efficiency $\chi_{i,\varphi}^{(n)}$,
- the tolerated delay bound $H_{\max}$,
- the QoS performance of the delay-sensitive flows in terms of packet loss ratio, $\mathrm{plr}_i^{(n)}$ and of the best-effort flows in terms of time-averaged throughput $R_{i,\mathrm{ave}}^{(n)}$,

the scheduling problem is defined as: *How to allocate to the different users the M_{PRB} PRBs in each scheduling interval in order to fulfill the QoS requirements of each of the flows from different traffic classes so that a good trade-off between fairness and efficiency is achieved*.

In order to mathematically formulate the problem, let us define the following parameters:

$R_i^{(n)}$: Throughput achieved by flow i at scheduling instant n.

I: Total number of flows in the system. It is the sum of delay-sensitive $I_{\mathrm{delay-sensitive}}$ and best-effort $I_{\mathrm{best-effort}}$ flows.

$\mathrm{plr}_{\mathrm{i}}^{(\mathrm{n})}$: The packet loss ratio of flow i at scheduling instant n calculated over the moving average transmission window t_w:

$$\mathrm{plr}_i^{(n)} = \frac{\sum_{m=n-t_w}^{n} P_{\mathrm{drop}_i}^{(m)}}{\sum_{m=n-t_w}^{n} \left(P_{\mathrm{transmit}_i}^{(m)} + P_{\mathrm{drop}_i}^{(m)} \right)} \tag{3}$$

where

$P_{\mathrm{transmit}_i}^{(m)}$: Number of transmitted packets of flow i over the moving average transmission window t_w.

$P_{\mathrm{drop}_i}^{(m)}$: Number of dropped packets of flow i over the moving average transmission window t_w.

The main goal of the scheduler is to maximize the system throughput $R_{t_w}^{(n)}$, subject to the QoS constraints of the

delay-sensitive flows, over the moving average transmission window t_w:

$$R_{t_w}^{(n)} = \max\left(\sum_{i=1}^{I}\sum_{m=n-t_w}^{n} R_i^{(m)}\right) \tag{4}$$

subject to

$$\frac{1}{I_{\text{delay-sensitive}}}\sum_{i=1}^{I_{\text{delay-sensitive}}} \mathbb{1}\left(\text{plr}_i^{(n)} \le \text{plr}_{\text{thr}}\right) = 1 \tag{5}$$

where

$\mathbb{1}\left(\text{plr}_i^{(n)} \le \text{plr}_{\text{thr}}\right)$ is an indicator function equal to 1 if its argument is true, i.e., when the packet loss rate of flow i is lower or equal than the threshold value plr_{thr}. If the packet loss rate exceeds the threshold, then the indicator function is 0. It is important to note that fairness for delay-sensitive traffic is guaranteed when the PLR over a short moving average window [22], for instance one second, is below the prescribed threshold for each of the delay-sensitive flows in the system. As noted in [23], when the scheduler achieves short-term fairness, then the long-term fairness is guaranteed.

The optimal solution of the above problem is not possible without restrictive assumptions on the arrival process of the traffic and changes in channel quality. Therefore, we propose a novel scheduling framework relying on fuzzy logic. Fuzzy logic is ideally suited for problems where a definite mathematical solution is unavailable. The information about the changes in the radio channel and the traffic rate of each user is uncertain. Fuzzy logic can deal with such situations because of its capability to make approximate reasoning. In our proposed scheduling strategy, each PRB is assigned to the user maximizing a defined metric. Our proposed metric is composed of a PRB-independent part and a PRB-specific part. The PRB-independent part calculated for a user describes the 'urgency' of an assignment as time-domain priority, whereas the PRB-specific part describes the instantaneous channel quality of the PRB and its relative quality versus other PRBs.

3 Fuzzy composite scheduling framework

The FCS framework consists of fuzzy proactive, reactive, and DRC controllers. It is important to note that the designs of the proactive and reactive controllers are the same. The proactive controller processes the HoL delay whereas the reactive controller processes the QoS violation. In the following, we present a detailed design of the three fuzzy controllers:

3.1 Proactive controller

The goal of the proactive controller is to avoid delay bound violations. In order to consider the delay urgency in a dynamic wireless environment, we propose a novel concept of utilizing time-averaged channel quality over a small moving window by using the average wideband spectral efficiency $\overline{\chi}_i^{(n)}$ associated to the CQI feedback, defined in Equation 1. The proactive controller processes two inputs. One of these is the HoL packet delay $H_i^{(n)}$ normalized to the maximum tolerated HoL delay $H_{\max}$ of each traffic class. The normalized HoL delay input, H_i', for the proactive controller is:

$$H_i' = \begin{cases} \frac{H_{i,\text{VoIP}}^{(n)}}{H_{i,\max}}, & \text{for VoIP traffic class} \\ \frac{H_{i,\text{video}}^{(n)}}{H_{i,\max}}, & \text{for Video traffic class} \\ \frac{H_{i,\text{best-effort}}^{(n)}}{H_{i,\max}}, & \text{for best} - \text{effort traffic class} \end{cases} \tag{6}$$

The goal of the controller is to be proactive for any possible delay violations; hence, the second input is designed as the weighted sum of the normalized delay and the normalized average channel quality. It is mathematically defined as:

$$\chi_{H'} = 0.5(1 - H_i') + 0.5(\overline{\chi}_i^{(n)}) \tag{7}$$

The rationale behind the weighted sum Equation 7 is discussed in Section 3.1.1.

In fuzzy logic, the input membership function represents the magnitude of the inputs which are mapped to the output membership function through a set of rules [20]. The membership functions can be linear, exponential, bell shaped, or any other shape according to the system requirements. According to [12], the M-LWDF scheduling rule, linear function of the HoL packet delay, outperforms the EXP-PF scheduling rule which is an exponential function of HoL packet delay. Therefore, we select linear membership functions for the proactive and reactive controllers. The graphical representation of the input and output membership functions is shown in Figure 2a and 2b, respectively. The same input membership functions are used for both the inputs (H_i' and $\chi_{H'}$). It is important to note that users with better channel quality result in a higher frequency domain priority on each PRB φ, as there will be a higher number of PRBs with better channel quality. Therefore, the time domain priority should be higher for users with higher normalized HoL packet delay and lower normalized channel quality.

Now, we will utilize the flexibility of fuzzy logic by mapping the input membership functions to the output memberships functions through a set of rules. Let μ_p be the output of the proactive controller (defuzzified proactive priority value), the fuzzy rules for the proactive controller are as follows:

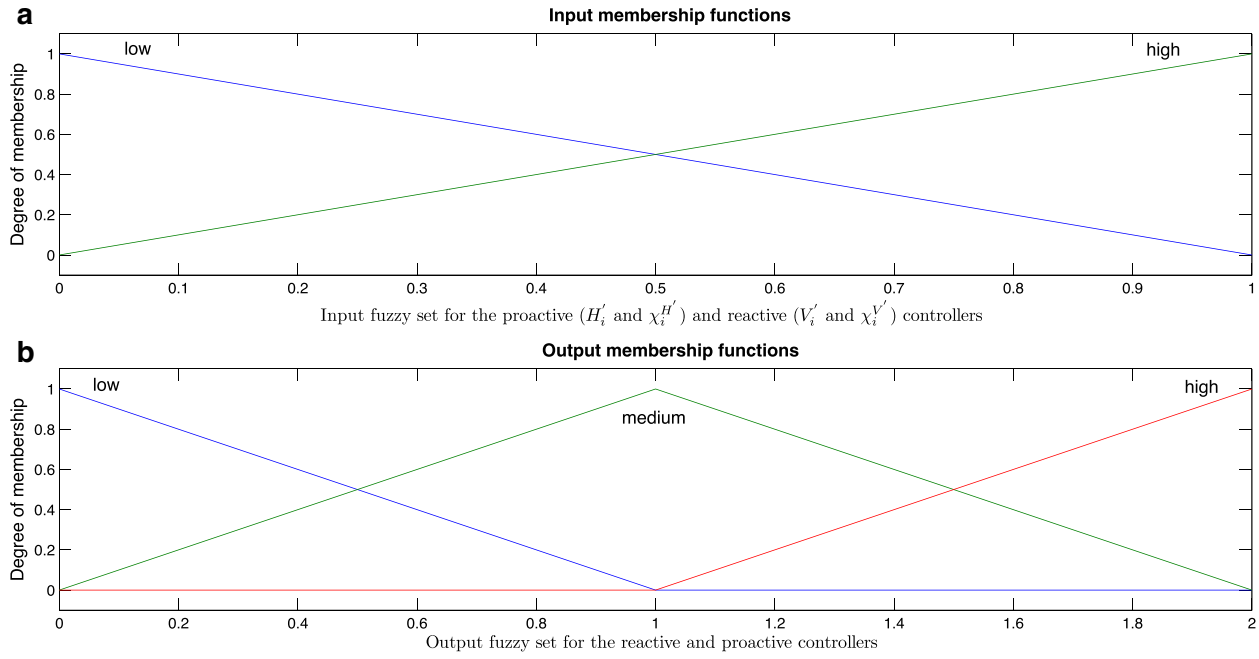

Figure 2 Input and output membership functions. (a) Input membership functions low and high of the proactive and reactive controllers. **(b)** Output membership functions low, medium, and high of the proactive and reactive controllers.

1. If H_i' is low AND $\chi_{H_i'}$ is low THEN μ_p is medium
2. If H_i' is low AND $\chi_{H_i'}$ is high THEN μ_p is low
3. If H_i' is high AND $\chi_{H_i'}$ is low THEN μ_p is high
4. If H_i' is high AND $\chi_{H_i'}$ is high THEN μ_p is medium

where *low*, *medium*, and *high* are the output membership functions as shown in Figure 2 and μ_p is the crisp output which along with the reactive controller output quantifies the time domain priority of each user. The main motivation of using the low, medium, and high output membership functions is to prioritize flows suffering from lower channel quality and higher HoL delay. The priority of the users with relatively good channel quality increases from low to medium as the HoL delay increases. On the other hand, the priority of users with lower channel quality increases from medium to high. Therefore, fairness is incorporated in the scheduling decisions through the output membership functions and rules of the fuzzy controllers. The main goal of the frequency domain priority is to improve the system efficiency whereas the time domain priority provides fairness through fuzzy proactive and reactive controllers.

The output fuzzy set of the membership functions, shown in Figure 2, determines the traffic priority of each traffic class. It is important to note that μ_p lies within the output fuzzy set. The proactive priority, μ_p, as a function of the inputs H_i' and $\chi_{H_i'}$ is shown in Figure 3.

The steps involved in producing a crisp output in the fuzzy logic system are described below.

1. Fuzzification. This is the process of converting fuzzy input values into a degree of membership for each linguistic term. Each linguistic term, high, medium, and low, characterizes a membership function. For instance, the proactive controller inputs, H_i' and $\chi_{H_i'}$, as shown in Figure 4, are fuzzified by the input membership functions low and high. In the figure, the four rows are the four rules of the proactive controller. Rule one comprises only low membership function, therefore input H_i' and $\chi_{H_i'}$ are fuzzified by the low membership function as shown in the figure.
2. Fuzzy inference. This is the core process of the fuzzy logic system, comprising a mapping from a given input to an output using the membership functions and logical operators in the *if-then-else* rules. According to Figure 4, the AND logical operation is performed, according to which the minimum of the two fuzzified inputs is mapped to the output membership function. The result of the fuzzy inference process is the degree of the output membership functions fulfilled by the logical operations between the fuzzified inputs. The result is the truncated output membership functions as shown in the third column of Figure 4.

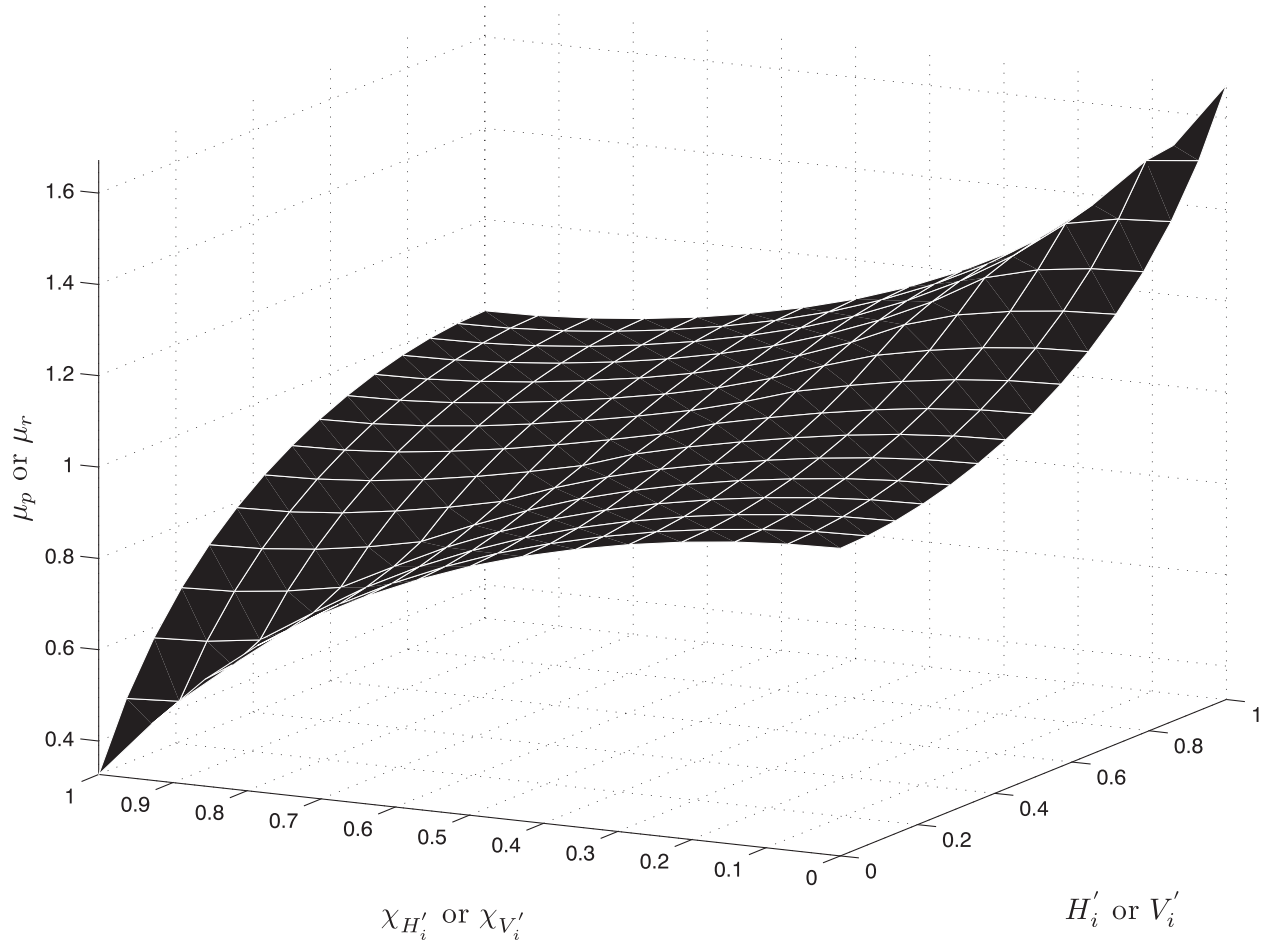

Figure 3 Proactive controller output, μ_p, w.r.t the inputs.

3. 'Defuzzification' and production of the final 'crisp' output. The crisp proactive priority output μ_p produced is shown in Figure 4. The output of each rule is combined to give the final fuzzy set, as shown in the fifth row and third column in Figure 4. The defuzzification process is simply the centroid calculation on the final fuzzy set as shown in Figure 4.

3.1.1 Rationale

The inputs H_i' and $\chi_{H_i'}$ are a function of the HoL delay; hence, the system is made more proactive for any possible delay violations. The second input, weighted sum of the normalized HoL delay and time-averaged channel quality, enhances the system fairness. For instance, consider two users - user 1 and 2 - having normalized average channel quality of 0.8 and 0.6, respectively, and normalized HoL packet delay of 0.4 and 0.8, respectively. If we select the second input as a simple function of the average channel quality, i.e., $\chi_{H_i'} = \overline{\chi}_i^{(n)}$, then the output of the proactive controller, μ_p (Figure 3), for user 1 and 2 is 0.844 ($H_i' = 0.4$, $\chi_{H_i'} = 0.8$) and 1.05 ($H_i' = 0.8$, $\chi_{H_i'} = 0.6$), respectively. The difference in the proactive priority of the two users is $1.05 - 0.844 = 0.206$. On the other hand, if the weighted sum Eq. 7 is used, then the output for user 1 and 2 is 0.872 ($H_i' = 0.4$, $\chi_{H_i'} = 0.6$) and 1.14 ($H_i' = 0.8$, $\chi_{H_i'} = 0.7$), with a difference in proactive priority of 0.268. A higher priority with weighted sum equation quantifies the urgency in the service needs of user 2 having relatively higher packet delay and lower channel quality. Therefore, the system is more sensitive to the HoL delay. If the instantaneous channel quality of the user improves, the system exploits it. For instance, consider Figure 5 where the channel quality increases at the *current scheduling instant*; the result is a higher time domain priority, quantifying lower time-averaged channel quality over a window of size n_c epochs and higher HoL packet delay. Because of the increase in the channel quality at the current scheduling instant, the frequency domain priority (function of current instantaneous channel quality) also increases with PRBs having

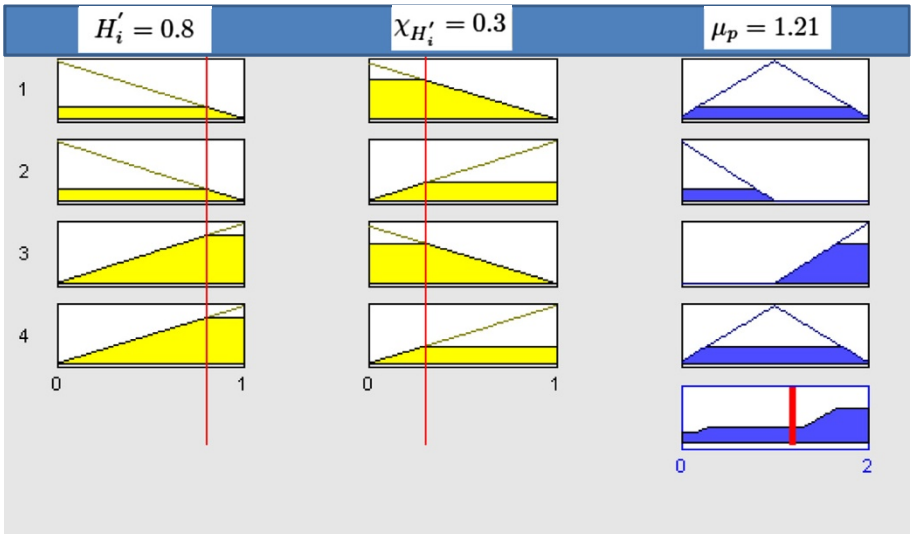

Figure 4 Fuzzy rules of the proactive controller.

better channel quality. Therefore, the weighted sum of the normalized HoL delay and the time-averaged channel quality with weights equal to 0.5 makes the system opportunistic (exploiting instantaneous channel improvements) and delay aware.

3.2 Reactive controller

Delay-sensitive applications can tolerate packet losses if they are below a given threshold. To provide fairness in multimedia traffic, packet losses should be kept below a given threshold for all users. The goal of the reactive controller is to distribute the packet losses proportionally equal across all the users. In order to define the inputs of the reactive controller, we utilize the packet loss ratio, $\text{plr}_i^{(n)}$, of user i given in Equation 3. The packet loss ratio can easily be calculated by using the number of dropped and transmitted packets over a small transmission window. The design of the reactive controller is

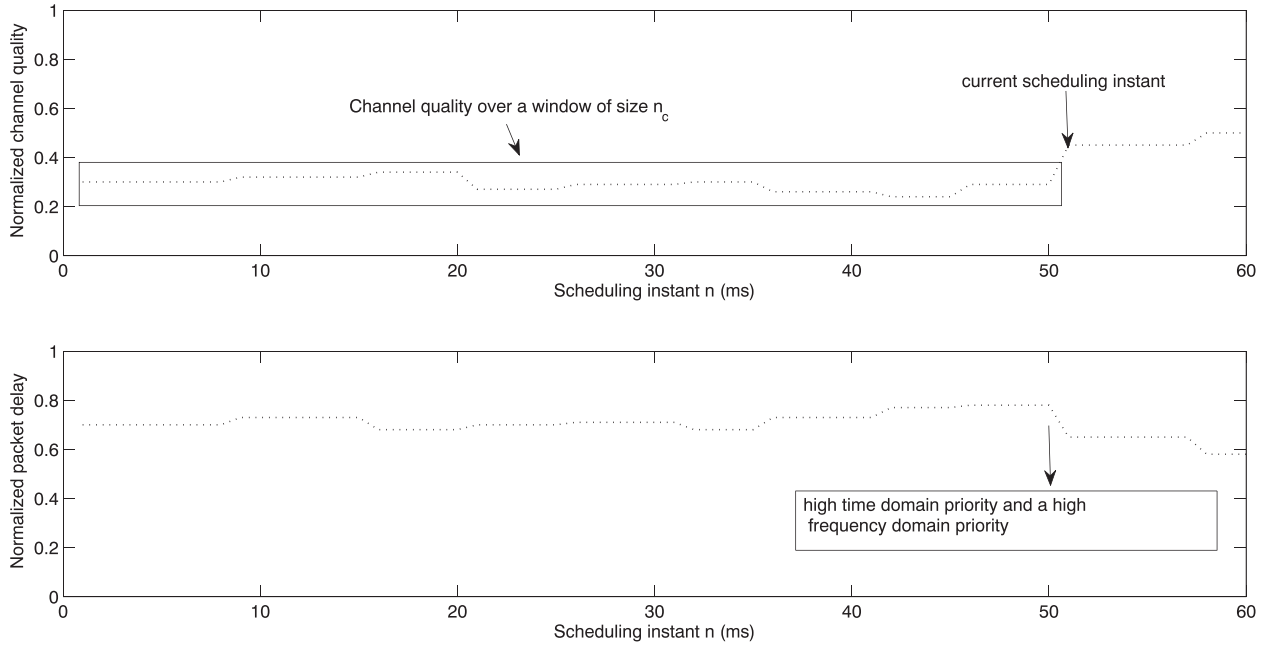

Figure 5 Rationale of the proactive controller design.

similar to the proactive controller except that the fuzzy inputs are based on the packet loss rate over a moving average transmission window. We consider a window size of 1 s in the simulation study. The amount of QoS violation in terms of packet loss ratio and tolerated packet loss threshold, plr_{thr}, of user i at scheduling instant n is:

$$V_{i,\text{delay}-\text{sensitive}}^{(n)} = \frac{\text{plr}_i^{(n)}}{\text{plr}_{i,\text{thr}}}. \tag{8}$$

The QoS parameter for the delay-sensitive traffic is the packet loss ratio, whereas for the best-effort flows, the QoS performance parameter is the ratio of minimum rate required to the achieved time-averaged throughput.

$$V_{i,\text{best}-\text{effort}}^{(n)} = \frac{R_{\min}}{R_{i,\text{ave}}^{(n)}} \tag{9}$$

where $R_{i,\text{ave}}^{(n)}$ is the time-averaged throughput and $R_{\min}$ is the minimum rate requirement. The QoS violation input, V_i', for the reactive controller is:

$$V_i' = \begin{cases} \frac{V_{i,\text{VoIP}}^{(n)}}{V_{j,\max}^{(n)}}, & \text{for VoIP traffic class} \\ \frac{V_{i,\text{Video}}^{(n)}}{V_{j,\max}^{(n)}}, & \text{for Video traffic class} \\ \frac{V_{i,\text{best}-\text{effort}}^{(n)}}{V_{j,\max}^{(n)}}, & \text{for best} - \text{effort traffic class} \end{cases} \tag{10}$$

It is a requirement of the fuzzy logic system that the inputs of the fuzzy controller should lie within the input fuzzy set, i.e., in between 0 and 1. Therefore, we normalize the input with respect to the flow having the maximum QoS violation, $V_{j,\max}^{(n)}$.

The second input, $\chi_{V_i'}$, of the reactive controller is designed as the weighted sum of the normalized QoS violations and the normalized average channel quality. Mathematically, it is defined as:

$$\chi_{V_i'} = 0.5(1 - V_i') + 0.5(\overline{\chi}_i^{(n)}). \tag{11}$$

3.2.1 Rationale

The rationale behind the design of the reactive controller is the same as that of the proactive controller discussed in Section 3.1.1. The weighted sum of the normalized QoS violations and the time-averaged channel quality with weights equal to 0.5 makes the system opportunistic (exploiting instantaneous channel improvements) and QoS aware as discussed in Section 3.1.1. The input and output membership functions and the output fuzzy set is the same as that of the proactive controller. It is important to note that we could have used all the inputs, i.e., the HoL packet delay, the QoS violations, and the time-averaged channel quality, and design a fuzzy priority scheme by defining a set of rules for these three inputs. However, this increases the complexity of the system because, with three inputs, eight rules and more than three output membership functions are required. A fuzzy logic system with two inputs is simpler in terms of implementation and processing. Therefore, by using the same rules and membership functions, the same fuzzy module is called for proactive (H_i' and $\chi_{H_i'}$) and reactive (V_i' and $\chi_{V_i'}$) inputs.

3.3 Dynamic resource controller

The best-effort traffic class is considered as the lowest priority class. Scheduling rules designed for delay-sensitive traffic, such as in [24,25] (see the time utility functions of different traffic classes), give low scheduling priority to the best-effort flows. High priority differentiation between the delay-sensitive and best-effort flows causes resource starvation for the best-effort flows [26,27]. In FCS scheduling framework, inter-class traffic priority differentiation is provided by output fuzzy set. The output fuzzy set represents the range of all possible output values that can be assigned to the proactive and reactive controllers. The larger the output fuzzy set, the higher the priority of the controller. In order to dynamically prioritize flows belonging to best-effort traffic class, we adapt the output fuzzy set of the best-effort flows according to the QoS performance of the delay-sensitive flows. The output fuzzy set of the delay-sensitive traffic class is fixed; thus, the amount of resource allocations between the delay-sensitive and best-effort flows is adaptable and controlled by the maximum limit of the output fuzzy set, μ_{max}, as given in Equation 12.

$$\mu_{r_{\text{best}-\text{effort}}} = \mu_{p_{\text{best}-\text{effort}}} \in \{0, \mu_{\max}\} \tag{12}$$

where $\mu_{r_{\text{best}-\text{effort}}}$ and $\mu_{p_{\text{best}-\text{effort}}}$ are the defuzzified outputs of the reactive and proactive controllers, respectively. As discussed in Sections 3.1 and 3.2, the design of both the controllers and the corresponding output fuzzy sets are the same. Flows from each traffic class utilize the same time domain priority by using the reactive and proactive controllers. The average delay and packet loss rate performance of the delay-sensitive flows are used to determine the maximum limit of the output fuzzy set for the best-effort traffic. Mathematically, the average

QoS parameters of the delay-sensitive flows are as follows:

$$\overline{H_i'} = \frac{1}{I_{\text{delay-sensitive}}} \sum_{i=1}^{I_{\text{delay-sensitive}}} H'_{i,\text{delay-sensitive}} \tag{13}$$

$$\overline{V_i'} = \frac{1}{I_{\text{delay-sensitive}}} \sum_{i=1}^{I_{\text{delay-sensitive}}} V'_{i,\text{delay-sensitive}} \tag{14}$$

where $I_{\text{delay-sensitive}}$ is the number of delay-sensitive users, $\overline{H_i'}$ is the average normalized delay, and $\overline{V_i'}$ is the average QoS violations of all the delay-sensitive users. The input and output membership functions of the DRC are shown in Figure 6a and 6b, respectively. The maximum limit, $\mu_{\max}$ is set according to the following fuzzy rules:

1. If $\overline{H_i'}$ is low AND $\overline{V_i'}$ is low THEN $\mu_{\max}$ is high
2. If $\overline{H_i'}$ is high AND $\overline{V_i'}$ is low THEN $\mu_{\max}$ is low
3. If $\overline{H_i'}$ is high AND $\overline{V_i'}$ is high THEN $\mu_{\max}$ is low
4. If $\overline{H_i'}$ is low AND $\overline{V_i'}$ is high THEN $\mu_{\max}$ is medium

The input degree of membership is determined by the trapezoidal input membership functions. A lower average packet delay and loss rate causes rule 1 to have a higher degree of membership. Therefore, $\mu_{\max}$ is maximum as given by the centroid of the highest area triangle membership function as shown in Figure 6. On the other hand, $\mu_{\max}$ is set to minimum when a higher average HoL delay and packet loss rate causes the smallest area triangle to be defuzzified through rule 2 and rule 3. If the normalized average delay is lower and average PLR is higher than the medium area, triangle is defuzzified as given in rule 4.

3.3.1 *Rationale*

The main rationale of utilizing DRC is to serve the following three goals:

- Utilization of delay tolerant nature of the best-effort traffic: according to the policy guidelines of the QoS architecture in the 3GPP standard, the resource allocation probability of the best-effort traffic class should be minimum in situations where the network becomes congested with delay-sensitive traffic. When the traffic load reaches the network capacity, the increase in average packet's latency of the delay-sensitive traffic decreases the maximum limit of the output fuzzy set for the best-effort flows as shown in Figure 7. Since best-effort traffic is delay tolerant, the decreased maximum limit of the output fuzzy set ensures delay-sensitive traffic gets priority over best-effort traffic.
- Channel diversity exploitation: the main goal of the scheduler is to maximize the system throughput

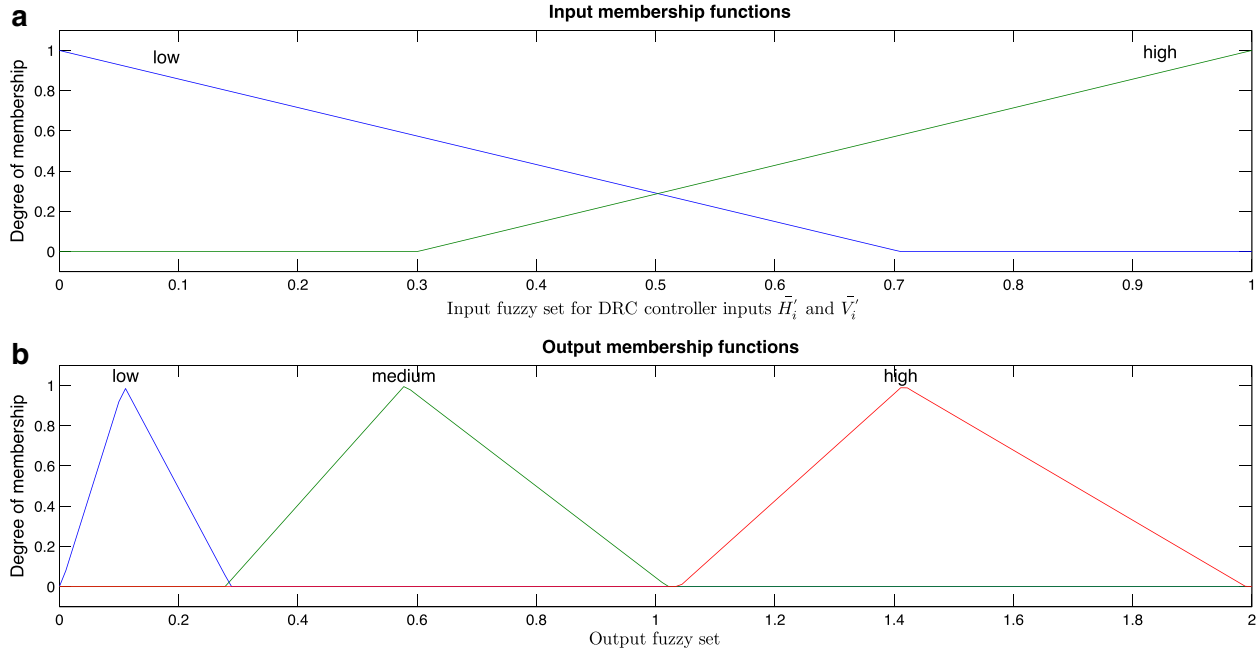

Figure 6 Input and output membership functions of the DRC. (a) Input membership functions low and high of the DRC. **(b)** Output membership functions low, medium, and high of the DRC.

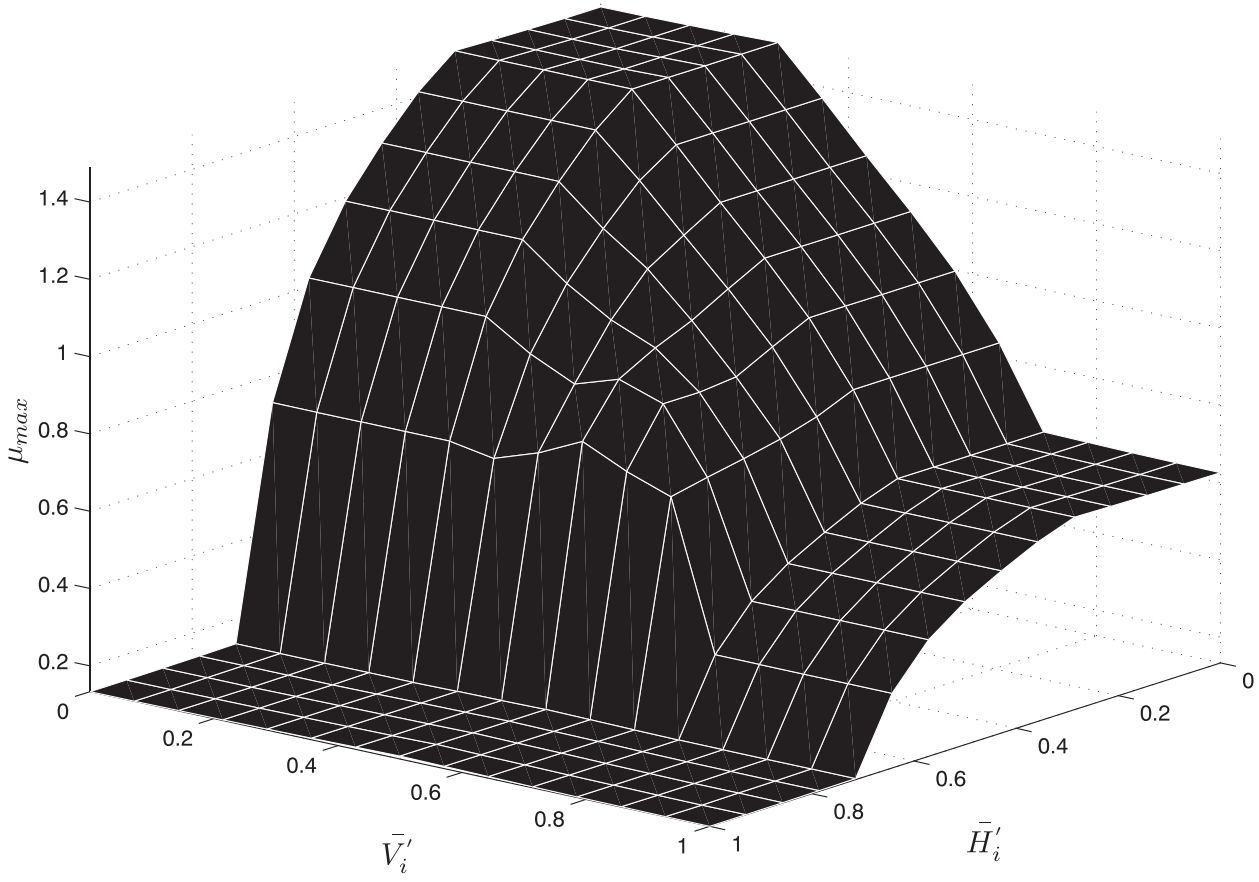

Figure 7 **DRC's response to the normalized average delay and average PLR.**

subject to maintaining the deadline violations below the prescribed threshold (Equation 5). At lower normalized average packet latency, the priority difference between the delay-sensitive and best-effort flows is minimal. Hence, flows from different traffic classes are scheduled based on their QoS performance and channel quality.

- Utilization of same output fuzzy set for the DRC, proactive, and reactive controllers: the prioritization of the delay-sensitive flows w.r.t the best-effort traffic can be achieved by using the same output fuzzy set for the proactive, reactive, and DRC controllers. When the output fuzzy set of these controllers are same, then the increase in latency of the delay-sensitive flows causes a reduction in the output fuzzy set of the best-effort traffic as shown in Figure 7. When the network becomes heavily congested, then delay bound violations occur for the delay-sensitive flows. The delay bound violation further reduces the output fuzzy set of the best-effort traffic as shown in Figure 7. Thus, decreasing the resource allocation probability of the best-effort traffic.

3.4 Time domain priority

The proactive controller output, μ_p, and the reactive controller output, μ_r, define the time domain priority of the scheduling rule. Let $\mu_i^{(n)}$ be the final time domain priority which is the product of the output of both the controllers given as:

$$\mu_i^{(n)} = (\mu_p \mu_r)^{\alpha_t} \tag{15}$$

where α_t is the time domain fairness parameter which enables the operator of the system to tune the fairness level. The higher the value of α_t, the higher will be the time domain priority of users suffering from relatively poor channel quality, higher HoL delay, and higher QoS violations.

3.5 Frequency domain priority

The time domain priority, by utilizing past and current CQI feedbacks, considers the channel quality over a small window. The goal of the time domain priority is to control the fairness among the users. On the other hand, the goal of the frequency domain priority is to improve the system

efficiency by considering only the current CQI feedback. Due to multipath propagation and interference from the neighboring users, there is a variable amount of fading on the PRBs of each user. Efficiency as well as fairness can be enhanced if this information is utilized. By employing the CQI feedbacks on each of the PRBs, information on the interference and multipath propagation can be obtained [28,29].

Hence, we adopt a parameter called relative strength of user i on PRB φ which is given as:

$$\theta_{i,\varphi}^{(n)} = \frac{\chi_{i,\varphi}^{(n)}}{\chi_i^{(n)}} \tag{16}$$

where $\theta_{i,\varphi}^{(n)}$ gives information on the variable amount of fading on the PRBs of each user. If a user is experiencing a high interference on some of the PRBs, this factor assigns a lower weight to such PRBs. On the other hand, the PRBs with the best channel quality for a user will be assigned a higher weight thus exploiting the independent multi-user frequency selective fading. The frequency domain priority, $\Gamma_{i,\varphi}^{(n)}$, of user i on PRB φ is the product of channel quality and relative strength:

$$\Gamma_{i,\varphi}^{(n)} = \chi_{i,\varphi}^{(n)} \theta_{i,\varphi}^{(n)}. \tag{17}$$

Replacing in (17), the expression of $\theta_{i,\varphi}^{(n)}$ given in (16), we get

$$\Gamma_{i,\varphi}^{(n)} = \frac{[\,\chi_{i,\varphi}^{(n)}]^2}{\chi_i^{(n)}}. \tag{18}$$

3.6 Final scheduling priority metric

It has been shown in [15] that a good trade-off between fairness and efficiency can be achieved by defining a priority function which is the product of the logarithmic function of the time-domain priority and a linear function of the instantaneous rate on each PRB. The time-domain priority used in the LOG rule in [15] is a function of the HoL packet delay. We use a time-domain priority which is derived from fuzzy logic and is a function of the user's HoL packet delay, time-averaged channel quality and packet loss rate. The final priority, $\mathrm{PRF}_{i,\varphi}^{(n)}$, of user i on PRB φ is a function of the logarithm of the time domain priority and it varies linearly with the frequency domain priority as given below:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \log(1+\mu_i^{(n)})\ \Gamma_{i,\varphi}^{(n)}. \tag{19}$$

User i^* is allocated a PRB φ satisfying the following rule:

$$i^* = \mathrm{argmax}\left(\mathrm{PRF}_{i,\varphi}^{(n)}\right). \tag{20}$$

It is important to note that state-of-the-art scheduling rules serve best-effort flows with the classical delay-insensitive PF rule and prioritize the delay-sensitive traffic by considering the HoL packet delay. We use the same priority equation given in 19 for all the traffic classes; dynamic prioritization between the delay-sensitive and best-effort classes is achieved by using the DRC. More details on the prioritization of different traffic classes is given in the following sections.

4 Performance evaluation

4.1 Benchmark scheduling rules

In order to assess the performance of the proposed FCS scheduling rule, we compare it the with the state-of-the-art strategies shown in Table 1.

According to the table, γ_i is a constant whose value is adjusted to account for different delay requirements for different flows. $N_{Q_i}^{(n)}$ is the number of packets residing in the queue of user i's flow at the eNodeB. In order to provide fairness, the delay-based rules (M-LWDF, EXP-PF, LOG-RULE, and EXP-RULE) use the HoL delay, whereas the queue-based rules (M-LWDFQ, EXP-PFQ) utilize the queue size of each flow. In the log (LOG-RULE) and exponential (EXP-RULE) rules, $b_i = \frac{1}{R_{i,\mathrm{ave}}^{(n)}}$ and a_i is a tunable parameter. The higher the value of a_i, the higher will be the priority of the delay-sensitive flows. All the priority rules shown in Table 1 are for delay-sensitive traffic. These rules calculate the priority for the best-effort traffic according to the PF rule given as:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}} \tag{21}$$

Table 1 Benchmark scheduling rules for delay-sensitive traffic

Strategy	Priority function
M-LWDF [12]	$\gamma_i \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}}\ H_i^{(n)}$
M-LWDFQ [39]	$\gamma_i \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}}\ N_{Q_i}^{(n)}$
EXP-PF [12]	$\gamma_i \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}}\ \exp\left(\frac{\gamma_i H_i^{(n)} - \overline{\gamma_i H_i}}{1+\sqrt{\overline{\gamma_i H_i}}}\right)$
EXP-PFQ [39]	$\gamma_i \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}}\ \exp\left(\frac{\gamma_i N_{Q_i}^{(n)} - \overline{\gamma_i N_{Q_i}^{(n)}}}{1+\sqrt{\overline{\gamma_i N_{Q_i}^{(n)}}}}\right)$
LOG-RULE [15]	$b_i\ \log\left[1.1+\left(a_i \frac{H_i^{(n)}}{H_{i,\max}^{(n)}}\right)\right]\ \chi_{i,\varphi}^{(n)}$
EXP-RULE [15]	$b_i * exp\left[\frac{\left(a_i \frac{H_i^{(n)}}{H_{i,\max}^{(n)}}\right)}{1+\sqrt{\overline{H_i}}}\right]\ \chi_{i,\varphi}^{(n)}$

where time-averaged throughput $R_{i,\text{ave}}^{(n)}$ is defined as:

$$R_{i,\text{ave}}^{(n)} = R_{i,\text{ave}}^{(n-1)} * (1 - \frac{1}{n_w}) + \frac{1}{n_w} * R_i^{(n-1)}. \quad (22)$$

where $R_{i,\text{ave}}^{(n-1)}$ is the average throughput at scheduling instant $n-1$. $R_i^{(n-1)}$ is the number of bits transmitted at scheduling instant $n-1$. n_w is the size of the time-average window also known as the exponential averaging constant. The higher the size of the time-average window, the higher the impact of the instantaneous channel quality.

4.2 Simulation scenario

In order to investigate the performance of the proposed scheduling algorithm, a link-level simulator [30,31] built on MATLAB's object-oriented features is selected as the simulation platform with all the basic features of an LTE link layer. The fuzzy controllers are designed by utilizing the Matlab's Fuzzy Logic Toolbox. The fuzzy logic priority scheduling, under varying load, has been studied in [21], where only video traffic was considered. In this work, delay-sensitive traffic is characterized by video and VoIP flows. In order to simulate best-effort traffic, CBR flows with the data rate of 400 Kbps are selected. On the other hand, 64 Kbps traffic with the threshold packet loss rate of 1% and maximum delay budget of 100 ms is selected for VoIP users. These QoS parameters are selected according to LTE QCI (QoS class indicators) [32]. Video traffic is generated from a trace file [33] with the average and peak traffic rates of 530 and 1,500 kbps, respectively. The maximum delay budget for video packets is 200 ms whereas the threshold packet loss rate is 5%. For non-critical video applications, 5% packet loss rate corresponds to a peak signal-to-noise ratio (PSNR) of approximately 29 to 30 dB [34]. Therefore, 5% is considered as the threshold packet loss rate for an acceptable video quality. Table 2 reports the simulation parameters adopted for the LTE system and the wireless channel. We simulate 18 video flows (9.54 Mbps), 27 VoIP flows (1.728 Mbps), and 9 best-effort flows (3.6 Mbps) corresponding to a total average input traffic rate of 14.868 Mbps. The main motivations for such traffic distribution are the following:

- It has been reported in [35] that by 2015, approximately 66% of mobile's traffic (in terms of petabytes per month) will be video and the proportion of VoIP traffic will be a minority. Therefore, the proportion of traffic in our simulation scenario is dominated by video followed by the best-effort and VoIP traffic. Specifically, we selected a loaded network with 64% video, 11% VoIP, and 25% best-effort traffic (in terms of average input traffic at the eNodeB).
- The proposed scenario corresponds to an average input traffic rate of 14.868 Mbps. In order to evaluate the channel utilization in terms of average spectral efficiency, we simulate an optimum sum rate maximization strategy. The optimum strategy maximizes the system throughput without considering the delay constraints. The average channel quality (in terms of SINR) of the users is set such that the total system throughput, sum throughput of all the flows, produced by the throughput maximization strategy [36] is 13.6 Mbps (2.72 bits/s/Hz). This corresponds to a heavily loaded system where the input traffic is approximately 110%, in terms of bits/s/Hz, of the maximum system capacity. Our main goal is to study the fairness and efficiency performance of the proposed and benchmark scheduling rules when the delay bound and packet loss threshold constraints are considered.

Table 2 Simulation parameters - downlink LTE scheduling for delay-sensitive and best-effort traffic

Parameters	Value
Simulator	LTE link level simulator [30,31]
Bandwidth, carrier frequency	5 MHz, 2.1 GHz
UE distribution, cell radius	Uniform, 1 km
Channel	3GPP-TU (typical urban)
Pathloss model	Hata-Cost-231 model
Shadowing model	Log-normal shadow fading
HARQ	Up to 3 synchronous retransmissions
Channel fading	Block fading (1 ms)
UE speed	15 to 100 km/h (users moving independently at variable speed)
CQI averaging method	Mutual information effective SNR mapping
	(MIESM) [37]
H_{max}, PLR_{thr} (video)	200 ms, 5% [34,38]
H_{max}, PLR_{thr} (VoIP)	100 ms, 1% [32]
H_{max}, R_{min} (best-effort)	300 ms, 200 Kbps [32]
Number of video, VoIP and best-effort users	18, 27, and 9
Average rate requirements for video,	530, 64, and 400 Kbps
VoIP and best-effort users	
n_c (Time-averaged channel quality window)	100 ms
and n_w (Time-averaged throughput window)	

We consider the time-averaged channel quality over the period, $n_c = 100$ ms. All the benchmark scheduling rules utilize the time-averaged throughput. In order to have a fair comparison, the exponential averaging constant n_w is set to 100 ms. In the literature, the optimum size of the

Table 3 Tunable parameters for FCS strategy

Strategy	Output fuzzy set for video	Output fuzzy set for VoIP	α_t
FCS1	{0, 2}	{0, 2}	2
FCS2	{0, 2}	{0, 2.2}	2
FCS3	{0, 2}	{0, 2.2}	3

exponential averaging constant is from 100 to 1,000, with the 100 being utilized in scenarios yielding high fairness in terms of throughput.

The FCS scheduling strategy has the following tunable parameters:

- The time domain fairness parameter α_t mainly used to adjust the fairness level.
- The output fuzzy set for the DRC, proactive, and reactive controllers. As discussed in Section 3.3.1, the same output fuzzy set is utilized for all the controllers. The membership functions and fuzzy rules of the DRC are set such that by utilizing the same output fuzzy set for all the controllers, dynamic prioritization is achieved between the delay-sensitive and best-effort traffic.

Table 3 reports the settings considered in the simulations for the tunable parameters. The table reports three examples. In the first one (FCS1), the output fuzzy set is the same for both VoIP and video classes, i.e., video and VoIP traffic flows have the same prioritization. It is important to note that the LTE QoS architecture specifies QCI for each of the considered traffic classes [32]. According to the QCI, the VoIP traffic class has the highest priority followed by the video and best-effort traffic class. Therefore, in the second case (FCS2), VoIP is prioritized by increasing the maximum limit of the output fuzzy set from 2 to 2.2. The time domain parameter α_t is set to 2. Finally, in the third case, we retain the VoIP priority and increase the time-domain parameter from 2 to 3. The FCS2 and FCS3 modes follow the QCI service architecture where VoIP is prioritized over video traffic, while the FCS1 mode gives same priority to both the delay-sensitive traffic classes.

4.3 Results and discussion

First, we analyze the fairness and efficiency performance of the FCS strategy according to the settings reported in Table 3. Next, we compare the FCS strategy with the benchmark scheduling strategies reported in Section 4.1.

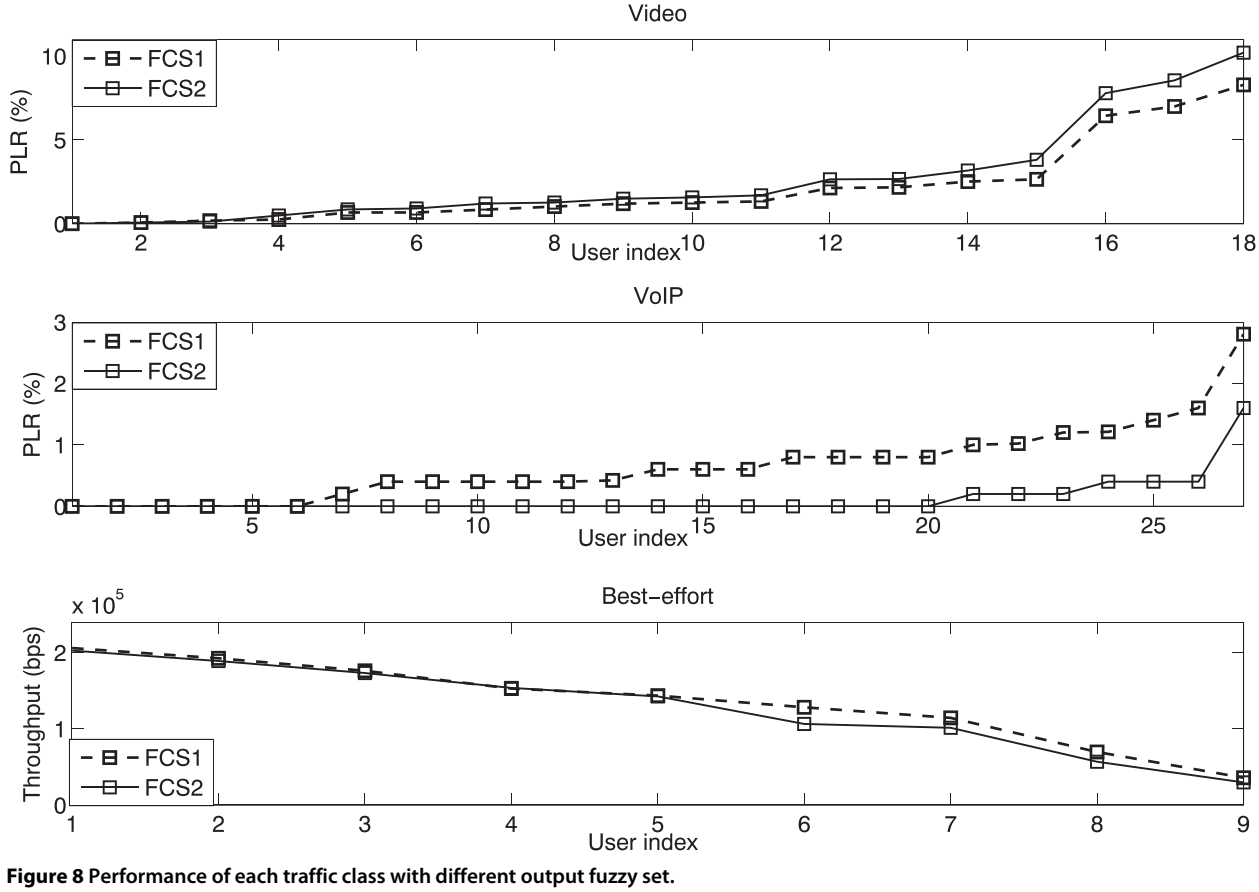

Figure 8 Performance of each traffic class with different output fuzzy set.

Results in terms of packet loss rate (delay-sensitive flows) and throughput (best-effort flows) for the proposed rule with different parameters are shown in Figures 8 and 9, where users are arranged in a decreasing order of the channel quality. The user with lowest index has the best channel quality which then decreases with the increase in user index.

Using the same prioritization (output fuzzy set) for video and VoIP (FCS1, Figure 8) results in a higher QoS violations for the VoIP flows, i.e., seven VoIP flows are violating the 1% PLR threshold. On the other hand, three video flows are violating the 5% PLR threshold.

In the second set of simulations (FCS2, Figure 8), there is a significant reduction in the PLR of the VoIP flows, i.e., only one VoIP flow has a delay bound violation (PLR) of more than 1%. In FCS2 mode, the impact of the time-domain priority is higher for the VoIP flows. The increase in the HoL delay and PLR prioritizes VoIP flows more than the video flows. The result is an increase in the PLR of the video flows as shown in Figure 8. There is also a slight reduction in the throughput of the best-effort flows. Higher limit of the fuzzy set, 2.2, for VoIP traffic serves well according to the QoS architecture of LTE as it is the highest traffic priority class. Any further increase in the maximum limit of the fuzzy set for VoIP traffic will penalize the video and best-effort flows by serving VoIP traffic.

Next, we analyze the impact of the time-domain parameter α_t. An increase in the time-domain priority parameter (FCS3, Figure 9) allocates relatively more resources to the worst channel flows since time-domain priority is a fuzzy function of the HoL packet delay, PLR, and time-averaged channel quality. It is important to note that the proportion of video traffic (18 flows with average rate requirements of 540 kbps) is high with respect to the VoIP traffic (27 flows with rate requirements of 64 kbps). Therefore, an increase in α_t results in a significant improvement for video flows as shown in Figure 9. In other words, more resources are allocated to the lower channel quality video flows and as a result, their PLR is reduced at the expense of a slight increase in the PLR of VoIP flows. There is also a marginal increase in the PLR of good channel video flows. According to the figure (FCS3, Figure 9), the worst served video flow has a PLR of approximately

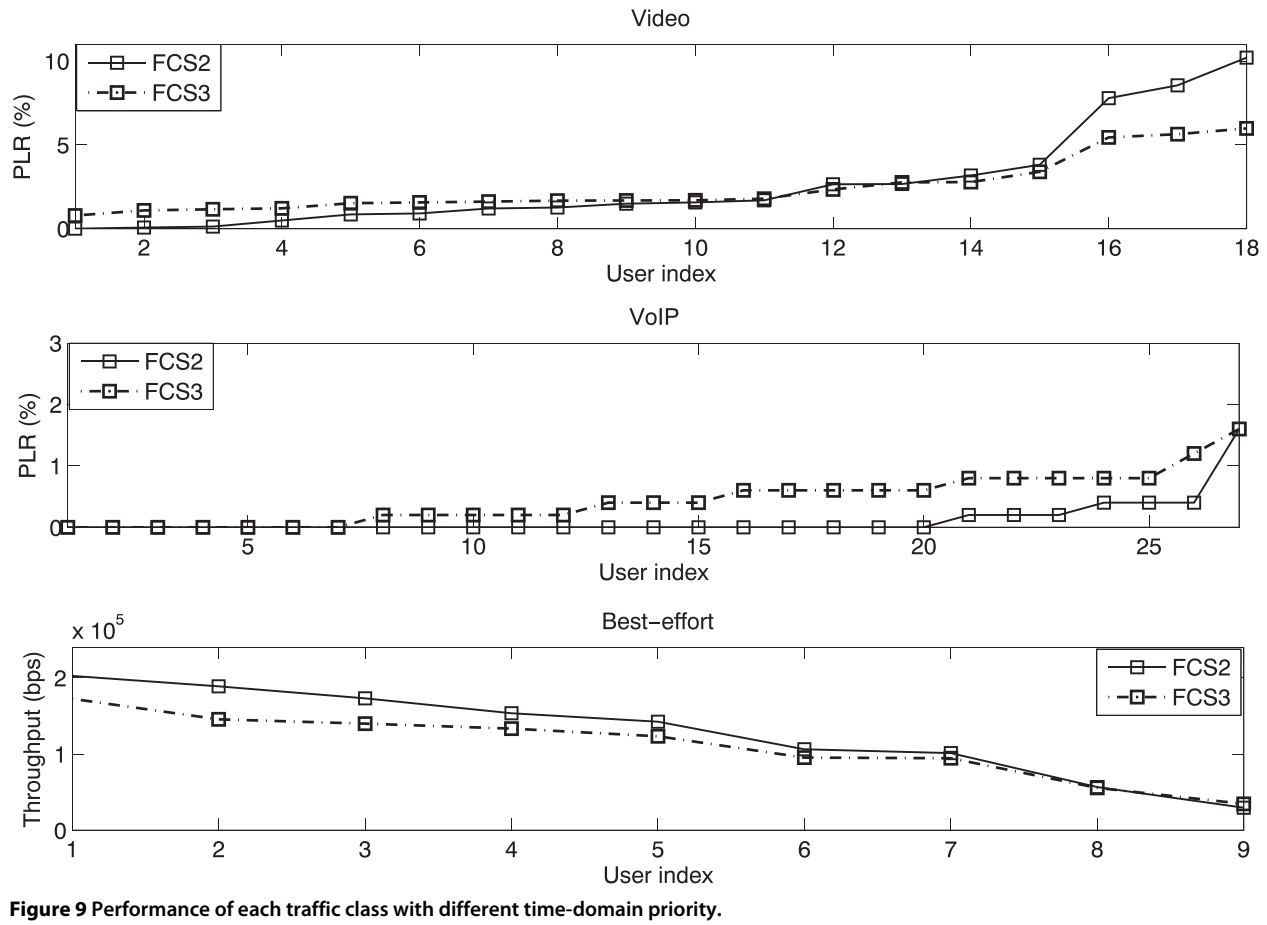

Figure 9 Performance of each traffic class with different time-domain priority.

5.4% and the worst served VoIP flow suffers from a PLR of 1.6%. Thus, the FCS3 mode results in an improved fairness performance for the delay-sensitive flows. Under high load, the time-domain priority of the delay-sensitive flows will always be higher than best-effort flows. Therefore, increase in α_t will further enhance the priority difference and results in a reduction in the throughput of best-effort flows.

After analyzing the performance of the FCS rule for different design parameters, we compare the FCS rule with the well-known scheduling rules. First, we discuss the performance of state-of-the-art scheduling rules for delay-sensitive traffic. Figures 10 and 11 analyze the performance of state-of-the-art scheduling rules for delay-sensitive traffic and compare it with the FCS rule. Although M-LWDF is generally considered as the best scheduling rule for delay-sensitive traffic [12], the PLR performance of the M-LWDF scheduling rule for low channel quality video flows is very poor. According to Figure 10, the PLR of the worst served user is as high as 20% and approximately seven flows suffer from the QoS violation, i.e., having PLR above the 5% threshold. Higher QoS violations for video flows stem from the fact that the M-LWDF rule exploits time diversity by considering the time-averaged throughput. The video flows exhibit variable bit rate (higher peak-to-average rate ratio) characteristics. Therefore, the higher time-averaged throughput in the scheduling decision of the M-LWDF rule increases the probability of delay bound violations for the video flows having lower channel quality. Hence, high rate delay-sensitive flows with lower channel quality suffer from HoL delay bound violations. On the other hand, none of the VoIP flows suffer from delay bound violations (Figure 11) mainly because lower time-averaged throughput prioritizes the VoIP flows irrespective of their channel quality. M-LWDFQ reduces the QoS violations of the video flows by considering the queue size based on virtual token mechanism [39]. The PLR of the worst served user is approximately 12%, and there are only three flows violating the PLR threshold of 5%. The improved performance for video flows is mainly due the fact that the M-LWDFQ rule prioritizes high rate flows by considering the number of packets in the queue based on virtual token mechanism, as compared to the M-LWDF rule which relies on the HoL delay. As a result, flows having fewer packets in the queue are penalized if their channel quality is low. Figure 11 shows that seven VoIP flows have PLR of more than 1%. Therefore, the M-LWDFQ rule increases the QoS violation for the VoIP flows as compared to the M-LWDF scheduling rule.

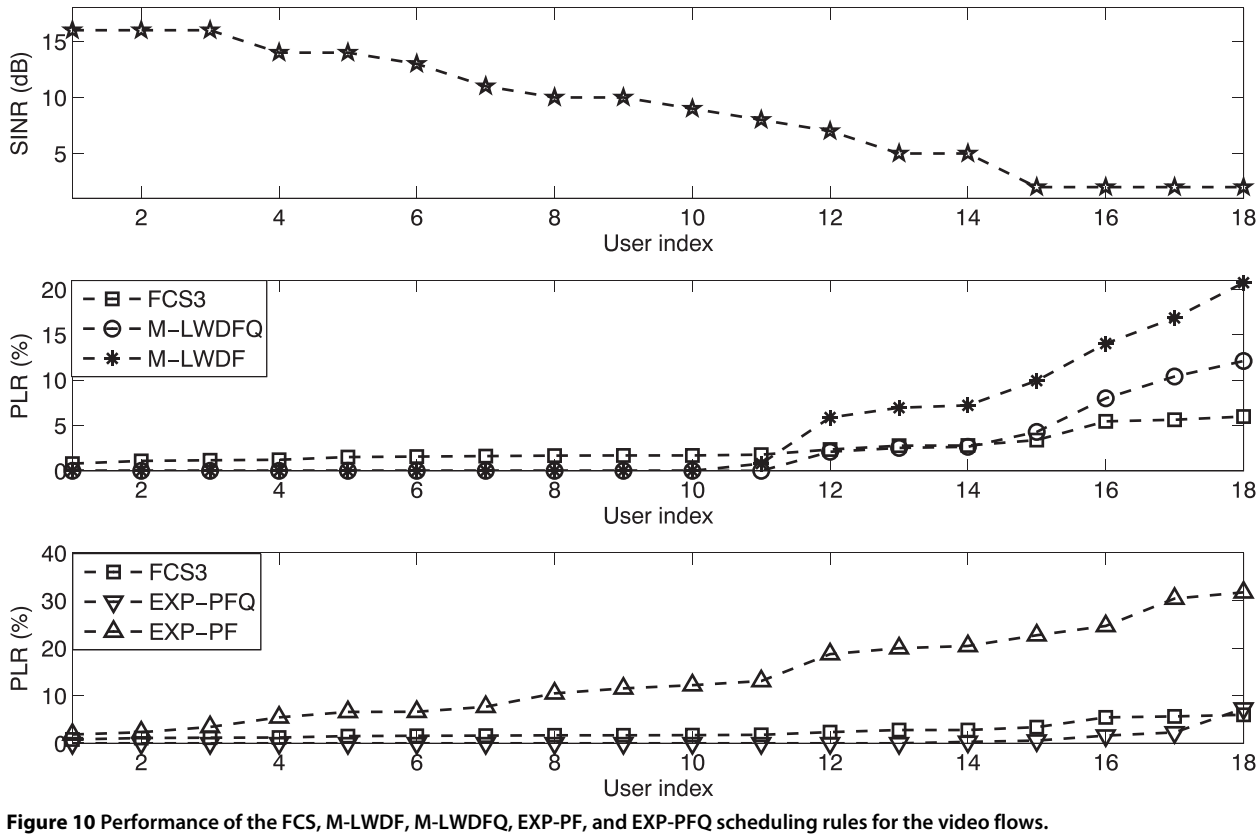

Figure 10 Performance of the FCS, M-LWDF, M-LWDFQ, EXP-PF, and EXP-PFQ scheduling rules for the video flows.

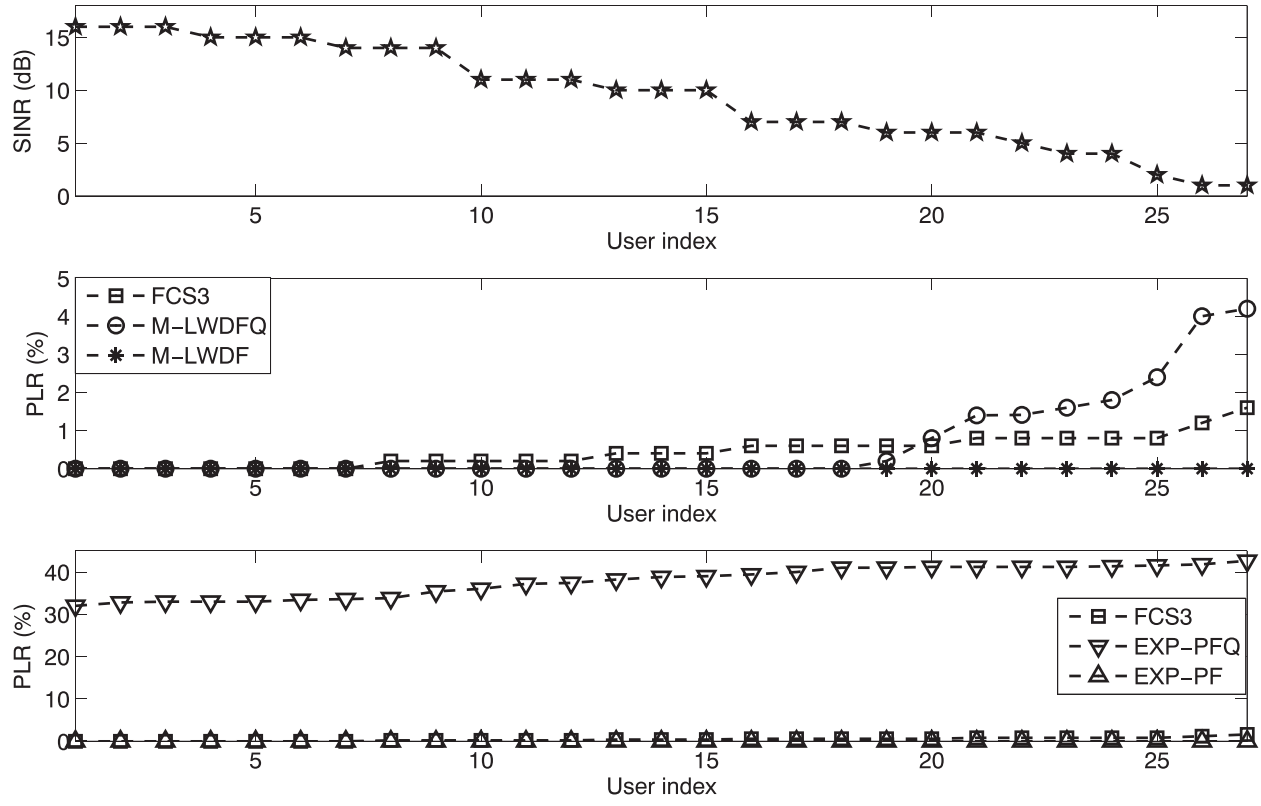

Figure 11 Performance of the FCS, M-LWDF, M-LWDFQ, EXP-PF, and EXP-PFQ scheduling rules for the VoIP flows.

When compared to the state-of-the-art scheduling rules, the FCS strategy improves the fairness performance for delay-sensitive flows mainly due to the fact that this scheduling rule considers the channel quality of a user in a novel way, by taking into account the past and current CQI feedbacks in the time domain priority metric. This allows the users with relatively low channel quality and high HoL delay to be prioritized in the time domain. As a result, the difference in the average waiting time of each flow's packet is low. On the other hand, state-of-the-art scheduling rules favor the good channel quality flows by serving them way before their packet's delay bound. These scheduling rules are highly unfair for the cell edge users as they require a substantial increase in the SINR of the cell edge users so that their packet's delay bound requirements are met. In the FCS scheduling strategy, the PLR over the moving average window is kept below the threshold for each of the delay-sensitive flows in the system. Therefore, this rule balance different flows' probabilities of QoS violations. It is important to note that the FCS strategy requires an admission controller to limit the arrival rate of delay-sensitive traffic within the achievable rate region. Since fairness is incorporated in the scheduling decisions, an increase in the arrival rate above the system capacity violates the QoS performance of the flows already being served.

The performance of the EXP-PF and EXP-PFQ scheduling rules for video and VoIP traffic classes is shown in Figures 10 and 11, respectively. For video flows, the EXP-PF scheduling rule increases the QoS violations significantly, i.e., approximately all the flows have delay violations of more than 5%. The performance of the EXP-PF rule for VoIP flows is the same as that of the M-LWDF rule. The M-LWDF and EXP-PF rule are delay-based schedulers. These scheduling rules prioritize VoIP flows mainly because of the lower rate requirements. The token-based version [39] of these scheduling rules penalizes the VoIP flows more because of the higher queue size of the video flows. EXP-PFQ performs worst for the VoIP flows mainly because the queue size of the VoIP flows always remains lower than the Video flows, which causes an exponential increase in the priority of the video flows. Therefore, all performance gain obtained by video flows penalizes to the VoIP and best-effort flows as shown in Figures 11 and 12. For best-effort flows (Figure 12), the performance of the EXP-PF rule is significantly better than other scheduling rules. The EXP-PFQ scheduling rule performs best for the video flows. However, it severely penalizes the VoIP flows.

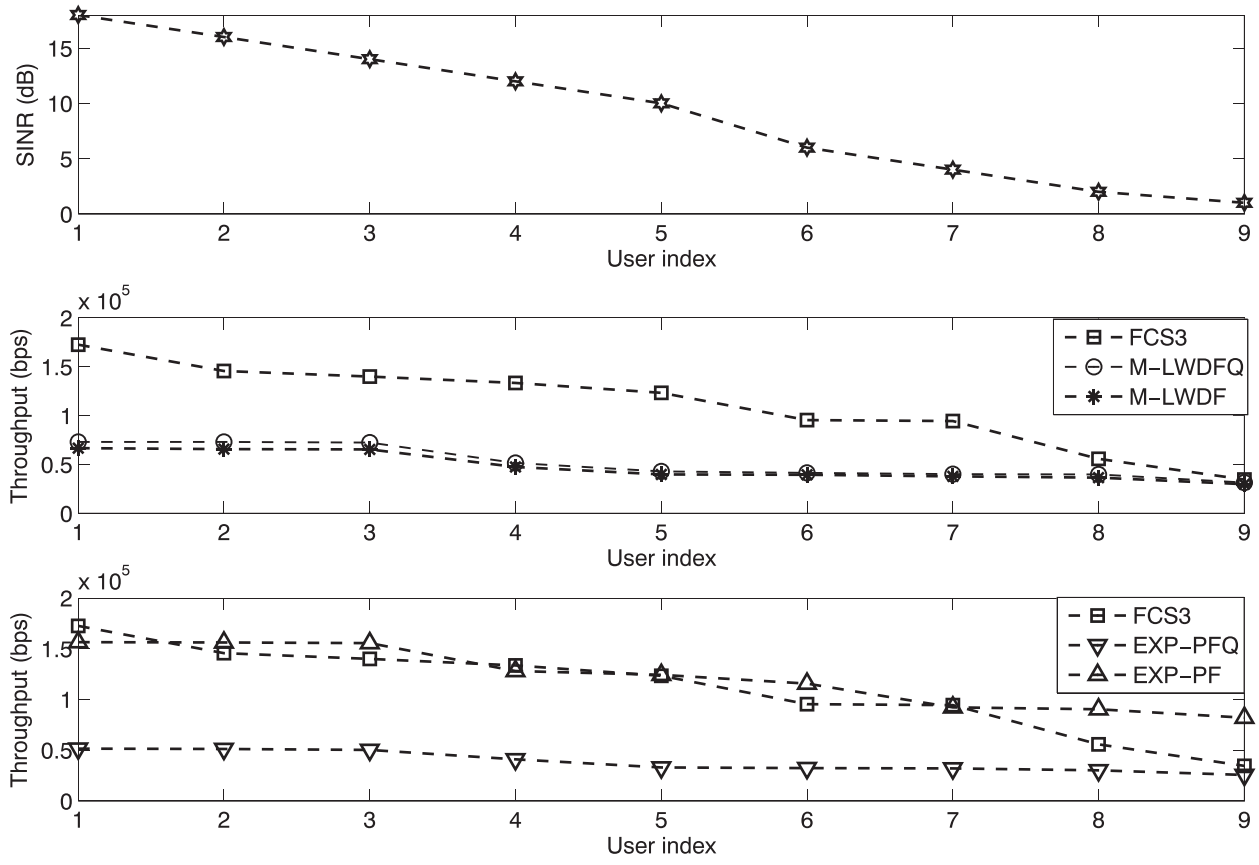

Figure 12 Performance of the FCS, M-LWDF, M-LWDFQ, EXP-PF, and EXP-PFQ scheduling rules for the best-effort flows.

All state-of-the-art scheduling rules prioritize best-effort flows by using the classical proportional fair Eq. 21. These rules prioritize delay-sensitive flows by using the linear, logarithmic, or exponential functions of the HoL delay as reported in Section 4.1. On the other hand, the FCS scheduling rule uses the same priority function for the best-effort and delay-sensitive flows, as given in Equation 19. The priority differentiation between the best-effort and delay-sensitive traffic classes is controlled by adapting the maximum limit of the output fuzzy set. The same priority function for each traffic class allows the exploitation of multi-user channel diversity across all the flows. This allows FCS rule to achieve intra-class and inter-class fairness which is not the case in state-of-the-art scheduling rules. The priority of the best-effort traffic class is dynamic and changes according to the QoS performance of the delay-sensitive flows.

The log-rule achieves the best performance for the best-effort flows but it is highly inefficient for the video flows as shown in Figure 13. It is important to note that the log rule uses the normalized HoL delay as the time-domain priority. The logarithmic variation of the normalized HoL delay has a marginal increase in the priority of the lower channel quality flows. For instance, if we set the tunable parameter a_i to 50 (log-rule equation in Table 1) and analyze the priority difference between the 3 and 15 dB flows, the good channel quality flow with a normalized averaged delay of 0.3 results in a time domain priority of $\log[1.1 + (50 \times 0.3)] = 2.78$. On the other hand, the poor channel quality flow having normalized HoL delay of 0.9 results in a time-domain priority of $\log[1.1 + (50 \times 0.9)] = 3.83$. The logarithmic function marginalizes the priority of the poor channel quality flow as the delay urgency does not proportionately increases its priority. It is evident from Figure 13 that the log rule increases the PLR of the lower channel quality flows. The figure also shows the performance of the Exp-rule. According to the figure, the Exp-rule achieves better performance than the log-rule but it is highly unfair for the best-effort flows. The exponential function of the normalized HoL delay caters the delay urgency of the delay-sensitive flows better than the log rule. It is important to note that the FCS strategy increases the PLR of the VoIP flows as compared to the linear, logarithmic, and exponential delay based scheduling rules. However, the VoIP traffic class is packet loss tolerant and can tolerate the

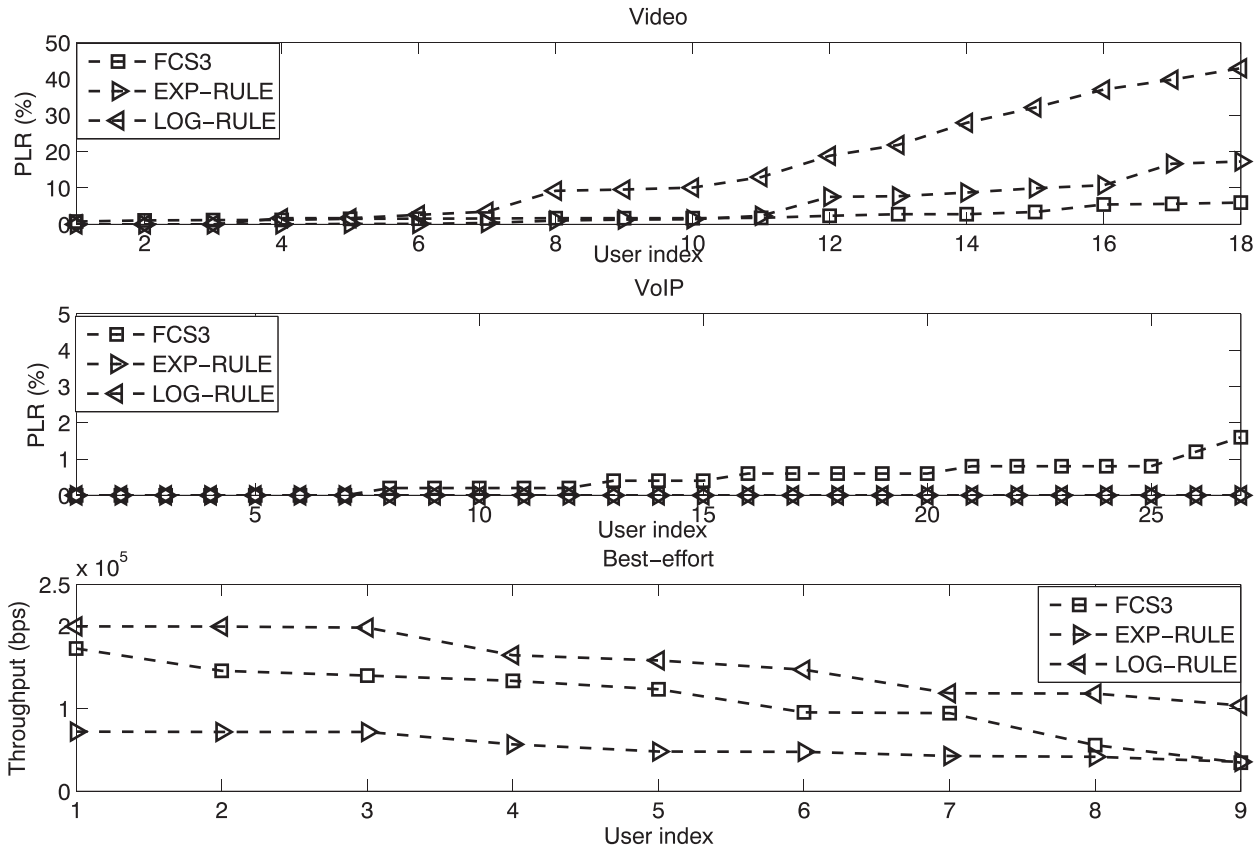

Figure 13 Performance comparison of the FCS strategy with the LOG and EXP scheduling rules.

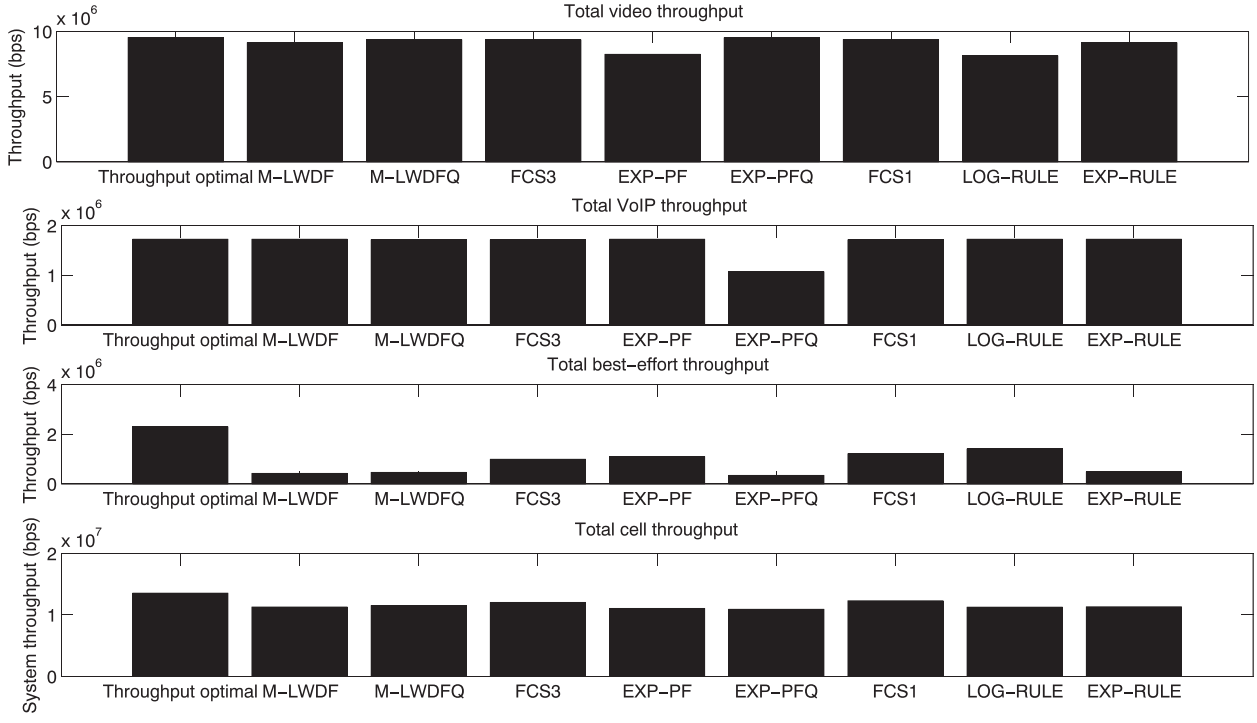

Figure 14 Throughput performance of all the considered scheduling rules.

PLR threshold of 1%. The FCS strategy marginally violates the packet loss threshold of two VoIP flows, i.e., the worst served VoIP flows have a PLR of 1.2% and 1.6%.

Figure 14 summarizes the performance of all the scheduling rules. The figure reports the throughput achieved by each of the traffic classes. The last sub-figure shows the total system throughput, which is simply the sum of the throughput achieved by each of the traffic classes. When compared to the optimum channel utilization strategy, the FCS scheduler compromises approximately 10.5% of the total cell throughput while providing fairness and QoS provisions. It is clear from the figure that among all the aforementioned QoS-aware scheduling rules, the FCS scheduling rule achieves the best inter-class fairness in terms of the throughput achieved by each traffic class.

5 Conclusions

We proposed a composite scheduling strategy for downlink scheduling at the MAC layer for delay-sensitive traffic in wireless systems based on OFDMA. This strategy uses novel concept of providing fairness using fuzzy logic membership functions and its rule base, instead of relying on the rate based proportional fair strategies employed in the literature. Furthermore, we provide a framework for service class differentiation among different traffic classes by utilizing the fuzzy logic priority scheme. Our approach leads to a framework which provides intra-class as well as inter-class fairness. The design of the scheduling rule is robust, and it serves well in diverse channel and rate requirements.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Rd, Kingston upon Thames, Surrey KT1 2EE,UK. [2]DOCOMO Euro-Labs, Landsberger Str. 312, 80687 Munich, Germany.

Received: 10 March 2013 Accepted: 13 July 2014
Published: 21 August 2014

References

1. S Sesia, I Toufik, M Baker, *LTE–The UMTS Long Term Evolution, From Theory to Practice*. (Wiley, New york, USA, 2009)
2. CY Wong, RS Cheng, KB Letaief, RD Murch, Multicarrier OFDM with adaptive subcarrier, bit, and power allocation. IEEE J. Selected Areas, Commun. **17**(10), 1747–1758 (1999)
3. J Jang, KB Lee, Transmit power adaptation for multiuser OFDM systems. IEEE J. Selected Areas, Commun. **21**(2), 171–178 (2003)
4. W Rhee, JM Cioffi, Increasing in capacity of multiuser OFDM system using dynamic subchannel allocation, in *IEEE Vehicular Technology Conference (VTC)* (California, USA, 2000)
5. J Jang, KB Lee, Adaptive resource allocation in multiuser OFDM systems with proportional fairness. IEEE Trans. Wireless Commun. **21**(2), 171–178 (2003)
6. J Gross, H Karl, F Fitzek, A Wolisz, Intl. Conf. on Wireless Networks (ICWN) (Las Vegas, USA, 2003)
7. J Gross, J Klaue, H Karl, A Wolisz, Subcarrier allocation for variable bit rate video streams in wireless OFDM systems, in *IEEE Vehicular Technology Conference (VTC)* (Florida, USA, 2003)
8. AKF Khattab, KMF Elsayed, Opportunistic subcarrier management for delay-sensitive traffic in OFDMA-based wireless multimedia networks, in *IST Mobile and Wireless Communications Summit* (Dresden, Germany, 2005)
9. A Jalali, R Padovani, R Pankaj, Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system, in *IEEE Vehicular Technology Conference (VTC)* (Tokyo, Japan, 2000)
10. M Andrews, K Kumaran, K Ramanan, A Stolyar, P Whiting, R Vijayakumar, Providing quality of service over a shared wireless link. IEEE Commun. Mag. **39**(2), 150–154 (2001)
11. JH Rhee, JM Holtzman, DK Kim, Performance analysis of the adaptive EXP/PF channel scheduler in an AMC/TDM system. IEEE Commun. Lett. **8**(8), 497–499 (2004)
12. HAM Ramli, R Basukala, K Sandrasegaran, R Patachaianand, Performance of well known packet scheduling algorithms in downlink 3GPP LTE system, in *IEEE Malaysia International Conference on Communications (MICC)* (Kuala Lumpur, Malaysia, 2009)
13. F Capozzi, G Piro, L Grieco, G Boggia, P Camarda, Downlink packet scheduling in LTE cellular networks: key design issues and a survey. IEEE Commun. Surv. Tutorials. **15**(2), 1–23 (2012)
14. N Khan, MG Martini, Z Bharucha, G Auer, Opportunistic packet loss fair scheduling for delay-sensitive applications over LTE systems, in *IEEE Wireless Communications and Networking Conference (WCNC)* (Paris, France, 2012)
15. B Sadiq, R Madan, A Sampath, Downlink scheduling for multiclass traffic in LTE. EURASIP J. Wireless Commun. Netw. **2009** (2009)
16. D Liu, YH Lee, An efficient scheduling discipline for packet switching networks using earliest deadline first round robin, in *International Conference on Computer Communications and Networks (ICCCN)* (Dallas, USA, 2003)
17. J Holtzman, Asymptotic analysis of proportional fair algorithm, in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)* (San Diego, USA, 2001)
18. G Fodor, A Furuskar, P Skillermark, J Yang, On the impact of uplink scheduling on intercell interference variation in MIMO OFDM systems, in *IEEE Wireless Communications and Networking Conference (WCNC)* (Budapest, Hungary, 2009)
19. K Beh, S Armour, A Doufexi, Joint time-frequency domain proportional fair scheduler with HARQ for 3GPP LTE systems, in *IEEE Vehicular Technology Conference (VTC)* (Calgary AB, Canada, 2008)
20. J Jantzen, Tutorial on fuzzy logic. Technical University of Denmark, Dept. of Automation, Tech. rep. 98-E-868 (1998)
21. N Khan, MG Martini, D Staehle, Opportunistic QoS-aware fair downlink scheduling for delay-sensitive applications using fuzzy reactive and proactive controllers, in *IEEE Vehicular Technology Conference (VTC)* (Las Vegas, USA, 2013)
22. S Shin, BH Ryu, Packet loss fair scheduling scheme for real-time traffic in OFDMA systems. ETRI J. **26**(5), 391–396 (2004)
23. C Koksal, H Kassab, H Balakrishnan, An analysis of short term fairness in wireless media access protocol, in *ACM SIGMETRICS* (California, USA, 2000)
24. RG Garroppo, S Giordano, D Iacono, L Tavanti, Game theory and time utility functions for a radio aware scheduling algorithm for WiMAX networks. Wireless Netw. **17**(6), 1441–1469 (2011)
25. RG Garroppo, S Giordano, D Iacono, Radio-aware scheduler for WiMAX systems based on time-utility function and game theory, in *IEEE Global Communications Conference (GLOBECOM)* (Hawaii, USA, 2009)
26. S Ali, M Zeeshan, A utility based resource allocation scheme with delay scheduler for LTE service-class support, in *IEEE Wireless Communications and Networking Conference (WCNC)* (Paris, France, 2012)
27. S Ali, M Zeeshan, A Naveed, A capacity and minimum guarantee-based service class-oriented scheduler for LTE networks. Eur. J. Wireless Commun. Netw. **2013** (2013)
28. G Monghal, KI Pedersen, IZ Kovács, PE Mogensen, QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution, in *IEEE Vehicular Technology Conference (VTC)* (Singapore, 2008)
29. M Assaad, A Mourad, New frequency-time scheduling algorithms for 3GPP/LTE-like OFDMA air interface in the Downlink, in *IEEE Vehicular Technology Conference (VTC)* (Singapore, 2008)

30. C Mehlfuhrer, M Wrulich, JC Ikuno, D Bosanska, M Rupp, Simulating the long term evolution physical layer, in *European Signal Processing Conference (EUSIPCO)* (Glasgow, Scotland, UK, 2009)
31. C Mehlfuhrer, JC Ikuno, M Simko, S Schwarz, M Wrulich, M Rupp, The Vienna LTE simulators - enabling reproducibility in wireless communications research. EUR. J. Adv. Signal Process. **2011**(29) (2011)
32. Tech. Specif. Group services and system aspects - policy and charging control architecture, 3GPP TS 23.203 V9.3.9, Release 9 (2009)
33. P Seeling, M Reisslein, Video transport evaluation with H.264 video traces. IEEE Commun. Surv. Tutorials, in print. **14**(4), 1142–1165 (2012). [Traces available at trace.eas.asu.edu]
34. J Shin, JW Kim, CCJ Kuo, Quality-of-service mapping mechanism for packet video in differentiated services network. IEEE Trans. Multimedia. **3**(2), 219–231 (2001)
35. Cisco visual networking index: global mobile data traffic forecast update 2011-2016. White Paper, Cisco (2012)
36. S Schwarz, C Mehlfuhrer, M Rupp, Throughput maximizing multiuser scheduling with adjustable fairness, in *IEEE International Conference on Communications (ICC)* (Kyoto, Japan, 2011)
37. X He, K Niu, Z He, J Lin, Link layer abstraction in MIMO-OFDM system, in *International Workshop on Cross Layer Design (IWCLD)* (Jinan, China, 2007)
38. AKF Khattab, KMF Elsayed, Opportunistic scheduling of delay-sensitive traffic in OFDMA-based wireless networks, in *International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM)* (Buffalo-NY, USA, 2006)
39. M Iturralde, A Wei, T Yahiya, A Beylot, Performance study of multimedia services using virtual token mechanism for resource allocation in LTE networks, in *IEEE Vehicular Technology Conference (VTC)* (San Francisco, USA, 2011), pp. 1–5

doi:10.1186/1687-1499-2014-138
Cite this article as: Khan *et al.*: QoS-aware composite scheduling using fuzzy proactive and reactive controllers. *EURASIP Journal on Wireless Communications and Networking* 2014 **2014**:138.